Evolution and Vaccination of Influenza Virus

Ham Ching Lam*, **Srinand Sreevatsan, and Daniel Boley**
University of Minnesota-Twin Cities Campus
Minnesota, MN 55455
United States of America

Abstract

In this study, we present an application paradigm in which an unsupervised machine learning approach is applied to the high dimensional influenza genetic sequences in order to investigate whether vaccine is a driving force to the evolution of influenza virus. We first used a visualization approach to visualize the evolutionary paths of vaccine-controlled and non-vaccine controlled influenza viruses in low dimensional space. We then quantified the evolutionary differences between their evolutionary trajectories through the use of within and between scatter matrices computation in order to provide the statistical confidence to support the visualization results. We used the influenza surface Hemagglutinin (HA) gene for this study as the HA gene is the major target of the immune system. The visualization is achieved without using any clustering methods or prior information about the influenza sequences. Our results clearly showed that the evolutionary trajectories between vaccine-controlled and non-vaccine controlled influenza viruses are different and vaccine as an evolution driving force cannot be completely eliminated.

1 Introduction

The rapid growth of the influenza genome sequence data due to the advanced development of sequencing technology in recent years has provided the opportunity for a more comprehensive sequence analysis of the influenza virus. The difficulty in sieving through and making sense of this mountain of data relying solely on phylogenetic approaches has become increasingly limited in part due to the poor scalability of the relevant algorithms [Nicholas, 2007]. Therefore, a different methodology needs to be utilized in order to take advantage of the massive amount of available data but at the same time be able to expose important information or structure within the data. Here, we present an application paradigm in which an unsupervised machine learning approach is applied to the high dimensional influenza genetic sequences so that the evolution of the vaccine controlled and non-vaccine controlled influenza viruses in the past century can be visualized. The main objectives of this study are twofold: (1) to visualize the evolution trajectories of influenza under vaccine pressure and in the wild without using any prior information about the viruses and (2) to provide statistical confidence to support the visualization results. Influenza virus is thought to have originated from a natural reservoir consisting of wild aquatic birds[Taubenberger and Kash, 2010; Webster *et al.*, 1992].

The influenza A virus is divided into subtypes based on differences in the surface proteins hemagglutinin (HA) and neuraminidase (NA), which are targets of the human immune system. Antigenic variants or immunologically distinct strains of A/H1N1, A/H3N2, and Type B have continued to emerge since its introduction into humans [Schweiger *et al.*, 2002]. Vaccination is the main strategy in stopping the infection and transmission of the virus in humans[Hannoun, 2013]. There are three components in a seasonal flu vaccine: (1) A/H1N1, (2) A/H3N2 and (3) Type B influenza. Each component is designed to fight the specific strain in each subtype that is predicted to be the dominant circulating strain in the upcoming flu season. Over the years, there have been over 20 vaccine updates for the A/H3N2 strain, over 16 updates for the Type B strain and 10 updates for the A/H1N1 strain. Each vaccine update is designed to provide immunity to the new antigenic variant that has emerged from the previous flu season. However, the long term effects of vaccination on the evolution of the virus itself is not clear. In order to shed light on this seemingly unsuspected problem, we used the nucleotide sequences from seasonal human A/H3N2 influenza virus from 1971 to 2009 as an example to demonstrate the evolutionary progress of this influenza virus against each successive vaccine introductions from 1971 to 2009. Figure 1 shows progression of influenza evolution based on the nonsynonymous substitutions (dN) and synonymous substitutions (dS) ratio analysis using the HA1 domain of the HA gene from A/H3N2 virus. The HA1 domain is a hypervariable domain of the HA gene where constant mutational changes can be observed due to the immune pressure generated from the host. A dN/dS ratio greater than 1 indicates the site is under positive selection pressure and is undergoing molecular adaption. In Figure 1, a constant shift of positively selected sites (blue color: dN/dS ratio greater 1) could be observed whenever a new vaccine (green square)

