Computational Analysis and Visualization of the Evolution of Influenza Virus

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Ham Ching, Lam

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctor of Philosophy

Daniel L. Boley

August, 2014

Acknowledgements

I wish to express my sincere gratitude to my advisor Professor Daniel Boley for allowing me the freedom to work on the problem of developing computational visualization methods to study the evolution of influenza viruses. I would also like to thank him for his advice, patience, guidance, encouragement, and support during my entire graduate studies. I have benefited from your teaching and I wish to say a heartfelt thank you. If I were allowed to express my appreciation and gratitude in a 'linear algebra' way, I would express it as:

$$\begin{bmatrix} Thank \\ \\ You \end{bmatrix} = \begin{bmatrix} sUpport & gUidance \\ \\ sUpport & gUidance \end{bmatrix} \times \begin{bmatrix} \Sigma ncouragement \\ \\ \Sigma ourage \end{bmatrix} \times \begin{bmatrix} Valuable - adVice \end{bmatrix}^T$$

I would like to express my sincere gratitude to Dr. Sreevatsan for his guidance and support and many useful and constructive dialogues. Without him, this doctoral work would not have been possible.

I also wish to thank my committee members Dr. Kuang and Dr. Myers for their time and suggestions during the course of my graduate studies.

Last but not least, I wish to thank my family, near and far, for always being there for me. Thank my best buddy "Daniel", my next best buddy "Benjamin", and the soon to be my next next best buddy "Renae".

Dedication

To those who give so others can succeed.

Abstract

Influenza viruses can infect a large variety of birds and mammals including humans, pigs, domestic poultry, marine mammals, cats, dogs, horses, and wild carnivores [1]. Surveillance for influenza viruses circulating in humans has been gradually increased and expanded to many areas around the world. These surveillance programs have produced large amount of influenza genomic data which facilitates the study of the virus by computational methods that are efficient and cost saving.

The main focus of this dissertation research is the development of visualization methods to understand the evolution of influenza viruses circulating in humans and other mammals. The methods developed have been applied to different human influenza A subtypes, swine influenza viruses, and avian influenza viruses. The methods are based on unsupervised dimensional reduction techniques which can be applied to each individual genome segments or to the complete genome sequence of the virus. These methods are a departure from the traditional phylogenetic tree construction paradigm because very large number of high dimensional input sequences can be processed and results are viewed directly in a two or three dimensional Euclidean space.

We reproduced the evolutionary trajectory of the seasonal human influenza A/H3N2 virus since its introduction to humans in 1968 on a 2D PCA space. The observed pathway led us to hypothesize that vaccination serves as a primary evolutionary pressure

on this virus. We provided visual, simulation results, and statistical results to support this hypothesis.

The North American swine influenza H3N2 viruses were also studied using the developed visualization methods. The diversity of this virus is changing since the 2009 H1N1 pandemic outbreak. Five main clusters were observed from the visualization results. The mutations at two positive selected sites on the HA gene were identified as the potential driver for clusters segregation of this virus after the pandemic.

A visualization method was developed to visually detect reassortant influenza virus. A reassortant influenza virus is difficult to detect because it consists of genome segments from different parental origin. As two different strains of influenza coinfect a single cell, the capability to exchange genome segments between these two strains can lead to progeny carrying different parental segments within its genome. In order to detect such progeny, a PCA projection based visualization method that is able to examine the full genome sequence of a reference and test strains simultaneously was developed in order to detect any reassorted segments within a full genome.

Besides the development of visualization methods, we have also developed a compact Markov Chain model to estimate the probability of viruses with high genetic similarity found after a very large time gap. This model is a two components model where we combined a Markov Chain with a Poisson model. The Markov model uses Hamming distance as the evolution process of the virus and a computed mutation rate as the input to the Poisson model, combined together, we simulated the evolution process of

the influenza virus under the neutral evolution process. The computational results from this model led us to conclude that the existence of reservoirs preserving viruses for decades cannot be completely eliminated.

In short, our primary goal has been to develop visualization based approaches to understand the evolution of the influenza viruses from different hosts. The results we have so far suggested that the power of visualization paves the way to gain deeper understanding and insight of the evolution of the virus as we utilize the rapidly growing amount of the genomic data of the virus.

Contents

Acknowledgements i

Dedication iii

Abstract iv

List of Tables xii

List of Figures xiii

1 Introduction 1

1.1 Introduction 1

1.2 Problem of Interest 2

1.3 Contributions 6

1.4 Limitations of the study 7

1.5 Dissertation organization 8

1.6 Influenza virus 9

1.6.1 Evolution of Influenza A Virus 10
1.6.2 Hemaggluttinin gene 12
1.6.3 Internal proteins 13
1.7 History of Influenza Virus and Vaccine 15

2 Background and Review of Literature **18**

2.1 Introduction 18
2.2 Genetic sequence data 19
2.3 Computational methods 25
2.4 Cluster Analysis 26
2.5 Influenza Antigenic Distance 27
2.6 Influenza Antigenic Cartography 28
2.7 Influenza Virus Evolution and Vaccine 30
2.8 Phylogenetic Analysis 33
2.8.1 Phylogenetic tree construction 33
2.8.2 Methods 34
2.8.3 Challenges 37

3 Novel Markov Model **40**

3.1 Introduction 40
3.2 Compact Markov Model 41
3.3 Hemagglutinin Sequence Data 44

3.4 Results . . . 46
3.5 Conclusions . . . 55

4 Influenza Evolution Analysis with Binary Encoding Approach 56

4.1 Introduction . . . 56
4.2 Background . . . 57
4.3 Genetic Sequence Conversion . . . 58
4.3.1 Incorporating amino acid biophysical information . . . 59
4.4 Principal Component Analysis . . . 60
4.4.1 Limitations of PCA . . . 62
4.4.2 Application of PCA to Influenza Data . . . 63
4.4.3 Singular Value Decomposition . . . 67
4.4.4 Low-Rank Approximations . . . 68
4.5 Clumpiness measure . . . 69
4.5.1 Multi-class scatter computation . . . 70
4.5.2 Class separateness measure . . . 71
4.6 Materials: influenza genetic sequence data . . . 73
4.7 Results . . . 75
4.7.1 Seasonal human influenza H3N2 virus . . . 75
4.7.2 Seasonal human Type B influenza virus . . . 77
4.7.3 Seasonal human influenza A/H1N1 virus . . . 84
4.7.4 Human influenza H5N1 virus . . . 87

4.7.5 Avian H5 influenza viruses . . . 89
4.8 Discussion and Conclusions . . . 93

5 Influenza Reassortant Prediction 100

5.1 Introduction . . . 100
5.2 Background . . . 102
5.3 Method . . . 103
5.3.1 Data Processing . . . 104
5.3.2 Reassortants Detection . . . 104
5.3.3 Automated detection . . . 105
5.4 Results . . . 107
5.4.1 Swine influenza H3N2 Reassortant Virus . . . 107
5.5 Discussion and Conclusions . . . 111

6 North American Swine H3N2 Influenza 113

6.1 Introduction . . . 113
6.2 Background . . . 114
6.3 MATERIALS AND METHODS . . . 115
6.3.1 Sequence data . . . 115
6.3.2 Methods . . . 115
6.3.3 Multiple sequence alignment and conversion . . . 115
6.3.4 Weight assignment . . . 116

6.3.5 Phylogenetic analysis 116
6.4 RESULTS 117
6.4.1 Clustering analysis 117
6.5 DISCUSSION 125

7 Conclusions **127**

7.1 Summary of contributions 127
7.2 Compact markov model 128
7.3 Influenza virus and vaccine 128
7.4 Influenza reassortant detection method 131
7.5 Cluster analysis of North American swine influenza virus 131
7.6 Future directions 132
7.6.1 Investigate other genome segments 132
7.6.2 Vaccine strain/New antigenic variant prediction 132
7.6.3 Kernel PCA application 133
7.7 Final remarks 134

