A Method for Clustering Hemagglutinin Influenza Protein Sequences

X. Li, H. Jankowski, X. Wang, J. Heffernan

September 13, 2014

Abstract

Each year influenza causes widespread disease globally. To combat the virus, vaccination programs are in place, but with the high mutation rate of influenza this vaccine needs to be updated every year. There is a high level of variability in the Hemagglutinin (HA) glycoprotein component of seasonal influenza strains. A better understanding of HA evolution over influenza seasons is needed to better advise vaccine strain development. We have developed a new method for clustering influenza viral sequences. Herewithin, we apply the method to the highly variable influenza A H3N2 HA viral glycoprotein. Our data comprises 1960 viral protein sequences active from 1998 to 2012, and our methodology aggregates these sequences into 23 clusters. Based on these clusters, we provide an investigation of past vaccines and the dominant cluster in each influenza season. We also investigate evolutionary pressures of closely matched circulating and vaccine strains HA glycoproteins. We end with a discussion of future work.

1 Introduction

Influenza viruses are negative stranded and segmented RNA viruses which cause serious and recurrent respiratory disease globally. The influenza season (or 'Flu Season'), which occurs each year in the winter months of each hemisphere, is associated with significant human mortality and morbidity worldwide[1]. Influenza A and B strains cause seasonal influenza, however, in recent years, influenza A has caused much of the burden[1].

Influenza A viruses are divided into subtypes based the hemagglutinin (HA) and neurominidase (NA) proteins that lie on the surface of the virus. The HA protein has been identified to be the major antigenic component of the influenza virus[2]. Seasonal flu epidemics are normally associated with small mutations in the HA protein (antigenic

drift), which allows the virus to evade host immune systems and increases the lifetime susceptibility of the host[2].

Vaccination is the best way to prevent or lessen the severity of seasonal flu. Each year, the seasonal flu vaccine is updated based on a study of the previous year's circulating strains[1]. However, because of the high mutation rate of the seasonal influenza strains it is difficult to determine what the appropriate vaccine strains should be[3]. A study of the HA's evolutionary course can better inform vaccine recommendations.

Phylogenetic analysis of virus strains can help determine evolutionary patterns. Recent advances in molecular biology and computational tools have enabled phylogenetic analysis of small components, to the entire influenza genome[4, 5, 6, 7, 8]. However, few pay attention to the influenza swarms or clades regarded as the main target driven by evolutionary forces [9, 10]. Additionally, in those studies that do consider the determination of clades, the statistical methods can be improved. For example, in [9], some components of the methodology include user-based decisions, so the method is not fully automatic. Also, in [10] the method is computationally complex, and could be improved.

We present a new formal cluster-based technique that can be used to study the evolution of influenza. The method is fully automatic with computational complexity O(N). Herewithin, we employ our methodology to study the evolution of the HA component of the influenza A H3N2 virus, (a major cause of seasonal influenza), and the relationships between this circulating virus and the recommended vaccine strains. We present our data acquisition and methodology in Section 2. In Section 3, we show that our new method can be used to uncover interesting trends in HA evolution, including a relationship between the vaccine and dominant circulating strains. Our results are discussed in Section 4, and this section ends with a critique of our methodology and directions for future work.

2 Data Description and Methodology

2.1 Data acquisition

The Influenza Research Database (IRD) is an online repository of influenza sequences obtained through voluntary contributions[11]. It is publicly available online at `www.fludb.org`. The sequences employed in this study were obtained from this data base using the criteria listed in Table 2.1.

This search yields a collection of 1947 sequences of the H3 type HA gene isolated between September 1998 and July 2012 from locations around the globe. Each sequence is made up of 1698 nucleotides plus a stop codon.

To understand the relationship between the observed influenza A strains and the vaccines used, we also include vaccine strains in the data set. The vaccine sequence

Table 1: IRD criteria: All other settings kept default or blank.

Option	Criteria
"Data to return":	protein
"Virus type":	A
"Sub type":	H3N2
"Select segments":	HA
"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

information was obtained from the World Health Organization[12], and is listed in Table 2.1. It includes all vaccine strains used from September 1998 to July 2012. Three vaccine strains were already included in the dataset, namely "A/Brisbane/10/2007", "A/Perth/16/2009" and "A/Texas/50/2012". All others were input manually. Note that this resulted in a total of 1960 sequences in the database.