*Corresponding author email: hamching@cs.umn.edu

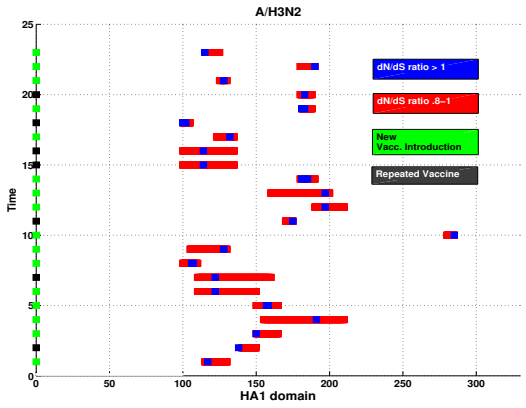

Figure 1: Seasonal human A/H3N2 influenza dN/dS ratio analysis against time of vaccine introduction. A constant shift of positively selected site location when a new vaccine was introduced. Horizontal axis represents the position of HA1 domain of the HA gene. Vertical axis represents time progression from 1971 (bottom) to 2009 (top) when each new (green square) and repeated (black square) vaccine was introduced. Red color bars denote the range of positions with dN/dS ratio from $0.8 - 1$. Blue color bars denote the range of positions with dN/dS ratio greater than 1.

was introduced which indicated that a new antigenic variant had emerged. When a repeated vaccine was introduced, the positively selected sites identified from the previous season remain unchanged. Given the results from the dN/dS ratio analysis, we compared the evolution trajectories of vaccine controlled to non-vaccine controlled influenza viruses and sought to better understand the effect of vaccination has on the evolution of influenza virus. In the present study, we used the human A/H3N2, A/H1N1, Type B, and avian H5 HA sequences as the vaccine-controlled samples. We used the human H5N1 and avian H5N1 HA sequences as the non-vaccine controlled samples.

2 Background

Influenza viruses have the ability to infect a very broad range of avian and mammalian hosts. Their genomic diversity is acquired through two biological mechanisms: antigenic drift and antigenic shift [Webster *et al.*, 1992]. Antigenic drift consists of the accumulated and continual mutations on surface proteins, resulting in the generation of antigenic variants. Of these surface proteins, we are focused on the hemagglutinin protein. Antigenic shift occurs when complete gene segments are exchanged among different subtypes of influenza viruses within a host cell, resulting in what effectively amounts to a whole new influenza virus genome. Both antigenic drift and antigenic shift allow for the virus to evade the host's immune response and rapidly adapt to new hosts [Caron *et al.*, 2009; Suzuki, 2006]. The evolution of influenza A virus is driven by the high rate of mutations and the ability to reassort gene segments. Because of its high rate of mutation combined with the lack of error correcting mechanisms during replication, influenza virus can easily generate different phenotypes that have the ability to survive within its host and infect others. To keep track of the evolution of the virus, annual update to the influenza vaccine composition is needed in order to provide a vaccine induced immunity to the general public [Boni, 2008]. The main process in influenza vaccine strain selection is to assess the match between the vaccine strain and the currently circulating strains and the potential new antigenic variant [Russell *et al.*, 2008]. If the vaccine strain does not match the currently circulating strains or the new antigenic variant that is likely to be the major variant in the upcoming influenza season, the vaccine composition is updated to contain a representative of the new variant [Russell *et al.*, 2008]. Each vaccine update is designed to provide immunity to the new antigenic variant that has emerged from the previous flu season. The seasonal influenza vaccine is used to prevent the infection and transmission of the virus, but its effect on the evolution of the virus itself is not clear.

3 Materials and Methods

In this study, utilizing the online NCBI influenza database[Bao *et al.*, 2008], we collected HA sequences from human A/H3N2, A/H1N1, Type B, and avian H5 HA sequences that represent the vaccine-controlled samples. We also collected human H5N1 and avian H5 HA sequences that represent the non-vaccine controlled samples. Table 1 lists the year range and number of HA nucleotide sequences from each sample.

Table 1: Vaccine controlled and non-vaccine* controlled human and avian sequences.

Samples	Year	Seqs
Human A/H1N1	1918-13	2140
Human A/H3N2	1968-09	175
Human Type B (Vic/Yam)	1970-13	818
*Human H5N1	1997-12	127
Avian H5 (Mexico)	1994-02	32
*Avian H5 (China)	1997-02	32