References **136**

List of Tables

1.1 Influenza genome 10

3.1 H1N1 strains with long time gap 46

3.2 H2 subtype long time gap strains 48

4.1 **Human and avian datasets** 75

4.2 Human samples 95

4.3 Avian samples 95

List of Figures

1.1 Hemagglutinin protein 3

1.2 Hemagglutinin protein evolution trend 4

1.3 Hemagglutinin protein dN/dS ratio 5

2.1 Sequence profile of an influenza HA protein. 22

2.2 Visualization based on codon vector approach. 25

2.3 Antigenic map of seasonal A/H3N2 influenza virus 31

3.1 Seasonal human influenza H1N1 virus pairwise distance 47

3.2 Poisson process distribution plot 49

3.3 H2 subtype histogram plot 50

3.4 Model prediction plot 52

3.5 H2 strains probability plot 53

3.6 Histogram of H1 from 1996-2006 54

4.1 PCA projection without hydrophobicity 60

4.2 PCA projection with hydrophobicity 61

4.3 Pairwise distance in 2D PCA space and Full space 66
4.4 Correlation between Hamming distance and PCA distance 67
4.5 Class scatter visualization 70
4.6 Influenza A/H3N2 evolution trajectory 77
4.7 Influenza A/H3N2 evolution in 3D space 78
4.8 A/H3N2 Class labels randomization simulation result 79
4.9 A/H3N2 Class labels distribution 79
4.10 Influenza B evolution 80
4.11 Influenza B evolution in 3D space 81
4.12 Type B (Yamagata) Class labels randomization simulation result 82
4.13 Type B (Yamagata) Class labels distribution 82
4.14 Type B (Victoria) Class labels randomization simulation result 83
4.15 Type B (Victoria) Class labels distribution 83
4.16 Seasonal human influenza A/H1N1 virus evolution visualization 85
4.17 A/H1N1 Class labels randomization simulation result 86
4.18 A/H1N1 Class labels distribution 86
4.19 Human H5N1 in 3D space 88
4.20 Human H5N1 Class labels randomization simulation result 88
4.21 Avian H5 evolution 90
4.22 Avian H5 Class labels randomization simulation result 91
4.23 Vaccinated avian H5 sample 92

4.24 Avian H5 (vaccine controlled) Class labels randomization simulation result 93

4.25 Singular Value Decomposition 96

5.1 Swine H3N2 double reassortant virus 108

5.2 Swine influenza reassortant virus 109

5.3 Swine influenza reassortant virus 2 110

6.1 HA weight distribution 117

6.2 North American swine influenza H3N2 phylogenetic tree 118

6.3 North American swine influenza H3N2 virus clusters 120

6.4 2013 swine H3N2 influenza viruses colored by isolation year 121

6.5 North American swine influenza H3N2 locations 122

6.6 2013 swine H3N2v 123

6.7 Weighted PCA clusters 125

7.1 Singular Values from A/H3N2 dataset 134

Chapter 1

Introduction

1.1 Introduction

Computational biology utilizes a broad spectrum of computational techniques to study varies problems arising from basic biology. As more genetic sequences become available for the model organisms, more basic research questions about the organisms can be answered using computational methods instead of relying on time consuming and costly benchwork. The model organism of this research study is the influenza virus (IV). Influenza virus is a persistent threat to global health which can kill up to fifty thousand people in the United States each year alone[2]. Influenza virus has been the subject of active research and study throughout the past century. The World Health Organization (WHO) Influenza Surveillance Network has established a network of influenza virus monitors and sample collection stations around the world to collect and

analyzes virological and epidemiological data related to influenza outbreaks. A direct consequence of this effort is the massive amount of influenza genetic sequences being collected, accumulated and deposited to online databases. This, in turn, has provided us with valuable data which makes the study of influenza virus through computational approach possible. The central focus of this dissertation is the study of the evolution of influenza virus by applying unsupervised statistical machine learning methods.

1.2 Problem of Interest

The main problem of interest is to study the evolutionary dynamics of the influenza virus through the use of computational tools. Specifically, we are interested in the type A influenza virus that circulates among humans and how does this type of virus evolves within a protected environment. A glimpse of the problem being studied in this dissertation arises from the observation presented in Figure 1.2. When hemaggluttinin sequence samples originated from vaccinated avian H5 (1994-2002) influenza and unvaccinated avian H5 (1997-2002) influenza was subjected to polymorphism analysis, we observed that the number of mutations were higher in the vaccinated sample than in the unvaccinated sample. In addition, more mutations are observed at the region of antibody binding sites. These sites are recognized by host's antibodies. When mutations are presented within these sites, antibody can no longer bind to the HA protein for neutralization, thus the virus can become an immune escape mutant.

Observation from Figure 1.2, Figure 1.2 and Figure 1.3 led us to hypothesize that

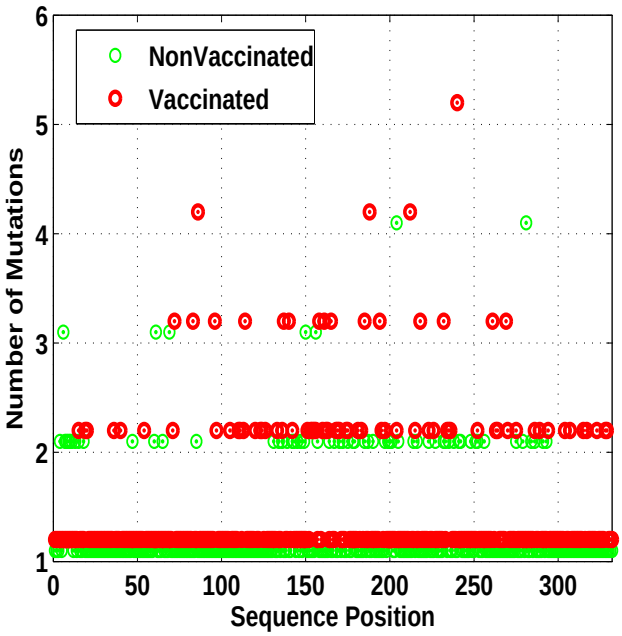

Figure 1.1: Hemagglutinin protein from vaccinated and unvaccinated influenza sample shows the number of mutations is larger in the vaccinated sample than in the unvaccinated sample. In addition, number of mutations are found to be higher in the hypervariable region of the HA protein of the vaccinated sample.

influenza virus evolves differently in a vaccinated environment than in a unvaccinated environment. In Figure 1.3, one sees the 'changing location' of the high dN/dS ratio sites as new vaccines are being introduced. A high dN/dS ratio (> 1) indicates the site is under heavy selection pressure and that the mutations on such particular site contributes to nonsynonymous changes that results in changes in the encoded amino acid. The dN/dS ratio was computed using vaccine strain against each season's dominant circulating strain from year 1968 to 2009. From this figure, we see that whenever a shift

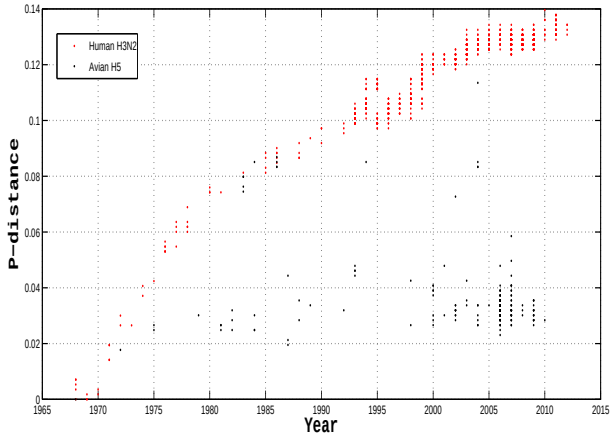

Figure 1.2: Based on the hemaggluttinin protein from 1968-2010, human H3N2 influenza sample shows that the p-distance is gradually increasing between the newer strains and its oldest strain. Avian H5 virus sample shows the p-distance between the newer and oldest strain stays at a constant throughout the 1968-2010 time period.

on the antigenic site (dN/dS ratio > 1) occurred, a new vaccine was introduced to target such antigenic variant. When the selection pressure remained the same on the same antigenic sites from the previous season, a repeated vaccine was used. A study by [3] indicated that the antigenic changes correspond to modifications in the virus and host relationship. Other factors that can drive mutation that leads to antigenicity change of the virus includes the rapid mutation rate of the RNA virus coupled with the lack of an error correcting mechanism during the replication cycle. This can result in progeny viruses with a fitness advantage that is able to survive the antibody neutralization from the host.