The resulting 1960 sequences were then translated into the corresponding amino acids using Perl or MEGA software[13]. This was done separately for the virus and vaccine strains. The transformation resulted in 566 amino acids in each of the 1960 sequences. We then performed multiple alignment for all of the 1960 protein sequences using MUSCLE software[14]. All sequences were easily aligned with only a few gaps present. Finally, we converted the character records of the amino acids into numerical values using Perl. This results in 1960 observations with 566 categorical variables, each containing 21 categorical states (20 for each kind of amino acid and one to represent a gap).

Each of these 1960 sequences is related to a calendar year, country and city of isolation, inferred from the sequence name (see, for example, Table 2.1). For the 1947 sequences obtained from the IRD, we can also obtain the date of isolation, which allowed us to partition the data into influenza seasons (October 1st through September 30th).

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`.

Table 2: 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

2.2 Clustering the sequences

To identify clusters of viral sequences, we use the following methodology.

Distances between any two sequences are calculated using Hamming distance[15]. For two sequences, $A = \{a_1, \ldots, a_{566}\}$ and $B = \{b_1, \ldots, b_{566}\}$ the Hamming distance is defined as

$$d_s(A, B) \quad = \quad \sum_{i=1}^{566} I(a_i \neq b_i), \tag{1}$$

where $I(E)$ is the indicator function equal to one if E is true, and is zero otherwise. Note that although $a_i, b_i \in \{1, \ldots, 21\}$, they are categorical in nature, and the Hamming distance preserves this property. Heuristically, the Hamming distance between any two HA sequences is the number of locations with different amino acid expressions.

The dimension of our data set is 21^{566}, which is not computationally manageable. We therefore perform some dimension reduction before proceeding with our clustering approach. The main idea behind our dimension reduction is to select a smaller number of sites (from the 566 amino acid sites) to use in our analysis. This is a reasonable approach, as is it well known that parts of HA sequences are well-conserved[16]. To identify the most variable (equivalently, least conserved) sites, we use the notion of entropy.

Suppose that X denotes a categorical random variable with distribution given by $P(X = k) = p_k$, $k = 1, 2, \ldots, 21$. The entropy of X is then defined as

$$H(X) \quad = \quad -\sum_{k=1}^{21} p_k \log p_k \,. \tag{2}$$

In our context, X represents the amino acid state with $k = 1, \ldots, 21$ possible categories, and p_k denotes the observed relative frequency at the site (with frequencies obtained from the 1960 observations). Using (2), we compute the entropy for each of the 566 sites. Note that $H(X)$ is always positive, and larger entropy indicates greater variability at a site. In other words, sites with larger entropy contain more variability in their amino acid states. In our approach, the 566 entropies (one for each site) are sorted increasingly. Sites with entropy equal to zero were removed, as there is no amino acid variability in these sites and hence no useful information for clustering. Sites where 1959 observations were equal, were also removed. Finally, we used a Gaussian mixture model to cluster the remaining entropies[17]. The algorithm results in 5 classes of entropies. We identify the class with the largest entropies as the sites with the greatest variability, and use these sites to cluster the sequences. This class contains 62 sites, which allows us to reduce the dimension to 21^{62}. That is, our data is now made up of 1960 sequences, where each sequence is of length 62.

To cluster the dimension reduced data set, we use the Hamming distance vector (HD vector) algorithm[20]. To understand the approach, consider a general set-up where p nominal categorical attributes are of interest and the jth attribute is categorized by m_j levels. The categorical sample space, Ω, is defined as the collection of all possible p-dimensional vectors of states. For us, $m_j = 21$ for each j, and $j = 1, \ldots, 62$. Therefore, each sequence can be seen as a vector of length 62 ($p = 62$), and each element of the vector is a value taken from one of 21 ($m_j = 21$ for all j) possible categories.

Any given dataset, which in our case can be represented as $\{A_1, \ldots, A_{1960}\}$, gives a distribution of distances on the sample space Ω from a fixed reference position in Ω. We denote this fixed reference position as $S = \{s_1, \ldots, s_p\}$. For a general dataset, we use n to denote the sample size (here, $n = 1960$). Recall the definition of Hamming distance given in (1), and note that it will take values in $0, 1, 2, \ldots, p$. The algorithm relies on the HD vector, which is defined as a $(p+1)$-element vector $U(S) = \{U_0(S), U_1(S), \ldots, U_p(S), \}$ where

$$U_q(S) \quad = \quad \sum_{j=1}^{n} 1(d_s(A_j, S) = q), \quad q = 0, \ldots, p.$$

Thus, $U_q(S)$ counts the number of all observations with Hamming distance to the given reference position S equal to exactly q.