3.1 Influenza evolution visualization

All genetic sequences were first converted into binary strings according to the method outlined in [Lam *et al.*, 2012]. Nucleotide sequences are represented by strings of characters out of an alphabet of four letters: A, C, G, T. To obtain the binary string, each letter is replaced by a code of 4 bits: 1000, 0100, 0010, 0001, respectively. All binary strings were collected into a matrix to which Principal Component Analysis (PCA) [Jolliffe, 2002] was applied to extract the dominant variation from the dataset. Here, we briefly outline the sequence of steps involved in the PCA analysis. Consider a data matrix $X_{m,n}$ of dimensions m by n with m being the number of strains and n being the number of sites or positions (in this case, $n = 987 \times 4 = 3948$ for nucleotide sequences). Each row of X corresponds to a strain of virus and each column of X corresponds to a particular position. We first center the columns of the data matrix X with $\hat{X} = X - \frac{1}{m} e e^T X$

where e is a column vector of all ones, and then obtain the sample covariance matrix C from $\hat{X}$ by $C = \frac{1}{(m-1)}\hat{X}^T\hat{X}$. C is a square symmetric $n \times n$ matrix whose diagonal entries are the variances of the individual sites across strains and the off-diagonal terms are the covariances between different sites. The PCA algorithm is then applied to matrix C. The result is then visualized by plotting the top two or three principal components of the projected data. Since each strain is encoded as a binary string and PCA works at the binary data level, the pairwise distance relationship between the strains in a reduced space can be understood as follows: Let $\|s-t\|_H$ denote the pairwise Hamming distance between two strains s, t (number of differences in genetic sequences). Let $\|s-t\|_{bin\,1}$, $\|s-t\|_{bin\,2}$ denote the distance between the binary encodings of the two sequences (1-norm and 2-norm, respectively), and let $\|s-t\|_{proj}$ denote the 2-norm distance in lower dimensional space after projection onto the leading principal components. Every single change in the genetic sequence alphabet corresponds to changes to 2 bits in the binary encoding. Hence we have the relation between the distance in the lower dimensional space shown on the plots with the Hamming distance among the original sequences: $\|s-t\|^2_{proj} \leq \|s-t\|^2_{bin\,2} = \|s-t\|_{bin\,1} = 2\|s-t\|_H$.

3.2 Quantification

In order to provide statistical support to the graphical results obtained, we performed a statistical analysis based on a method that combined a multi-class scatter matrix computation and class labels randomization. The projected data points served as the viruses' 2-D coordinates and the year of isolation of each virus served as the class label. The multiclass scatter matrix involves the computation of Between-class matrix (**B**) and Within-class matrix (**W**) (Box 1). These computed matrices were not used explicitly as we only sought the trace of **B** and **W**. These are just the scalar scatter values: sum of squared distances between points and their respective centers. The class separateness measure λ_o is the ratio of trace **B** over trace **W**. A large λ_o indicates that the classes or clusters are well separated between each other and that elements within a cluster are strongly related or share the same property. This is basically an estimate on how well a multi-class Fisher's linear discriminant could separate the classes [Alpaydin, 2010]. A class label randomization algorithm (Alg I) provided the "distance measure" as a surrogate for the probability of observing the observed λ_o by chance. This is because the area under the tail of the randomized λ distributions beyond the observed separateness values was below rounding error of 10^{-16} which made the computation of p-value not possible. The larger the 'distance', the less likely the observed λ_o is generated by chance.

Box 1:
Virus isolation year as class label
C: Number of Classes
N_i number of data points in class $i = 1, 2, ...C$

- $\lambda = \frac{tr(B)}{tr(W)}$
- B : Between Class scatter matrix
  - $\sum_i^C (u_i - M)(u_i - M)^T$
  - $M = \frac{1}{c}\sum_i^C u_i$ "global mean of dataset"
- W : Within Class scatter matrix
  - $\sum_i^C \sum_j^{N_i} (x_j - u_i)(x_j - u_i)^T$
  - u_i: mean of class i.

Alg. I: Estimate Separateness Measures
Let $\lambda_o = \frac{tr(B_o)}{tr(W_o)}$ be the observed separateness value.
Repeat $j = 1 : K2$
  Repeat $i = 1 : K1$
    generate a randomization of the class labels
    compute the within-cluster scatter W
    compute the ratio $\lambda_i = \frac{tr(B)}{tr(W)} = \frac{tr(T)-tr(W)}{tr(W)}$
  compute the mean μ and std σ for all $\lambda_{i=1,..K1}$
  compute the distance $d_j = \frac{\mu-\lambda_0}{\sigma}$
Compute the mean $\bar{d}$ and std $\hat{d}$ of all $d_{j=1..K2}$
Report the distance of λ_o from the mean in the form of $\bar{d} \pm \hat{d}$