Our primary research interest is to develop computational methods to investigate

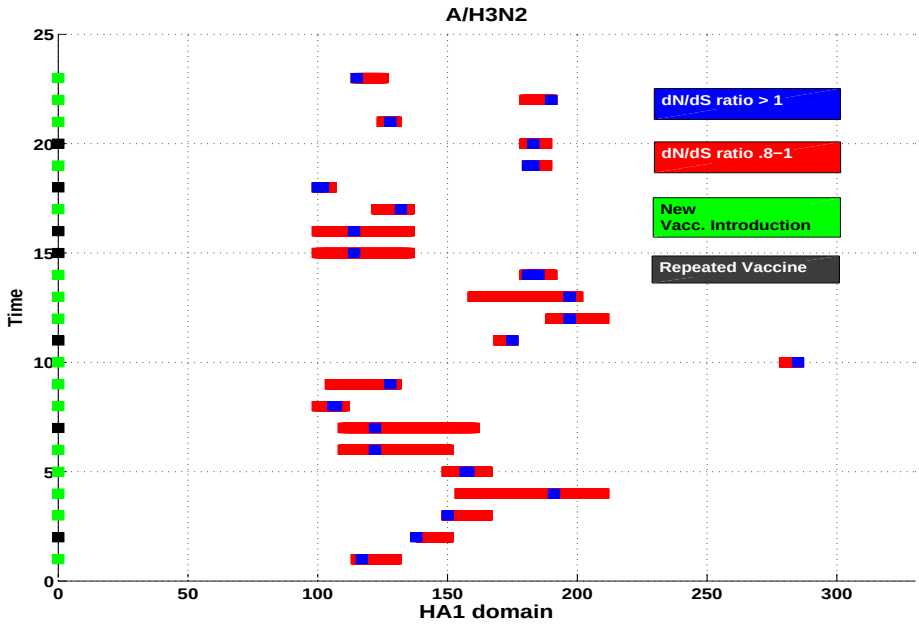

Figure 1.3: A/H3N2 HA gene. X-axis represents the location of each amino acid residue of the HA1 domain of the HA gene. Y-axis represents time in terms of new vaccine introductions (green squares) and repeated vaccine component (black square). Blue color indicates the highest dN/dS ratio observed on the HA1 domain and red color indicates a higher than average dN/dS ratio at the locations observed on the HA1 domain. The blue color squares are mostly located on the antigenic binding sites A and B on the HA1 domain.

the evolution of the human influenza virus in both vaccinated and unvaccinated environment. This dissertation also encompasses three other areas of interest: antigenic drift simulation using a stochastic model, reassortant detection based on a PCA projection technique and cluster analysis of North American swine influenza viruses. These three areas will bring depth to the study of the influenza virus and at the same time provide a broader coverage in understanding the evolution of the influenza virus in general.

1.3 Contributions

The effectiveness of influenza vaccine has been studied by many researchers and the focus has always remained to measure the vaccine effectiveness from the host's point of view. A subtle, yet important aspect that is often missed or not being studied is the effect of vaccination have on the evolution of the virus itself. Limited studies [4, 5] using phylogenetic analysis on very small number of avian samples collected within a short time span indicated that vaccination could play a role in the evolution of the virus. These studies have shown that antigenic drift of avian influenza viruses under vaccination pressure are similar to those observed with human influenza viruses. The findings from these studies encourage one to utilize the massive amount of collected influenza genetic sequence data for further and more in-depth investigation into the evolution of or antigenic drift of human influenza virus under vaccine pressure. Here, we have taken a computational approach to the investigation of human influenza under vaccine pressure as the main research objective. Below is a list of the contributions made:

- Developed a compact Markov Model to model the evolution of H1N1 influenza virus.
- Developed a high dimensional data visualization method to visualize the evolutionary trend of human influenza virus.
- Developed a projection method to predict influenza reassortant virus by visual

inspection.

- Provided computational evidence suggesting that influenza virus evolves differently in a protected environment than in the wild.
- Provided computational evidence which showed that North American swine H3N2 influenza virus segregation could be explained by two positively selected sites on the Hemagglutinin protein.

1.4 Limitations of the study

Often time, a research study is not without its limitations. There are a few unavoidable limitations that can potentially affect the quality of this study. The first such limitation in our study originates from the data source. The majority of the genetic sequences that were used in the study were downloaded from the NCBI online influenza database [6]. The database often contains incomplete or redundant sequences and the sparsity of samples from early years which can make the study difficult. These issues can cause bias in sequence database due to data curation bias or data sparsity. This bias is not uncommon but can potentially impact the results of the study if not treated with care. The second limitation are the factors that have not be considered in the study. These factors include the difference in host's life span, difference in the geographical location of the sample collected and the difference in host species. The third limitation is the extreme difficulty in finding vaccinated samples for the study. The fourth limitation is

that our study did not use antigenic cartography to directly capture the vaccine antibody associations. Our analyses were based on publicly available genetic sequences alone. Other factors may also have played a role in driving the evolution of influenza. These factors include 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 protected and wild environments.

1.5 Dissertation organization

The organization of this dissertation is as follows:

- Chapter 1 introduces goals and contributions to the field of computational influenza virus study and provides background information on influenza virus.
- Chapter 2 is a literature review chapter focuses on computational methods applied to the study of influenza virus.
- Chapter 3 describes a novel Markov model used to model influenza virus evolution.
- Chapter 4 describes the binary encoding technique used to study the evolution of influenza viruses.

- Chapter 5 describes a PCA projection based reassortant virus detection method.
- Chapter 6 presents a cluster determinant analysis of the North American swine influenza virus.
- Chapter 7 concludes this thesis with the final discussion and outlines the future directions.

1.6 Influenza virus

The influenza virus (IV) has been the focus of intensive research for the past century ever since the 1918 Spanish flu pandemic that killed from three to five percent of the world population at the time. This is one of the most deadliest pandemics in human history[7]. Influenza viruses have the ability to infect a very broad range of avian and mammalian species. Their genomic diversity is acquired through two biological mechanisms: antigenic drift (see section 1.6.1 and antigenic shift (see section 1.6.1). Both antigenic drift and shift allow for the virus to evade host's immune response and can rapidly adapt to the new host [8]. Influenza virus belongs to the viral family *Orthomyxoviridae* and classified into five genera: influenza A, influenza B, influenza C, Thogotovirus, and Isavirus [9]. All influenza viruses contain a segmented negative sense single stranded RNA genome. Currently, there are 17 A subtypes (H1 to H17) identified in the influenza A virus family. Influenza A has eight unique RNA segments [10] that encode 11 different gene products (PB1 polymersae, PB2 polymerase, PA polymerases,

PA-X polymerases, Hemagglutini (HA), Nucleoprotein (NP), Neuraminidase (NA), Matrix M1 and M2 proteins, and Nonstructural NS1 and NS2 proteins. Table 1.1 lists the eight segments of the influenza virus.

Table 1.1: Influenza genome

Segment	Length (NT)	Protein Name	Encoded polypeptide
1	2341	Polymerase basic 2	PB2
2	2341	Polymerase basic 1	PB1
			PB1-F2
3	2233	Polymerase acidic	PA
			PA-X
4	1778	Hemaggluttinin	HA
5	1565	Nucleoprotein	NP
6	1413	Neuraminidase	NA
7	1027	Matrix 1	M1
		Matrix 2	M2
8	890	Nonstructural 1	NS1
		Nonstructural 2	NS2

1.6.1 Evolution of Influenza A Virus

The evolution of influenza A virus is driven by two fundamental mechanisms: high rate of mutations and the ability to reassort gene segments. Gene reassortment is the exchange of the complete matching gene segments between two or more influenza viruses. This often give rise to the emergence of reassortant viruses and it is termed antigenic shift. For example, when the surface HA and NA genes are swapped, a new subtype of influenza [11] virus can emerge. Due to the lack of error correcting mechanism during replication [12, 13], difference genotypes can emerge that have the ability to survive

within the host. This mutational change in the viral genomic sequence caused by high rate of mutations over time allows influenza virus to escape antibody neutralization. The HA and the NA genes of the surface proteins of the virus are most likely to undergo mutation over time [14, 15].