Using the HD vector as a measure of distance, the algorithm proceeds iteratively. First, it uses Pearson's chi-squared statistic to test whether there exist clustering patterns. If all the data points are not uniformly distributed in the sample space, we can use the modified chi-squared statistic to calculate the cluster centre and cluster radius. This proposed HD vector algorithm detects one cluster at a time, and this cluster will be deleted from the existing dataset before the next search. Note that the algorithm is thus fully automatic, selecting both the clusters and the number of clusters.

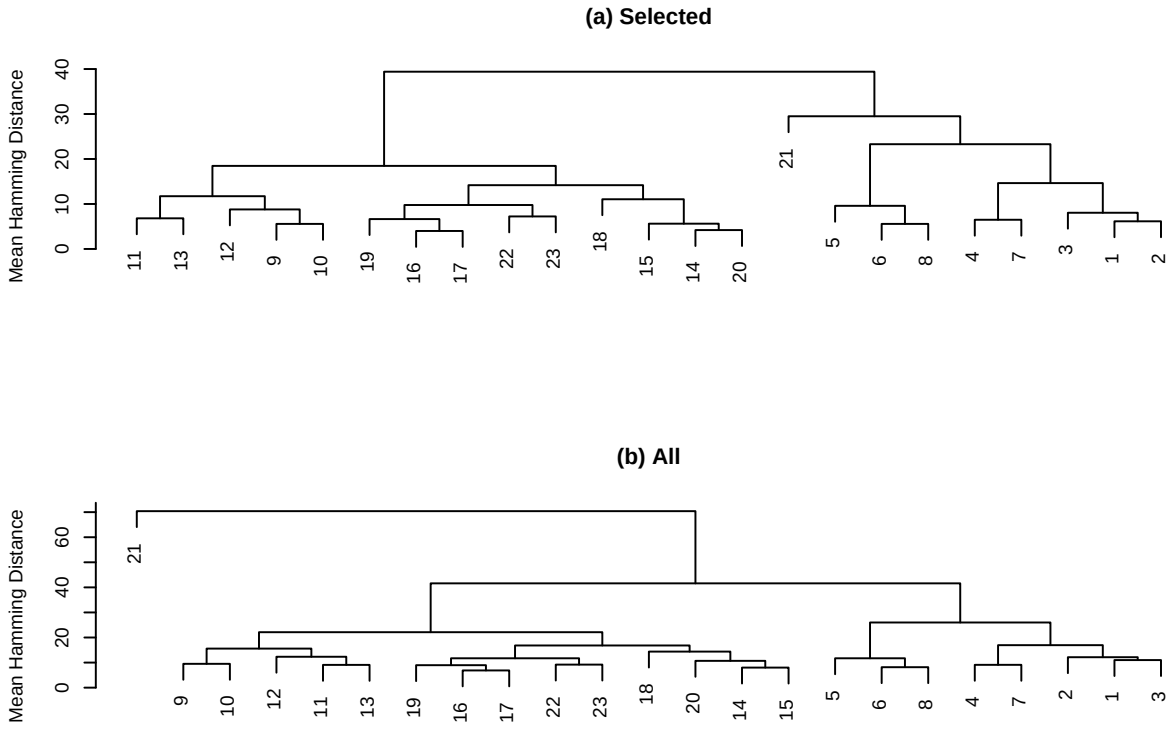

Figure 1: Dendrograms of clusters by mean Hamming distance. This plot is drawn using hierarchical cluster analysis with complete linkage. The top plot uses Hamming distance based only on the 62 highest entropy sites, whereas the bottom plot uses all 566 sites to calculate the Hamming distance.

The results of applying the HD vector algorithm to our data set of sequences is discussed in the following section. This discussion uses the notion of mean Hamming distance between two clusters, which we now define. Consider two clusters, $\mathcal{C}_1$ and $\mathcal{C}_2$. Each cluster is made of up of a number of sequences, say, $\mathcal{C}_1 = \{A_1, \ldots, A_{\kappa_1}\}$ and $\mathcal{C}_2 = \{B_1, \ldots, B_{\kappa_2}\}$. The mean Hamming distance is then

$$d_c(\mathcal{C}_1, \mathcal{C}_2) \quad = \quad \sum_{i,j} \frac{d_s(A_i, B_j)}{\kappa_1 \kappa_2},$$

if $\mathcal{C}_1$ and $\mathcal{C}_2$ are two different clusters. If $\mathcal{C}_1 = \mathcal{C}_2$, then we use instead

$$d_c(\mathcal{C}_1, \mathcal{C}_1) \quad = \quad \sum_{i<j} \frac{d_s(A_i, A_j)}{\kappa_1(\kappa_1 - 1)/2} \quad = \quad \sum_{i \neq j} \frac{d_s(A_i, A_j)}{\kappa_1(\kappa_1 - 1)}$$

This modification is due to the fact that when comparing the same cluster, all distances "along the diagonal" will always be equal to zero.