4 Results

The application of high-throughput unsupervised method to the high dimensional influenza virus genetic sequence data has made possible the visualization of the evolution of the influenza virus in the span of almost half a century. In this study, we present the graphical results from visualization of vaccine and non-vaccine controlled influenza viruses based on their genetic sequences alone. The human influenza A/H3N2 has the highest number of vaccine updates among the three vaccine controlled influenza viruses circulating in humans. Given the observation that constant shifting of positively selected sites whenever a new vaccine was introduced, we sought to visualize the evolution trajectories of vaccine and non-vaccine controlled influenza samples. We also set out to compute the class or clusters separateness values for both vaccine and non-vaccine controlled samples using the multi-class scatter matrix computation method for both the before and after class labels randomization process. We performed 1000 runs of Alg I on these samples and listed the results in Table 2. The observed separateness values λ_o of vaccine controlled samples are consistently higher than the non-vaccine controlled samples. This suggested that the vaccinated samples have very good separability by isolation years.

In Figure 2, we observed that the human A/H3N2 viruses clustered around vaccine seed strains chronologically since their introduction into humans in 1968. The evolution trajectory is directional going from lower left to lower right in the figure. In Figure 3, two separate lineages of human Type B influenza are co-circulating and that each lineage shows

the same observational characteristics as the A/H3N2, Type B viruses are also clustered around vaccine seed strains. For the human H1N1 influenza virus, a single lineage (black) can be seen that corresponds to the pre-2009 swine H1N1 pandemic. A sudden jump or gap is illustrated in the visualization due to the fact that the pandemic swine H1N1 strain had replaced the classical A/H1N1 and began to evolve (directional trajectory) as it circulated among humans. A vaccinated avian sample was used (avian H5) to further understand the evolution characteristic of vaccine controlled influenza.

In late 1993, an outbreak of avian H5 influenza in poultry in Mexico was detected and a long term vaccination program was implemented in hope to bring the outbreak under control and to eradicate the virus [Lee *et al.*, 2004; Escorcia *et al.*, 2008]. The vaccination program was in effect for over 13 years but an increase in respiratory signs of disease was observed in vaccinated chickens [Escorcia *et al.*, 2008]. In other words, the vaccine strain used in the vaccination program no longer matched the circulating strain in the field. The vaccine strain (A/Ck/Mexico/CPA-232/1994) was isolated in 1993 and has been in used for the duration of the program for over a decade. Using the available genetic HA sequences from these vaccinated chicken, we produced a 3 dimensional PCA plot (Figure 5) to show the evolution of the field isolates from 1994 to 2002. The first observation from Figure 5 is that a directional evolutionary trend similar to other vaccinated samples can be seen in this figure. Second, a chronological pattern is obvious indicating that the virus had undergone constant evolution or antigenic drifted away from the early strains. A split in the evolutionary path can be seen occurring in the 1990s. This split or divergence has been reported in studies by [Lee *et al.*, 2004; Escorcia *et al.*, 2008] based on phylogenetic analyses conducted on the same sequence sample.

Figure 6 illustrates the evolution trajectory of the non-vaccine controlled human H5N1 influenza from 1997 to 2002. We included the human H5N1 virus as the 'control' since this subtype is not currently being vaccinated against in humans but is under active research due to its high mortality rate in infected humans. Figure 6 suggests that this subtype has evolved into a few dominant clusters since 1997. Three major evolutionary trends or clustering patterns can be seen originating from the center cluster which contains viruses from 1997. This also implies this influenza subtype has undergone HA gene diversification. Although it has diversified since 1997, the specific H5 HA gene identified in 1997 has remained present in these days [Wei *et al.*, 2012].

Figure 7 shows the evolution of non-vaccine controlled avian H5 influenza virus. The overall observation that arises from this figure is that rather than forming a restricted directional trend, the evolution of the virus is characterized by a collection of clusters scattered on the plot. The collection of clusters suggests a diverse pool of the genetic diversity of the virus. For the avian H5 subtype, a less focused evolutionary trend than vaccine controlled influenza viruses can be observed. The increased genetic diversity since 2000 has been observed by [Garcia *et al.*, 1997] and is captured in this figure with clusters scattered to the left and extended to upper and lower corner at almost the same time. This clearly suggests the co-circulation of multiple clades or sublineages of the avian H5 subtype. The diverse genetic diversity of the avian H5 represented by multiple clusters across a long time period indicated that the avian subtype in the wild evolves much slower than seasonal human influenza viruses.