Antigenic drift

Antigenic drift is the term that is often used to describe the mutations on the surface proteins of the influenza virus. The changes usually are found on the antibody binding sites of the HA protein. Once sufficient changes are made or accumulated, the antigenicity of the virus will change leading to a new antigenic variant. Antigenic cartography technique is used to establish the antigenicity of a virus. The data for constructing the cartography can come from HI binding assay or micro-neutralization assy tests of the homologous and heterologous HA types on sera from vaccinated individuals. The major concern about antigenic drift is that the human seasonal influenza vaccines requires yearly reformulation in order to provide protected immunity from the new antigenic strains [16, 17].

Antigenic shift

Antigenic shift is the ability of the virus to reassort gene segments and it is often thought of as the mechanism of introducing an antigenically distinct virus within a population that is different from currently circulating strains [16]. A reassortment event can take

place when a host cell is co-infected by two different influenza subtypes. Due to the segmented genome, influenza viruses can easily 'swap' their complete individual gene segments between one and another during replication. The reassorted virus can spread rapidly within human population and sometimes lead to a pandemic [7]. The most recent pandemic is the H1N1 swine origin from 2009 which originated from Mexico and spread around the globe within a very short time. The 2009 H1N1 swine pandemic virus H1N1pdm09 is a reassortant of genes (PB1, PB2, PA, HA, NP, and NS) from North American swine lineage and genes (M and NA) from Eurasian swine lineage [18]. A more recent reassortment event is the emergence of H3N2 variant, a classical TRIG virus that infects swine reassorted with pdmH1N109 to acquire the pandemic matrix gene [19]. These H3N2v strains have a higher velocity of spread within pig populations and has been reported in zoonotic transmission to humans [20].

1.6.2 Hemagglutinin gene

Hemagglutinin (HA) is the major envelope glycoprotein of the influenza A virus. The classification of influenza A virus into different subtypes is based on the antigenic specificity of each individual HA. Antigenic specificity refers to the ability of the immune system to recognize an antigen as a unique molecular entity and differentiates it from the other [21]. During membrane fusion, HA is cleaved into HA1 (about 329 amino acids) and HA2 (about 237 amino acids). The HA1 is a receptor binding protein and the major target of immune responses. HA2 is an anchor protein that mediates the

fusion of protein envelope and the cellular endosomal membrane [22, 23].

The virulence and pathogenicity of influenza A virus are often associated with its hemagglutinin gene. The human influenza A H3 protein is said to have five specific epitope regions (A, B, C, D, and E located on the globular head of the HA protein [24]. These regions tend to accumulate amino acid changes due to antigenic drift over time and eventually preventing antibody binding to the epitope region. This allows the virus to escape the host immune response and continue its replication and transmission.

Understanding the evolution of the HA genes is importance as it is a target of the current influenza vaccine. In terms of genetic similarity among all the HA subtypes, H13 and H16 share high degree of genetic similarity. H7 and H15 also are very similar in their genetic composition [25, 26]. Chen et. al., also showed that human influenza viruses have diverged into difference clades during the past decades. Results from Chen et. al., suggested that human viruses update rapidly and the viruses are maintained almost exclusively by humans themselves[26]. In total, 60 lineages and 83 sublineages within influenza A virus circulating around the globe based on HA protein alone have been identified by phylogenetic analysis [26].

1.6.3 Internal proteins

Although our research is focused on the surface proteins of the virus, a short background introduction of the internal genes of the virus is necessary. Influenza genome segments 1, 2, 3, 5, 7, and 8 encode the internal genes of the virus. The functions of

all the internal genes are not fully understood. Influenza polymerase proteins (PB1, PB2 and PA) encoded by segments 1, 2, and 3 are responsible for the replication of the eight different uncapped, non-polyadenylated, negative-sense RNA segments (vRNAs) that make up the viral genome [9]. Prior to translation by the host cell, a negative sense (3' to 5') viral RNA needs to be first transcribed into a positive sense RNA by an RNA polymerase. The nonstructural proteins NS1 and NS2 encoded in segment 8 are thought to be associated with the pathogenicity of the virus. Studies have shown that influenza virus with partial deletions in NS1 proteins are attenuated and do not cause disease [27, 28]. The attenuated influenza virus through partial deletions of the NS1 protein presents a different pathway through which live attenuated influenza vaccines for human can be designed and developed [29]. The NP protein is encoded by segment 5 and its primary function is to encapsulate the virus genome for the purposes of RNA transcription, replication and packaging [9]. Phylogenetic analysis of virus strains isolated from different hosts reveals that the NP protein is relatively well conserved across time, with a maximum amino acid difference of less than 11 percent [30]. Its role in the influenza virus life cycle involves polymerase interactions, M1 interactions, and Homo-oligomerization/NPNP interactions [30]. The broad spectrum of activities of this protein suggests that this NP gene is a key functional component of the virus. The membrane proteins M1 and M2 of influenza A viruses are thought to have established four major host related lineages based on phylogenetic analysis: (1) Equine, which has the most divergent M gene; (2) a lineage containing only H13 avian viruses; (3) a lineage

containing both human and classical swine viruses; and (4) an avian lineage subdivided into North American avian and avian viruses in general [31]. The M1 protein is evolving very slowly in all lineages, whereas the M2 protein shows significant evolution in human and swine lineages but virtually none in avian lineages[31]. The M protein is also a proton gating apparatus on the virion and is critical to viral replication. This feature may explain why M protein alone of pandemci H1N1 confers a wider host range.

1.7 History of Influenza Virus and Vaccine

Currently there are two dominant influenza A subtypes (A/H1N1 and A/H3N2) that are circulating in human populations. The A/H1N1 has the longest history since its first emergence in humans during the 1918 Spanish flu pandemic outbreak [32, 33]. The A/H1N1 abruptly disappeared from humans in 1957 and reappeared in 1977 in the former Soviet Union, Hong Kong, and northeastern China [33]. The disappearance of A/H1N1 was due to the replacement by the A/H2N2 strain which contained mixed genome segments from avian source and the A/H1N1 strain of 1918 Spanish pandemic lineage. A/H2N2 circulated in humans from 1957 to 1977 until A/H1N1 was detected again in 1977. The reason for the complete disappearance of this A/H2N2 was not clear [33]. A new A/H1N1pdm09 strain that caused the 2009 pandemic was first discovered in Mexico in the April of 2009 and spread around the globe in less than six months [34]. This new A/H1N1/pdm09 contains gene segments from triple reassortant swine influenza virus lineage and Eurasian influenza A (H1N1) swine virus lineage. This new

A/H1N1pdm09 is now the dominant A/H1N1 subtype circulating in human populations since 2009. Influenza A/H3N2 appeared in Hong Kong in 1968 and it has been in continual global circulation ever since. Although A/H3N2 strains appeared later than the A/H1N1 strains, there have been well over 20 annual vaccine updates associated with this virus since its outbreak in 1968. This shows that this virus has a much higher antigenic drift rate than the A/H1N1 and any other vaccine controlled influenza viruses. The fast rate of antigenic drift of this virus has remained one of the most challenging questions to date [35]. Influenza B emerged in the late 1970s and split into two antigenically distinct lineages since the early 1980s. These two lineages are referenced as Vic87 and Yam 88 respectively [36]. The viruses from these two antigenic distinct lineages have been cocirculating in human populations in particular time and regions [37]. Influenza B virus mutates at a rate that is 2 to 3 times slower and is genetically less diverse than influenza A [38].

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 [39]. 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 [17]. 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 [17]. 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.

Avian influenza (AI) or bird flu, is an infectious viral disease of birds, and most do not infect humans. Historically, human infections with avian influenza viruses have been rare and most of these viruses have caused only mild illness [40]. However, AI H5N1 and H7N9 have caused infections in people that have led to death [41, 42]. The majority of these cases of A(H5N1) and A(H7N9) infection have been found to be tied to the direct contact with live or dead poultry by the victims [41, 42]. Avian influenza A outbreaks occur in poultry from time to time around the world and in North America. Culling or depopulation of infected flocks is usually the preferred control and eradication methods when avian H5 or H7 influenza outbreaks occur in poultry. The usage of vaccination to control and to prevent infection from avian influenza viruses in poultry is generally banned or discouraged [43]. However, vaccination against the avian H5 and H7 influenza viruses in several occasions in recent years with the general objective of controlling and in some cases eradicating the disease has been used in isolated outbreaks in poultry farms. This does not mean that vaccination program against avian influenza viruses has been widely adapted or implemented across different poultry farms or regions. This is because vaccination as the primary control method has been unsuccessful on the larger scale [44, 45, 46, 47, 48, 43].