3 Results

3.1 Sites of Variability

As mentioned previously, to reduce the dimension of the system, we identified sites of high variability across the HA sequence. 62 sites were found to have the largest

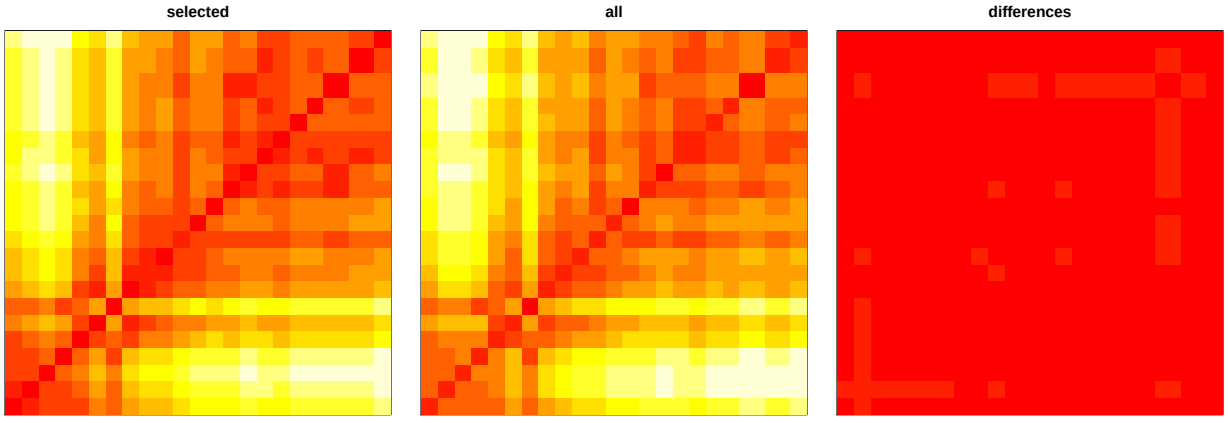

Figure 2: From left to right: mean HD matrix of 62 selected most varied sites, mean HD matrix of the whole sequence with 566 sites, absolute differences of the two matrices. In these plots cluster 21 is excluded. Both matrices (left and centre) have been standardized by dividing by their corresponding maximum values.

variability. Of the 62 sites, 52 lie within the HA1 domain. The remaining 10 sites lie within the HA2 domain.

3.2 Sequence Clusters

Using the method described in the previous section, the 1960 viral sequences were partitioned into 23 clusters. Figure 1 shows two dendrograms of the resulting clusters. The top dendrogram is based on the mean Hamming distance calculated only for the sites of maximal entropy, whereas in the bottom dendrogram the Hamming distance is calculated for all sites. Since the sites of highest entropy cover the mutation hot spots in the amino acid sequence it is expected that the dendrograms should be similar. Indeed, the dendrograms are largely consistent with regard to tree locations. Clusters 1 - 8 are grouped into a clade, while the remaining clusters except 21 are grouped into another clade. Although there exist subtle discrepancies in the specific clade location of some clusters, the mean distances are fairly small.

For a further comparison of the clusters using the entire amino acid sequence and the sites of maximal entropy we compare the Hamming distance matrices. Figure 2 (left, centre) shows the corresponding heat maps of the mean Hamming distance matrices, excluding cluster 21. Figure 2 (right) displays the absolute difference between these two distance matrices. We investigate further by considering also the ratio of the two Hamming distance matrices. The results are shown in Figure 3, and we can see that once cluster 21 is excluded, there appear only spurious inconsistencies off

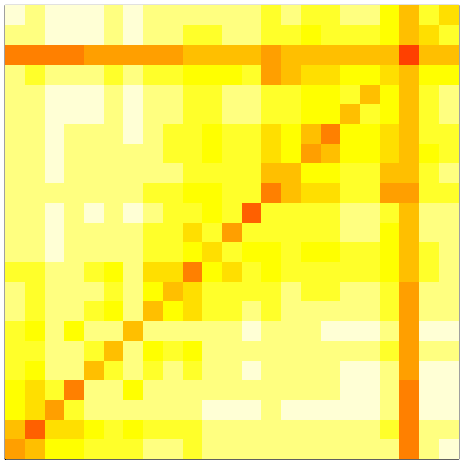

Figure 3: Ratio of distance matrices using the 62 most varied sites and all 566 sites in the Hamming distance calculations.