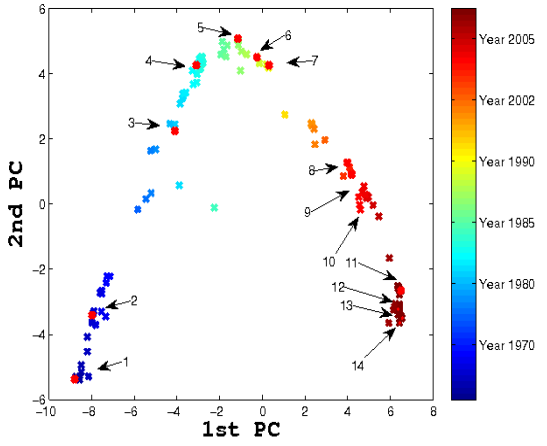

Figure 2: Seasonal human A/H3N2 influenza virus evolution trajectory. Each arrow points to a vaccine seed strain (red dot). The directional evolution can be seen as traveling from lower left to the top and then coming down to the lower right.

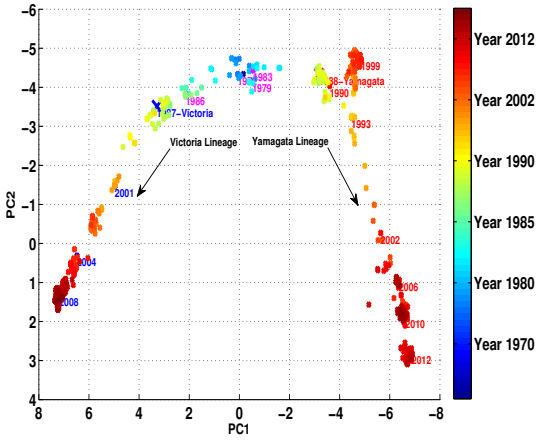

Figure 3: Seasonal human Type B influenza virus evolution trajectory. Two separate lineages (Victoria and Yamagata) are evolving simultaneously (top to lower left and to lower right). Vaccine introductions are indicated by year labels.

5 Discussions and Conclusions

Vaccination is the principal measure for preventing influenza and reducing its impact [Webby *et al.*, 2004; Wood *et al.*, 2001]. Almost a century ago after the isolation of the first influenza virus, influenza vaccines have been persis-

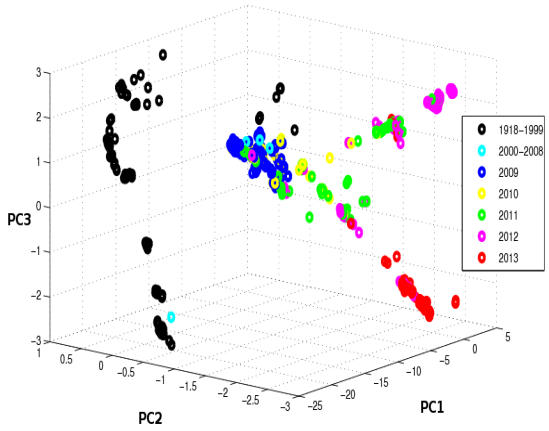

Figure 4: Seasonal human H1N1 influenza virus evolution trajectory in 3 dimensions. Pre-2009 pandemic viruses are in black. A clear separation can be seen after pandemic09 replaced the classic A/H1N1 strain. Separate lineages emerged indicating different genetic diversity.

Table 2: Class separateness results: Vaccine and non-vaccine* controlled human and avian samples

Sample (Human)	λ_o	**Distance**
A/H3N2 (1968-2009)	30.5	$978.3 \pm .031$
Type B:Victoria (1970-2013	26.3	$1310 \pm .02$
Type B:Yamagata (1970-2013)	25.3	$1327.8 \pm .019$
A/H1N1 (1918-2013)	24.7	$617.2 \pm .04$
*H5N1 (1997-2002)	1.01	$34.8 \pm .029$
Sample (Avian)	λ_o	**Distance**
Avian H5 Mexico (1994-2002)	1.7	$12.23 \pm .11$
*Avian H5N1 China (1997-2002)	0.268	$3.16 \pm .0.6$