Chapter 2

Background and Review of Literature

2.1 Introduction

This chapter reviews the current computational methods used in studying influenza virus. The purpose is to elaborate on the specifics of the current methods so that alternative methods offered by the current research can be properly assessed and valued. This review focuses on methods that work at the primary structural (sequence) level of the influenza virus, computational methods that target the secondary structure of the virus will be described if they serve to validate the results generated from sequence level analysis.

2.2 Genetic sequence data

We are interested in direct sequence conversion, which is the 'first source approach' instead of the 'second source approach'. For example, a second source approach is to give the pairwise hamming distance between each sequence in the dataset and subsequent data processing steps are performed based on the pairwise distance matrix.Data exploration as the initial step in analysis is beneficial and useful when little is known about the data. When presented with large datasets, visualization is often used as the first step in data exploration. A large dataset can be either large in size or high in dimensionality or both. A good visualization method can often reveal structure, trends, outliers, hidden patterns, and relationships in the dataset. Once an initial understanding of the data is achieved, one can often make hypothesis about the study at hand. Thus, initial data visualization can be viewed as a hypothesis generation process. There are a few properties a good visualization should facilitate: (1) finding outliers, finding connections, finding patterns, or reveal hidden structure of the data; (2) finding sparse representation of the data while minimizing information loss; and (3) diagnostics for model fit and evaluation. These three are often the main objectives in performing data visualization but depending on the nature of the data or application, one is often faced with the challenge of achieving all three in one go [49, 50].

Many methods can be utilized to visualize a dataset on a two dimensional screen that are both effective and cost saving (in terms of time) in data exploration. For this chapter, we concentrate on presenting high dimensional data visualization methods

based on unsupervised machine learning approaches. The reason is that unsupervised approaches often require minimal manual intervention and understanding of the data at hand. Unsupervised visualization method if done properly can provide good qualitative overview of the large and complex data or even help in identifying regions of interest for focused analysis. For example, clustering and classification algorithm coupled with visualization can help to generate a visual aid that serves as a rough blueprint or guide to provide users with insights. This type of visual aid can be very useful in order to have a better understanding of the data in order to make better decisions involving any further quantitative analysis.

Computational methods that operate at the sequence level often are presented with a few limitations. The first limitation is not from the method itself but rather it is originated from the data source. Most often it is commonly known to be the data source bias problem. Influenza surveillance programs give top priority to antigenic novel isolates to be sequenced and deposited to databases which may cause sample bias problem in the data. The prioritization of isolates means that sequences were deposited to the database that are not representative of the overall influenza community cases. If the study is targeted toward a specific issue dealing with a specific influenza virus subtype, then this type of problem is not as damaging to the result. The second limitation is that not every sequence that is available in the database is a complete sequence, meaning that partial sequence that is considered to be 'important' or contains antigenic sites of the virus is sequenced and made available. For example, HA gene has

two domains, HA1 and HA2 domain respectively. Often time only HA1 portion which is the most variable portion of the complete HA sequence is available for download. A third limitation is that a sequence can be a mixed subtype sequence, this type of sequence is not commonly found in the database but they do present a special challenge in the sense that it is a mixed infection sequence and unless a specific study concerning mixed infection otherwise sequence like this is best removed from the dataset.

Datasets that are not easily presented using two or three dimensional visualization techniques (bar graph, charts, histogram, line plots, scatter plots) are often more difficult to handle and can demand more computational resources. Such datasets are usually of very high dimension and very large in size. These type of dataset often presents challenges both to researchers and computational resources respectively. Often, this type of data starts out as a very large matrix either dense or sparse in nature. The visualization of this large matrix is complicated by the noise embedded in the data which is not obvious to the data analyst. In this chapter, we focus on the gene genetic sequence data of influenza viruses.

Genetic sequence data usually consists of long strings of alphabets which can not be easily represented by numbers. For example, influenza hemagglutinin nucleotide sequence consists of 1698 (566 amino acids) bases and horizontal scrolling is needed when viewing the sequences or sequence alignment on the screen. For simple content visualization of any genetic sequences, there are software that can display multiple sequence alignment or some simple Matlab programs [51] that can help to display the

profile of a set of sequences. For example, figure 2.1 below shows a visual profile of a set of influenza protein sequences. This profile view gives a quick rough 'feel' of the genetic composition of this set of sequences. The y-axis has the twenty amino acids and the x-axis represents the positions of the sequence. Depends on the application at hand, visualization of genetic sequences often sought to reveal sites that are related to the function of the protein. Sites that are conversed are often thought to be related to the core function of the protein while sites that show variation are thought to be not important in maintaining the function of the protein.

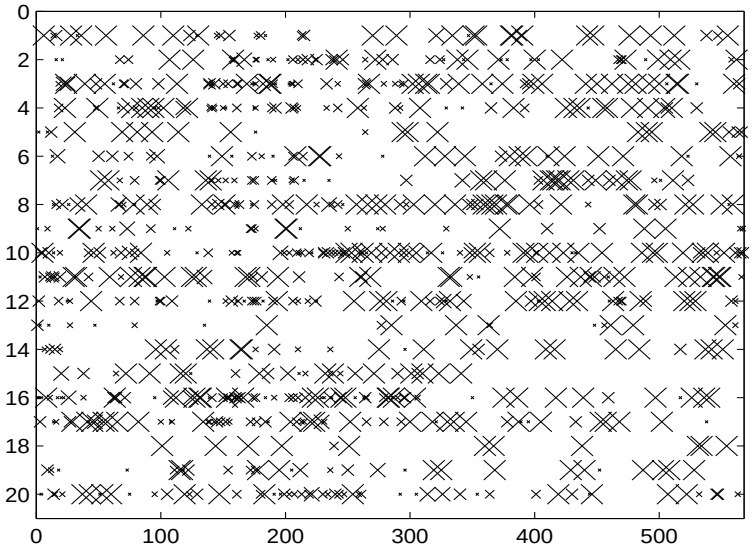

Figure 2.1: Sequence profile of an influenza HA protein.

Often, a conversion method can be used to convert each letter or each sequence into a numerical vector. The length of the numerical vector depends on the conversion

scheme used and often time a trade off exists in the conversion scheme such that all important information of the genetic sequences may not be captured. Here, we list a few conversion schemes that have been applied to convert genetic sequence data to numerical vectors for computation.

k-mer vector

The k-mer vector converts a genetic sequence to a vector by using the k subsequence to describe the data. For a nucleotide sequence, it is converted to a vector of length 4^k where 4 equals the number of letter in the alphabet $[AGCG]$. If $k = 3$, then the vector length is $4^3 = 64$. The k-mer conversion scheme also works on protein sequences with the converted vector length of 20^k. The higher the k value, the longer the vector. However, often this k-mer scheme is applied to the nucleotide sequences. All the genetic sequences converted to numerical vectors can form a matrix of size $M - by - N$ with M being the number of sequences and $N = 2^k$ the features. Standard algorithms can then be applied to this matrix for analysis purposes. A drawback of this conversion scheme is that certain positions on a genetic sequence may carry specific signal or play a functional role in the organism's evolution or life cycle and this information is not captured with the $k - mer$ conversion method. It is because any genetic sequence is broken up into $k - mer$ subsequences and any positional signal is lost during the conversion process.