the main diagonal. Hence, the ratio is relatively constant (again, excluding cluster 21), which implies that the two distances (using the 62 sites and using all 566 sites) are approximately equivalent. Therefore, the plots indicate that, once cluster 21 is removed, the 62 most varied sites can stand for the whole sequence effectively with regard to the distances between sequences. In what follows, we proceed with our analysis of the results, keeping cluster 21.

3.3 Spatio-Temporal Evolution of the Clusters

Figure 4 shows the number of sequences in each cluster sorted by the first year of isolation. The clusters that house the vaccine strains are also indicated. The vaccine locations are consistent with the calendar year according to their strain name (and year closeby) eg. "A/California/7/2004" and "A/Wisconsin/67/2005", "A/Victoria/361/2011" and "A/Texas/50/2012" are clustered together. This helps to verify the validity of the cluster pattern.

We can observe from Figure 4 that large clusters are generally surrounded by clusters of much smaller size. Thus, the dominant clusters over many years can be identified. We can also observe that more small clusters are generated in recent years. This may be due to higher reporting rates, as rapid sequencing technologies have become increasingly available.

Each cluster houses strains that exist over one or more influenza seasons. In Figure 5, we plot the number of sequences in each cluster as a function of their isolation year. The size of each cluster is indicated by line thickness (with large clusters indicated with thicker lines). It is observed that some clusters are significantly more long-lived than others, but that no cluster spans more than seven years. It is also observed that clusters first increase and then decrease in size over their lifespan. Dominant clusters of viral sequences replace one another every 2-5 years, but once HA evolves away from a given region of the sequence space, it does not later revisit that region. This agrees with previous studies of influenza evolution[2, 9].

3.4 Evaluation of recommended vaccines for each season

In Table 3, we identify which clusters contain which vaccines, focusing on the clusters that exist over the calendar years 2000-2012. The dominant cluster, vaccine strain, and cluster that houses the vaccine strain for each year are provided for each vaccine. Ideally, the strain used as the basis for a vaccine each year would correspond to the dominant cluster. Considering the time lag that exists between the disease outbreak and time of isolation, the cluster housing the vaccine strain should be as close to the dominant cluster as possible.

A common observation over all of the years shown is that the relationship between