tent and have evolved to respond to the evolution of the influenza viruses evolving in humans. [Gunn *et al.*, 2010; Hannoun, 2013]. Antigenic drift of influenza viruses occurs frequently among circulating strains that leads to new antigenic variants. However, whether the drift mechanism occurs with the presence of vaccine pressure is an important question that needs to be addressed at different level as vaccination is the primary method in prevention and protection for humans against influenza virus. Two studies [Hensley *et al.*, 2009; Lee *et al.*, 2004] have shown that vaccination forces mutations on the HA protein of the influenza virus. These mutations changed the way in which the virus gradually evolved and adapted to a new vaccine protected environment. Here, we extended the spectrum of analysis to include vaccine controlled human and avian samples and non-vaccine controlled human and avian samples to better compare and contrast and understand the evolutionary dynamic of influenza viruses under vaccine pressure. Using vaccinated and non-vaccinated samples from both human and avian hosts, we hope to minimize potential data selection bias and at the same time to provide a fair comparison across hosts under vaccination pressure. Our method utilized only the genetic composition of the HA sequences alone without using any specific clustering algorithms. As mentioned above and shown in Figure 1, genetic sequences contain important signals to detect evolutionary trends between different influenza subtypes under vaccination pressure. The genetic composition combined with the implicit positional information of the HA gene is enough to provide clues that the vaccine-controlled influenza viruses are under pressure to mutate in order to escape immune responses. Our method takes advantage of the binary coding of each sequence that preserves the positional information of each HA gene.

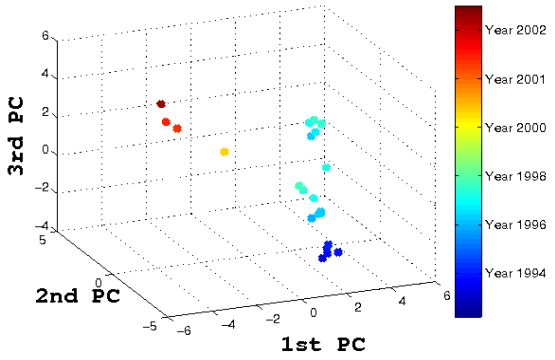

Figure 5: Vaccine controlled avian H5 influenza virus evolution trajectory in 3 dimensions. Vaccine was introduced in early 1990s and the virus slowly evolved away from the vaccine strain and established two separate lineages.

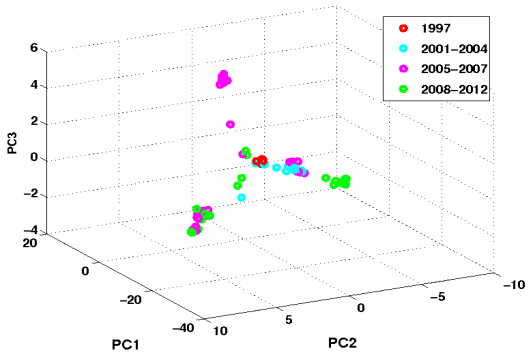

Figure 6: Non-vaccine controlled human H5N1 influenza virus evolution trajectory in 3 dimensions. The virus has evolved into a few dominant lineages since 1997. Three major evolutionary lineages can be seen originating from the center cluster which contains viruses from 1997. However, the specific H5 HA gene identified in 1997 has remained present in these days.

In this study, we have demonstrated that the evolutionary trajectories for vaccine controlled influenza are directional and restricted. The restricted directional evolutionary trends and clusters formation around the vaccine strains along the evolutionary paths exhibited by the vaccine controlled in-

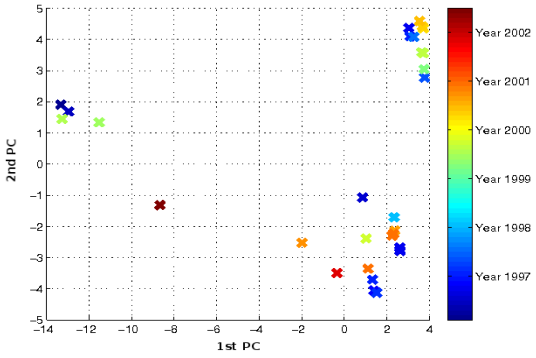

Figure 7: Non-vaccine controlled avian H5 influenza virus evolution trajectory in 3 dimensions. Multiple clusters scattered throughout sharing almost the same time periods suggesting the co-circulation of multiple clades or sublineages of the avian H5 subtype.

fluenza viruses are in sharp contrast to the non-vaccine controlled influenza viruses. Apart from this distinction, the naturally emerged chronological ordering of vaccine controlled influenza viruses in both two and three dimensional visualizations are much more noticeable than the non-vaccine controlled viruses. This natural chronological ordering reflects the active adaptation of the viruses to their changing environment. The class separateness measure exposes the fact that vaccine controlled influenza viruses that share the same isolation year have the tendency to cluster tightly together with good separateability. Each separate cluster or group represents a distinct genetic diversity of the virus group. In contrast, non-vaccine controlled influenza viruses isolated within the same time period appeared to be more scattered and the clusters exhibited much larger within cluster distance with no narrow restricted bands being observed. These observations suggested that the mutations on the HA gene were not restricted to certain sites alone and that the majority of these mutations most likely were synonymous nucleotide substitutions on the HA gene.