Codon vector

There are 64 genetic codes and each code defines how sequences of the nucleotide triplets, called codons, translates into amino acid. The codon vector scheme converts a nucleotide sequence into a numerical vector of length 64. Each component of the vector is a codon count of the nucleotide sequence. A M by 64 data matrix where M is the number of rows can then be constructed for analysis. Again, this scheme does not preserve positional information from the sequence and any significant information will be lost during the conversion process. As an example, a visualization of influenza sequence data based on this approach is illustrated in figure 2.2. The visualization was performed using Principal Component Analysis [52] on the data matrix and the top three principal components were selected for visualization purpose. Three different subtypes of influenza virus sequences were used in this example. As can be seen in the figure, distinct clusters can be observed which suggested that the genetic composition or genotype of these three subtypes is different. A small cluster of human H1N1 can be seen off to the lower center below a bigger cluster of the same subtype. This could indicate that the human H1N1 virus had evolved into a different genetic makeup. This example illustrates that the codon vector scheme can capture the genetic fingerprint of three subtypes of influenza viruses even those positional information was not kept.

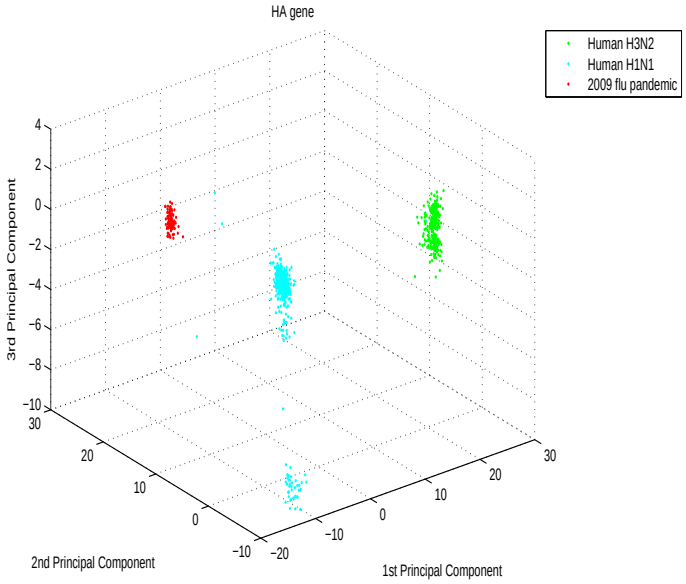

Figure 2.2: Visualization based on codon vector approach.

2.3 Computational methods

Here, we give a brief introduction on a few computational methods that have been developed to study influenza viruses. These methods shared a common theme in that they all utilized techniques from unsupervised machine learning paradigm. Detail for each method is provided in subsequent sections. Smith et. al [16] produced an antigenic map (cartography) showing the antigenic evolution of influenza A (H3N2) virus from its introduction into humans in 1968 to 2003. The map was generated using hemagglutination inhibition (HI) binding data utilizing a classical multidimensional scaling method[52] which revealed clusters of viruses along a chronological path from 1968 to

2003. Using computational method to predict evolution of the virus based on cartography took on a more comprehensive approach when Cai et. al [53] demonstrated that combining multiple HI datasets gave more accurate result which can be used favorably to predict vaccine strain. Their approach can be thought of as a two steps process in which they: (1) reduce the full rank HI data matrix by Singular Value Decomposition (SVD) to give a low-rank approximation of the data and (2) perform multidimensional scaling on the low-rank data to generate the cartography map of the virus. Because of combination of multiple HI datasets, the results were shown to be more accurate both in simulated and real HI data. The cartographic map successfully captures the evolution of the virus and it shows that influenza viruses tended to form clusters along a path. The cluster formation of the virus was first shown by [54] using a simple single-linkage clustering algorithm [52] based on the pairwise Hamming distance of the HA sequences downloaded from the NCBI influenza database [6]. Adapting unsupervised clustering algorithm to study influenza viruses have become more common.

2.4 Cluster Analysis

In 1992, a cluster analysis of 560 influenza H3 HA1 nucleotide sequences from 1968 to 2000 was performed using a single-linkage clustering algorithm by Plotkin et al., [54]. This is the first study using a clustering algorithm to cluster influenza virus sequences into disjoint groups. Their results suggested the existence of a natural scale of non-random aggregation of viral swarms. The viral swarms were closely matched to the time

series of the influenza vaccines recommended by World Health Organization (WHO) for influenza seasons. Through the cluster analysis, they also proposed a simple vaccine selection scheme based on a distance of 6(2) nucleotide(amino acid) changes between clusters. The most recent virus in the current season's most dominant cluster is selected as the vaccine strain for next flu season. Further, they showed that pattern of epitope changes corresponded to cluster jump. Each cluster jump was dominated by mutations on different epitope.

2.5 Influenza Antigenic Distance

The antigenic differences between influenza strains are quantified by the hemagglutination inhibition (HI) assays. These antigenic differences can be used as 'distance' between vaccine strains and current circulating strains when selecting vaccine strains for vaccine production. The antigenic measurement is best described by Ndifon [55] as: The degree to which antisera extracted from individuals infected by one strain (the infecting or homologous strain) prevent another strain (the heterologous strain) from agglutinating red blood cells (heterologous HI titer) is used to measure the antigenic difference between the two virus strains. However, HI titers measured in HI assays can be quite variable due to the fact that HI titers depend on factors, such as: capacity of strains to induce the production of antibodies [56], experimental conditions (temp, pH, etc), and the properties that are not directly related to antigenic difference.

Because of the variability issues in HI assays, sequence based distance measures

have been developed to evaluate the degree of match between the vaccine strains and the dominant circulating strains [57, 58]. The first such distance measure, $P_{epitope}$, considers only the antibody binding sites on the H3 HA surface protein. To use this $P_{epitope}$, one must know the dominant epitope of the HA under consideration. According to [57], a dominant epitope can be identified as the antibody binding site of a circulating strain with the largest fractional change in amino acid sequence relative to the vaccine strain. This measure is almost identical to the p-distance (hamming distance over the total length of the sequence) commonly used in molecular evolution except we are considering a small portion of the sequence to calculate the hamming distance. $P_{epitope} = \frac{h}{N}$ where h is the number of amino acid differences in dominant epitope and N is the total number of amino acids in the dominant epitope. Pan et. al., extended this formulation to $P_{all-epitope} = \frac{\hat{h}}{\hat{N}}$ where $\hat{h}$ is the number of changes in all five epitopes and $\hat{N}$ is the total number of amino acids in all five epitopes.

2.6 Influenza Antigenic Cartography

Influenza antigenic cartography (Figure 2.3) by Smith et .al., a two dimensional representation of the antigenic distance between strains that shows the antigenic variation of the virus across time. This computational technique is also used to help in seasonal influenza vaccine strain selection [59, 60, 61, 62, 16]. The influenza antigenic map produced by Smith et. al.,[16] was based on the multidimensional scaling (MDS) method published in [63]. In [63], the authors provided methods related to related to metric

and ordinal multidimensional scaling algorithms first developed in the mathematical psychology literature, to construct explicit, quantitative coordinates for points in two dimensional space given experimental data such as hemagglutination inhibition assays, or other general affinity assays. The points are the antigens and antibodies and that the coordinates of these points in the two dimensional space represent their physico-chemical properties related to binding. Distances between these points are assumed to be related to their affinity, with small distances corresponding to high affinity. As long as the experimental affinity data (antibodies to antibodies, antigens to antigens, and the affinity of antigens to antibodies) is available from binding assays, a cartography map can be produced. The drawbacks are that the binding assay data can contain missing values and the resolution of the data might not be optimal. In [16], the antigenic maps or cartography were produced based on a functional minimization approach and solved by using conjugate gradient optimization method with multiple random restarts. Briefly, the map is based on minimizing the function: $E(D_{i,j}, d_{i,j}) = \sum_{i,j}(D_{i,j} - d_{i,j})$. D_{ij} is the target distance between antigen i and antiserum j derived from the HI measurement $H_{i,j}$ from the binding assays. d_{ij} is the Euclidean distance between the coordinates of antigen i and antiserum j in the antigenic map. Basically, MDS tries to minimize the difference between the Euclidean distances of all embedded antigen and antiserum pairs in the map and the corresponding HI values [64]. Antigenic map/cartography can also be produced by a linear algebra approach in which we seek the eigenvectors of a symmetric pairwise distance matrix. This pairwise distance matrix can be computed from

the $D_{i,j}$ matrix using any distance functions. The most common distance measure is the Euclidean distance function $d_{i,j} = \sqrt{\sum_1^p (x_p^i - x_p^j)^2}$ where x_i and x_j are two vectors in p dimensions. The results produced by MDS using Euclidean distance measure is the same as Principal Component Analysis (PCA). However, MDS has the flexibility to use difference distance function besides the Euclidean distance function, for example, city-block distance or Mahalanobis distance.