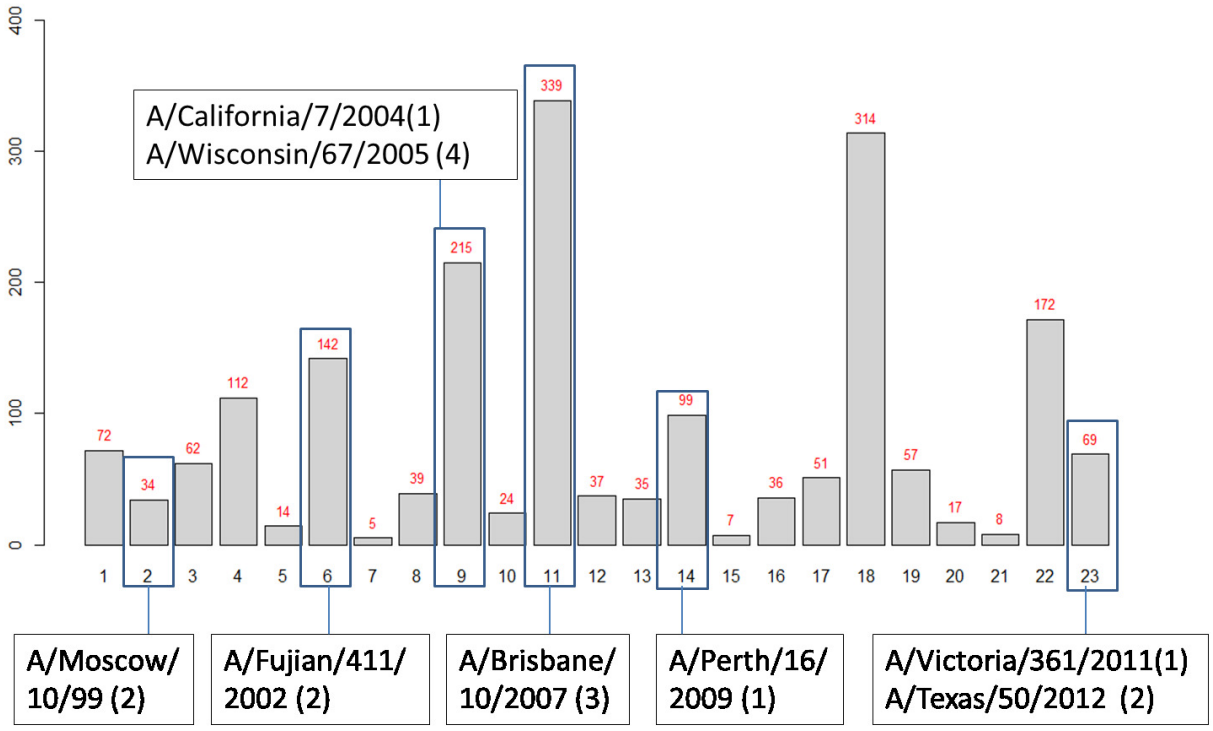

Figure 4: Histogram of cluster size and vaccine location. The clusters have been re-ordered by earliest year of isolation.

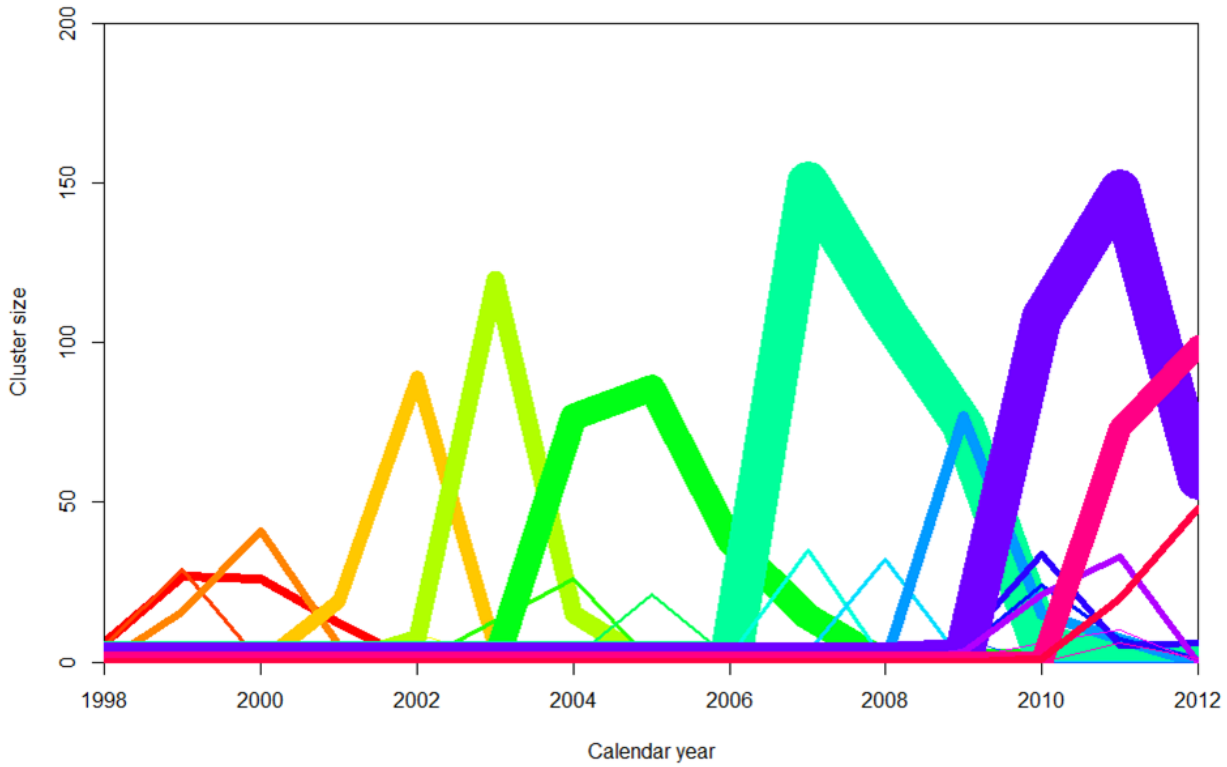

Figure 5: The number of HA protein sequences within each cluster plotted versus calendar year of isolation. Each cluster is indicated by a different colour, and the line width reflects the cluster size. The dominant sequence clusters tend to replace each other every 2-5 years.(The number of sequences each year is only for the unique sequences and does not reflect the severity of infections.)

the vaccine strain cluster and the cluster housing the dominant strain diverges (that is, is further apart on the dendrogram tree). For example, from 2000-2004 the same vaccine strain "A/Moscow/10/99" was used, however, the dominant cluster changes each year in this time period, moving from a mean distance of 6.15 amino acids (aa) from the vaccine sequence, to 8.37aa, and then 18.68aa in 2002-2004.