Also, the number of clusters observed are almost identical to the number of vaccine updates for the seasonal human A/H3N2 and influenza B viruses. The number of clusters observed in the seasonal human A/H1N1 is not the same as the number of vaccine updates but it does show the fact that this virus has been gradually evolving away from the vaccine strains as time passes. Since the A/H1N1pdm09 pandemic strain replaced the A/H1N1 strains in 2009 as the H1N1 vaccine component, the virus can be seen as slowly evolving but has not changed to a new antigenic variant. The very low value of λ_o computed from non-vaccine controlled influenza viruses has clearly captured the fact that non-vaccine controlled viruses are not actively evolving by the year. In contrast, the vaccine controlled influenza viruses have been actively evolving and adapting to the changing environment constantly as new vaccine composition is being introduced year after year. This is clearly reflected in the very high λ_o value for vaccine controlled influenza viruses. Although our analysis was based on genetic sequences alone, the results suggested that a clear difference existed among influenza viruses evolving in a vaccine protected environment than in the wild. This difference is shown through the multi-class scatter computation of their evolutionary paths. This quantitative measurement also serves as a basic statistical support to the observed differences in the evolution dynamics between vaccine controlled and non-vaccine controlled influenza viruses.

There are other potential factors besides vaccination that can affect the evolution of influenza viruses, such as host specific immune response, the large difference in life expectancy between humans and avian species, vaccine efficacy and effectiveness, the transmission channel of the virus in difference environment, and geographical regions. These factors have not been considered in this present study because our overall objective is to present a genetic sequence only approach as the first step in understanding the evolution of influenza viruses in a protected environment. Our approach works directly at the sequence level with no prior assumption about the evolution of the virus. It is a departure from traditional one dimensional phylogenetic approach in that we visualize influenza evolution in 2D and 3D space. All phylogenetic methods make or rely heavily upon the assumptions about underlying evolutionary process [Jenkins *et al.*, 2002]. By using methods that avoid making assumptions about the parentage relations among the strains, we can avoid possible misinterpretation of the results. As has been shown in this paper, a data driven approach with no prior assumptions about the evolution of the influenza virus affords us a different perspective in directly visualizing how the virus evolves in a span of over half a century. This perspective has given us insight into the way we think about the driving forces behind the emergence of human seasonal influenza antigenic variant strains season after season. Perhaps, vaccination did play a role in forcing the virus to undergo a different evolutionary path in order to continue to establish itself in its occupied host. A definitively scientific conclusion cannot be drawn without a thorough study of the virus in a controlled experiment for an extended period of time which should no less to include multiple influenza epidemics in humans.

Acknowledgments

This research was supported in part by NSF grants IIS-0916750, IIS 1319749. Influenza research in Srinand Sreevatsan lab is funded by the **National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services**, under Contract No. HHSN266200700007C. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or NSF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[Alpaydin, 2010] E. Alpaydin. *Introduction to Machine Learning*. MIT Press, 2nd edition, 2010.

[Bao *et al.*, 2008] Yiming Bao, Pavel Bolotov, Dmitry Dernovoy, Boris Kiryutin, Leonid Zaslavsky, Tatiana Tatusova, Jim Ostell, and David Lipman. The influenza virus resource at the national center for biotechnology information. *J Virol*, 82(2):596–601, Jan 2008.

[Boni, 2008] Maciej F. Boni. Vaccination and antigenic drift in influenza. *Vaccine*, 26 Suppl 3:C8–14, Jul 2008.

[Caron *et al.*, 2009] Alexandre Caron, Nicolas Gaidet, Michel de Garine-Wichatitsky, Serge Morand, and Elissa Z. Cameron. Evolutionary biology, community ecology and avian influenza research. *Infect Genet Evol*, 9(2):298–303, Mar 2009.