Cai et. al.,[53] developed a computational framework to construct cartography for influenza virus using Singular Value Decomposition (SVD) technique. Their approach utilized SVD method to generate the lower dimensional representation of the datasets and plotting the results on a two dimensional graph. The rationale is that because HI binding data tends to be noisy and often with missing entries in the dataset, a lower representation that can capture the essential signal from the full dataset is usually sufficient for general purpose visualization of the evolutionary trend of the influenza virus.

2.7 Influenza Virus Evolution and Vaccine

Research studies that focus on how vaccination affecting the evolution of influenza virus have been few and scarce. In this section, we will present research studies that have investigated the effects of vaccination have on the evolution of influenza viruses. Although these studies were only done on mice and chicken, but the results are very useful in understanding the evolution of influenza virus from a different point of view. Hensley

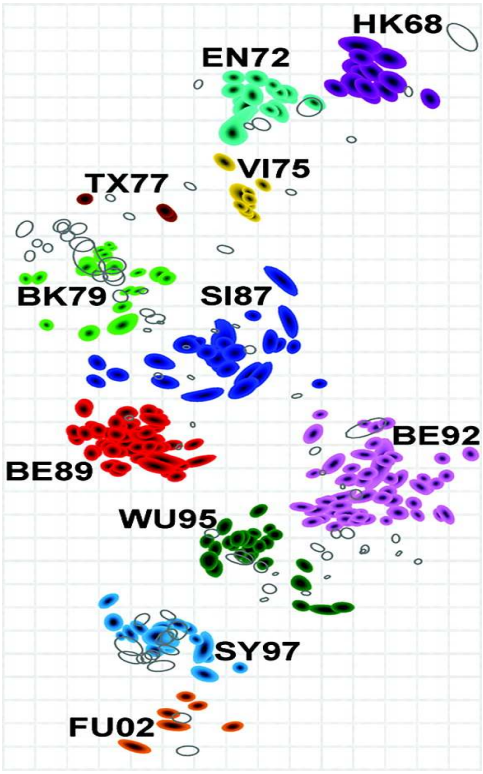

Figure 2.3: Antigenic map or influenza cartography. Generated using Hemagglutinin Inhibition binding data of seasonal A/H3N2 virus. The vertical and horizontal axes both represent antigenic distance and the orientation of the map is free. The spacing between the grid lines is 1 unit of antigenic distance (corresponding to a 2 fold dilution of antiserum in the HI assay. Two units is 4 fold dilution, 3 units is 8 fold). Each clump is designated by a location-year name on the map. The locations are in chronological order: Hong Kong 1968, England 1972, Victoria 1975, Texas 1977, Bangkok 1979, Sichuan 1987, Beijing 1989, 1992, Wuhan 1995, Sydney 1997 and Fujian 2002.

et. al., [65] infected mice with a seasonal influenza virus strain isolated in 1934 from Puerto Rico (A/Puerto Rico/8/1934 H1N1). A group of mice were vaccinated against this virus and developed antibodies against it, while another group were unvaccinated. After the infection of the vaccinated and unvaccinated mice with the 1934 influenza strain, the viruses were passed on to a new set of mice and this process was repeated

nine times. After the ninth passage, sequencing results of the HA surface protein revealed that the unvaccinated mice showed no mutation on the HA protein. In contrast, the HA gene in virus isolated from vaccinated mice had mutated. The mutation helped the virus to become more adherent to the receptors of the host cells. The mutated virus essentially developed a way to shield its hemagglutinin antigenic sites from antibody attack. In the second set of experiments, they infected a new set of unvaccinated mice with the mutant virus emerged in the first series of experiments. Due to the lack of vaccine pressure in these mice, the virus reverted to a 'low-affinity' form. Their results suggested that influenza virus 'reacted' differently to their environment and that vaccine pressure played a significant role in how the virus evolved in a vaccinated environment.

Studies [45, 4] to understand vaccine pressure on the evolution of the influenza virus using phylogenetic analysis have been performed on avian influenza H5 subtype viruses isolated in a Mexico chicken farm. Extended vaccination program was used in Mexico check farms in the early 1990's to mid 2000's and significant antigenic drift of avian influenza viruses has been observed in chickens. Isolates from years 2002-2006 show significant genetic drifts when compared with the vaccine strain. The studies also demonstrated that genetic drifts in the HA gene lineages followed a yearly trend, suggesting gradually cumulative sequence mutations. After vaccine introduction in 1992, multiple sublineages separated from vaccine lineage were detected which suggesting the virus has mutated away from vaccine strain. Continual genetic and antigenic drifts in avian influenza virus have not been detected before and are occurring after the vaccine

introduction in Mexican chickens. These findings suggested that antigenic drift and mutations are likely aided by homologous challenge strains alone.

2.8 Phylogenetic Analysis

Phylogenetic tree construction has always been the universally accepted method of studying the evolution of influenza viruses. Although we did not use this method in our study, a brief introduction would help readers to at least make aware of the utility of this widely used approach. Phylogenetic analysis of influenza viruses is essentially a two dimensional representation of the evolutionary relationships of the viruses with underlying assumptions about the evolution process applied in the analysis. Unlike our approach, we do not assume any evolutionary process before the analysis and hypotheses are only generated after seeing the results from the analysis.

2.8.1 Phylogenetic tree construction

In this section, we provide a brief introduction to phylogenetic tree construction. Phylogenetic tree provides the means to display the evolutionary relationships among species, genes, genomes, or any other entities that share a common ancestor, in a two dimensional graphical format [66]. To represent the relationships from a common ancestor to its descendants, phylogenetic tree construction algorithms build branching patterns with branch length that capture the 'changes' along their imputed evolutionary paths. One must bear in mind that the branching patterns that make up a phylogenetic tree can

rarely be observed directly [66]. Three widely used methods of phylogenetic tree construction are: (1) parsimony, (2) distance based, and (3) maximum likelihood method. Within all these methods, a scoring function is used to score all the possible trees to find the best tree that can arise from the data. In general, what is estimated by phylogenetic tree construction is the amount of evolutionary change between the inner nodes and the leaves of the tree [67]. Phylogenetic studies of the evolution of influenza viruses are mostly performed using the distance based and the maximum likelihood methods. This is because these two methods give more reliable results or predictions when analyzing viruses that are isolated across a long time period and that the variation of evolutionary rate can be taken into account.

2.8.2 Methods

In this section, we briefly describe each phylogenetic tree construction method currently being used in the field of molecular evolution studies.

Parsimony method

Simply put, the parsimony method is based on the idea that the best tree or branching pattern is the one that requires the fewest evolutionary changes to capture the complete evolutionary relationships between the ancestor and its descendants. Given an aligned sequence dataset, this method computes all the scores for all the possible branching patterns that can be generated based on the "simplest is the best idea" from the dataset.

The score is a measure of the number of evolutionary changes or substitutions that would be required from the ancestor to its descendants. The tree that has the lowest score is considered the most likely representation of the true evolutionary history of the sequences in the dataset. This tree is also called the most parsimonious tree. A major drawback of this method is that the number of possible trees increases exponentially as the number of sequences grows. This is a serious drawback by today's standard as next generation sequencing technology can easily generate over millions of sequences. The amount of sequence data that is available for analysis makes this method the least attractive among the three. One way to get around this limitation is to only score the shortest tree and not all the possible trees. To do this, heuristic search is implemented in the algorithm to search through the large tree space to find the shortest tree. In general, parsimony method works well with small number of sequences with strong sequence similarity and can produce reliable results. But as the number of sequences grows, the amount of memory and computational time required can become a significant factor that affect the accuracy of the results. This is because the heuristic search engine used in the algorithm can be trapped in a local minima in the tree space.

Distance based method

The main idea behind the distance based method is to infer evolutionary relationships from the patterns of similarity among organisms [66]. Given a sequence dataset, a distance matrix is generated by computing the pairwise distance between the sequences.