From Table 3 we can also see that cluster extinction often coincides with the existence or introduction of a well matched vaccine strain. In particular, the extinction of clusters 6 and 9 coincides with the introduction of vaccines housed in the same cluster. Something similar can be seen on clusters 1 and 11. Ultimate extinction of a cluster, however, is a result of a combination of various factors, including vaccine strain and competition between strains (that may have higher fitness). An exploration of strain fitness is a course for future work.

Table 3: Vaccines and Clusters by Year

Season	Vaccine	Cluster	Dominant cluster	Vaccine cluster
2000-2001	A/Moscow/10/99	**1** 3 4 5	1	2
2001-2002	A/Moscow/10/99	1 **4** 5 6	4	2
2002-2003	A/Moscow/10/99	4 5 **6** 7 8	6	2
2003-2004	A/Moscow/10/99	4 **6** 8 9	6	2
2004-2005	A/Fujian/411/2002	(6) 8 **9** 10	9	6
2005-2006	A/California/7/2004	1 (**9**) 10	9	9
2006-2007	A/Wisconsin/67/2005	(9) 10 **11** 11 12 13	11	9
2007-2008	A/Wisconsin/67/2005	(9) **11** 13	11	9
2008-2009	A/Brisbane/10/2007	(**11**) 14 15 16 17 19 20 21	11	11
2009-2010	A/Brisbane/10/2007	(11) 14 16 17 **18** 19 20	18	11
2010-2011	A/Perth/16/2009	11 (14) 16 17 **18** 19 20 21 22 23	18	14
2011-2012	A/Perth/16/2009	(14) 16 17 18 21 **22** 23	22	14

4 Discussion

In this paper, we have presented a new method for clustering protein sequences, and have applied the method to clustering of HA sequences of seasonal influenza A H3N2. Including vaccine sequences in the analysis, allows us to present important relationships between the vaccine and dominant influenza strain evolution.

In our approach, 62 sites of highest entropic variability were used in the clustering in lieu of the full sequence, greatly reducing the computational complexity of the problem. These 62 sites lie within the HA1 *and* HA2 regions of the HA genome. Our clustering methodology separated the HA dataset into 23 clusters. Analysis of these clusters found that HA generally clusters by year. Upon further analysis, we found that clusters replace one another every 2-5 years, that the evolution of the dominant cluster diverges from that of the cluster housing the vaccine strain, and that extinction of a dominant cluster often coincide with the existence or introduction of a well-matched vaccine.

Our results are highly consistent with previous studies of HA evolution (i.e., see [9, 10], and [26] for a review). Previously, Plotkin et al.[9] found that the persistence of clusters can be used to predict the next season's influenza sequences. In their analysis, however, only the HA1 component of the HA protein was considered. Therefore, some sites of high variability were neglected. Through choosing those most varied sites of the whole sequence, all the potential evolutionary hot spots can be taken into account. This can better help to find more diversified strains and evaluate the outbreak and epidemics of each cluster.

In a recent study, Luksza and Lassig [10] employed a new method using an ensemble of trees to infer the genealogy of influenza strains over time, trace the evolution of strain clades, and predict vaccine strains from year to year. A careful comparison of their results to that presented here is currently underway. It is important to note, however, that the computational complexity of their method is higher than the method we employ. Also, a key feature of our algorithm is that a statistical test is employed to determine whether a suspected cluster structure is justified. Such a test is not included in the methodology of Luksza and Lassig[10].

Our method can be applied to other components of the influenza virus genome. A similar study on the NA glycoprotein is currently underway.

Our methodology can be improved in several ways. The discussion in Section 3.2 indicates that cluster 21 behaves differently from the remaining clusters. The within and between mean Hamming distances based on the 62 high entropy sites of cluster 21 is similar to, or even smaller, than for other clusters. However, when the Hamming distance is calculated for all sites, the situation is considerably different. Here, cluster 21 has the greatest within and between distance values. This difference may be the result of numerous factors. It is possible that in the analysis, we have missed some hot spots or some diversifying mutations when choosing the amino acid sites using the

entropy approach. Alternatively, this could be the result of some inherent behaviour of cluster 21, which has yet to be understood. Note that cluster 21 is made up of only 8 sequences. It is important to investigate this discrepancy further, and is part of our ongoing research.