[Escorcia *et al.*, 2008] Magdalena Escorcia, Lourdes Vzquez, Sara T. Mndez, Andrea Rodrguez-Ropn, Eduardo Lucio, and Gerardo M. Nava. Avian influenza: genetic evolution under vaccination pressure. *Virol J*, 5:15, 2008.

[Garcia *et al.*, 1997] M Garcia, DL Suarez, JM Crawford, JW Latimer, RD Slemons, DE Swayne, and ML Perdue. Evolution of h5 subtype avian influenza a viruses in north america. *Virus research*, 51(2):115–124, 1997.

[Gunn *et al.*, 2010] Jennifer Lee Gunn, Susan Craddock, and Tamara Giles-Vernick. *Influenza and public health: Learning from past pandemics*. Earthscan, 2010.

[Hannoun, 2013] Claude Hannoun. The evolving history of influenza viruses and influenza vaccines. 2013.

[Hensley *et al.*, 2009] Scott E. Hensley, Suman R. Das, Adam L. Bailey, Loren M. Schmidt, Heather D. Hickman, Akila Jayaraman, Karthik Viswanathan, Rahul Raman, Ram Sasisekharan, Jack R. Bennink, and Jonathan W. Yewdell. Hemagglutinin receptor binding avidity drives influenza a virus antigenic drift. *Science*, 326(5953):734–736, Oct 2009.

[Jenkins *et al.*, 2002] Gareth M Jenkins, Andrew Rambaut, Oliver G Pybus, and Edward C Holmes. Rates of molecular evolution in rna viruses: a quantitative phylogenetic analysis. *Journal of molecular evolution*, 54(2):156–165, 2002.

[Jolliffe, 2002] Ian T Jolliffe. *Principal component analysis*. Springer verlag, 2002.

[Lam *et al.*, 2012] HamChing Lam, Srinand Sreevatsan, and Daniel Boley. Analyzing influenza virus sequences using binary encoding approach. *Scientific Programming*, 20:3–13, 2012.

[Lee *et al.*, 2004] Chang-Won Lee, Dennis A. Senne, and David L. Suarez. Effect of vaccine use in the evolution of mexican lineage h5n2 avian influenza virus. *J Virol*, 78(15):8372–8381, Aug 2004.

[Nicholas, 2007] Barton Nicholas. *Evolution*. Cold Spring Harbor Laboratory Press, 1st edition, 2007.

[Russell *et al.*, 2008] Colin A Russell, Terry C Jones, Ian G Barr, Nancy J Cox, Rebecca J Garten, Vicky Gregory, Ian D Gust, Alan W Hampson, Alan J Hay, Aeron C Hurt, et al. Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*, 26:D31–D34, 2008.

[Schweiger *et al.*, 2002] B. Schweiger, I. Zadow, and R. Heckler. Antigenic drift and variability of influenza viruses. *Med Microbiol Immunol*, 191(3-4):133–138, Dec 2002.

[Suzuki, 2006] Yoshiyuki Suzuki. Natural selection on the influenza virus genome. *Mol Biol Evol*, 23(10):1902–1911, Oct 2006.

[Taubenberger and Kash, 2010] Jeffery K Taubenberger and John C Kash. Influenza virus evolution, host adaptation, and pandemic formation. *Cell host & microbe*, 7(6):440–451, 2010.

[Webby *et al.*, 2004] RJ Webby, DR Perez, JS Coleman, Y Guan, JH Knight, EA Govorkova, LR McClain-Moss, JS Peiris, JE Rehg, EI Tuomanen, et al. Responsiveness to a pandemic alert: use of reverse genetics for rapid development of influenza vaccines. *the Lancet*, 363(9415):1099–1103, 2004.

[Webster *et al.*, 1992] R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka. Evolution and ecology of influenza a viruses. *Microbiol Rev*, 56(1):152–179, Mar 1992.

[Wei *et al.*, 2012] Kaifa Wei, Yanfeng Chen, Juan Chen, Lingjuan Wu, and Daoxin Xie. Evolution and adaptation of hemagglutinin gene of human h5n1 influenza virus. *Virus genes*, 44(3):450–458, 2012.

[Wood *et al.*, 2001] John M Wood, KG Nicholson, M Zambon, R Hinton, DL Major, RW Newman, U Dunleavy, D Melzack, JS Robertson, and GC Schild. Developing vaccines against potential pandemic influenza viruses. In *International Congress Series*, volume 1219, pages 751–759. Elsevier, 2001.