Secondly, we use the term "dominant" to denote clusters with the greatest number of sequences in a given season. However, this definition only indicates that the cluster contains the largest number of unique sequences. Therefore, the definition does not account for (a) the actual number of sequences reported in a season, or (b) their relationship with the frequency of the strain within the population. Although the first issue is relatively straightforward to fix, the second is more problematic. The influenza sequences available via IRD are based on voluntary contributions, and are therefore not the result of random sampling. It is thus possible that systematic biases exist in the data set, including yearly and regional variations[10]. Translating the observed sequences on IRD into an appropriate representation of population level frequencies is an important statistical problem which requires careful consideration in our future work.

Lastly, we point out that our analysis is based on the Hamming distance (1). This means that sequences close in Hamming distance (in amino acids) can be regarded as close in lineage evolution history. If we infer the genealogy of all the strains by ensemble of trees, those with small Hamming distance will be grouped into a clade, i.e. in the same trunk of the phylogenic tree. This approach is "purely mathematical" in that it does not include any potential information on the level of importance of specific amino acid differences, or their locations. Incorporating such additional information, will improve the quality of our analysis, and will therefore be included in future analysis.

References

[1] World Health Organization, http://www.who.int/influenza/en/.

[2] D.M. Knipe and P.M. Howley, *Philadelphia: Lippincott Williams and Wilkins.* (2007).

[3] F. Carrat, A. Flahault, *Vaccine* **25**, 6852-6862 (2007).

[4] W.M. Fitch, R.M. Bush, C.A. Bender and N.J. Cox, *Proc. Natl. Acad. Sci.* **94**, 7712-7718 (1997).

[5] W.M. Fitch, R.M. Bush, C.A. Bender, K. Subbarao and N.J. Cox, *J. Hered.* **91**, 183-185 (2000).

[6] T.T.-Y. Lam, Y.L. Chong, M. Shi, et al., *Infection, Genetics and Evolution* **18**, 367-378 (2013).

[7] K.B. Westgeest, C.A. Russell, X. Lin, et al., *J. Virol.* **88** 2844-2857 (2014).

[8] M. Escalera-Zamudio, M.I. Nelson, A.G.C Guemes, et al., *PLoS One* **9** e102453 (2014).

[9] Joshua B. Plotkin, Jonathan Dushoff, Simon A. Levin, *Proc. Natl. Acad. Sci.* **99**(9), 6263-6268 (2002).

[10] Marta Luksza and Michael Lassig, *Nature* **507**, 57-61 (2014).

[11] Influenza Research Database, www.fludb.org.

[12] World Health Organization, http://www.who.int/influenza/vaccines/virus/recommendations/en/.

[13] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar, *Mol. Bio. Evo.* **28**, 2731-2739 (2011).

[14] Robert C. Edgar, *Nucleic Acids Res.* **32**(5), 1792 - 1797 (2004)

[15] G. Forney, *Information Theory* **12**(2), 125-131 (1966).

[16] Zuzana Staneková, Eva Varečková, *Virology Journal* **7**, 351 (2010).

[17] B.S. Everitt and D.J. Hand. *New York: Springer.* (1981).

[18] Chris Fraley and Adrian E Raftery, *J. Amer. Statist. Assoc.* **97**, 611-631 (2002).

[19] Elodie Ghedin, Naomi A. Sengamalay, Martin Shumway, et al., *Nature* **437**, 1162-1166 (2005).

[20] Peng Zhang, Xiaogang Wang and Peter X.-K. Song, *J. Amer. Statist. Assoc.* **101**(473), 355-367 (2006).

[21] Arthur Chun-Chieh Shih, Tzu-Chang Hsiao, Mei-Shang Ho and Wen-Hsiung Li, *Proc. Natl. Acad. Sci.* **104**, 6283-6288 (2007).

[22] R Development Core Team, *R Foundation for Statistical Computing*, Vienna, Austria, http://www.R-project.org. (2008).

[23] 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).

[24] S. Bhatt, E.C. Holmes and O.G. Pybus, *Mol. Biol. Evol.* **28**, 2443-2451 (2011).

[25] Christopher J.R. Illingworth and Ville Mustonen, *PLoS Pathog.* **8**(12), e1003091 (2012).

[26] M.I. Nelson and E.C. Holmes, *Nature Rev.* **8** 196-205 (2007).