This distance matrix reflects the similarity among all the sequences in terms of changes observed from the data. However, one needs to adjust the distance matrix to correct it to account for the evolutionary events (single substitution, multiple substitutions, coincidental substitutions, parallel substitutions, convergent substitution, and back substitution) based on some pre-defined Markov model that describes the probabilities of each nucleotide change. For example, the Jukes-Cantor Markov model [67] is a commonly used model that can be applied for distance matrix adjustment or correction. Once the distance matrix is corrected, a phylogenetic tree is then constructed based on the distance values of this matrix. Two popular distance based phylogenetic methods are in use today for studying influenza virus. One, the UPGMA (unweighted pair group method with arithmetic mean) method. Two, the neighbor-joining method. UPGMA method assumes the rate of change along the branches is a constant and that the distances are approximately ultrametric, meaning that for three sequences (a,b, and c), $D_{ac} \leq max(D_{ab}, D_{ac})$ where D is the distance measure [67] between two sequences. The constant evolutionary rate assumption implies that all of the leaves in a tree are equidistant from the root of the tree.

The neighbor-joining (NJ) method is widely used in phylogenetic analysis of influenza virus because it allows for evolutionary rate variation on separate branches of the tree. This is especially suitable to analyze difference subtypes of the influenza virus. The NJ method works by first forming the corrected distance matrix, then join the two sequences that give the smallest distance value. This pair is then become a new 'entity'

and the distance matrix is recomputed based on this new entity. This step is repeated until all the entities are joined to form the phylogenetic tree.

Maximum likelihood method

Likelihood and Bayesian methods have been designed to provide a statistical framework for phylogenetic reconstruction [66] but at the expense of computational time. This method is computationally intensive because it considers all possible trees to find the best probable tree that fits the data best. The likelihood score L of a tree is written as $L = Prob(D|Tree)$ where $Tree$ are all the possible trees and D is the dataset given. Maximum likelihood method requires an evolutionary model (e.g. Jukes-Cantor or Kimura model) that estimates the rates of substitution of one base for another in a set of sequence data. Once the model is selected, the probability that the sequence data would be generated given a particular tree can be computed. The best tree is found as the one that has the highest probability of producing the observed sequence data based on the model selected.

2.8.3 Challenges

All phylogenetic tree construction methods will generate a tree or some trees, the issue is then how well does the constructed tree represent the underlying evolutionary relationships of the given data. Approaches for determining how well a particular tree represents the data have been proposed because of this issue. One such approach is the

bootstrapping approach. This approach resample the data repeatedly and reconstruct the phylogenetic tree to see how often the same result is obtained from the resampled dataset. If resampling the data without replacement, it is called the jackknifing technique. Another approach is to compare the trees generated with different methods and determine how similar they are to each other. One can compare parts of the tree with each other and score the number of differences in tree branching.

The success of phylogenetic analysis depends on the multiple sequence alignment algorithm that is applied to the data before trees are constructed. For sequences that have diverged considerably, it is a difficult issues because the multiple sequence alignment may not be optimal and can affect the reliability and topology of the constructed trees[67]. This raises the issue of how to choose a suitable multiple sequence alignment method for the data at hand before the phylogenetic analysis. If the sequences are closely related and there are no gaps or insertion/deletion in the sequence data, then any good multiple sequence alignment algorithms that emphasize on local or global alignment score can be used. If the sequences have high degree of variation, then a global alignment scheme might be more suitable.

Another aspect of challenge in phylogenetic trees construction is that inferring ancestry lineage with limited data can cause bias in interpretating of the result. The origin of the recent 2009 swine H1N1 pandemic outbreak was first inferred by using phylogenetic tree approach with swine influenza sequences which led to the conclusion of the origin of the A/H1N1pdm09 being a swine originated influenza virus [68] based on the

topology of the tree. However, further study by [69] concluded that the A/H1N1pdm09 strain actually came from a human source.

Phylogenetic tree construction or analysis is very useful to elucidate the genetic origins, selection pressures, evolution rates, reassortment histories, and population dynamics of influenza viruses in different host population. Given the rapid increase of the viral sequence data, this important tool faces the challenge of large computational costs that potentially limits its success in producing accurate and robust results.

Chapter 3

Novel Markov Model

3.1 Introduction

In this chapter, we present a compact Markov model that models the evolution of the influenza virus as a sequence of single point mutations, using a two-layered statistical model: a Markov chain for the mutations and a Poisson process for the timing of their occurrence. Modelling the mutations this way captures much of the process because the fixed length of the influenza genome segments mean very few insertions or deletions occur. This model allows us to estimate the probabilities of seeing similar influenza viruses after long time gap. Our working hypothesis is that after a long enough time gap, many site mutations should accumulate in the virus due to a lack of a proofreading function [70], leading to distinct modern variants. Our working assumption is based on the neutral theory of evolution [71] and that each amino acid or nucleotide site is

under a neutral mutation process. We test our hypothesis by combining a standard Poisson process with the Markov model. The Poisson process models the occurrences of mutations in a given time interval, and the Markov model estimates the probabilities of changes to the genetic distances due to mutations. We show that it is highly unlikely that very similar sequences would arise long after the original sequence. Given the observations of several pairs of very similar sequences separated by several decades, our results suggest that there must be some reservoir or evolutionary mechanism that is capable of preserving old virus strains, allowing them to reappear after extended time intervals.

3.2 Compact Markov Model

We model all mutations as the combination of several single point mutations and use a Poisson process to model the mutation rate. The Poisson process naturally admits more complex mutations, treating them as several single point mutations occurring in rapid succession. Then we build a compact Markov model to model the mutations themselves. Markov models have proven to be a powerful tool for phylogenetic inference and hypothesis testing when modeling transitions between amino acid states. Modeling amino acid transitions is complex since proteins are made of twenty amino acids. Because of this, we take a very different approach in building our Markov model. We are trying to avoid a Markov chain where each sequence is a state because this would give rise to an exponentially large number of states (20^n where n is the number of sites). In

our Markov model, we collect into a single state H_k all the protein sequences at given Hamming distance k from a given starting sequence $s_0 \in H_0$. The starting sequence s_0 can be chosen either as the earliest isolated sequence or the most recent one. Our Markov model assigns the probability of an arbitrary HA sequence $s_1 \in H_k$ mutating into a different HA sequence $s_2 \in H_l$ through a single point mutation, where l must be one of $k-1, k, k+1$.

Previous studies [72, 73] have shown that to better fit the model, conserved sites should be excluded in the analysis under the neutral theory framework. Here we have taken the same approach where we have limited the mutations captured by our Markov chain to the HA1 domain consisting of $n = 329$ sites, since this region is less conserved than the HA2 region [74, 75]. Therefore, our Markov model has only $n+1 = 330$ states instead of the 20^n states it would have if we kept each state and each possible transition separate.

Formally, consider a finite set of states labeled $\{H_0, ..., H_n\}$. In order to keep the Markov chain to a manageable size, we group all the sequences within Hamming distance of k from a start sequence into a single "super state" H_k. At each transition, we assume a single point mutation occurs, and that this mutation of amino acid replacement exhibits uniform rate of evolution throughout long periods of evolutionary time [76]. This assumption is particularly consistent with the concept of "molecular evolutionary clock" and is central to the neutral theory [77, 78, 79]. Because of the high rate at which RNA viruses evolve, it has been observed that these sequences show the typical

pattern of neutral evolution [78].

We denote by a the size of the alphabet of amino acids, in our case 20. For a sequence $s_1 \in H_k$, there is a probability k/n that the mutation occurs in one of the k positions where s_1 differs from s_0, and if this change occurs, there is a $1/(a-1)$ chance that the new amino acid in this position will match that in the same position of s_0. Hence the probability x_k of a transition from H_k to H_{k-1} is $x_k = \frac{k}{n} \cdot \frac{1}{a-1}$. Similar reasoning yields the probability y_k that a transition will remain at the same Hamming distance: $y_k = \frac{k}{n} \cdot \frac{a-2}{a-1}$. The probability that a mutation will be in one of the $n-k$ sites that still match s_0 is $z_k = 1 - \frac{k}{n}$, corresponding to a transition from H_k to H_{k+1}. The resulting probabilities x_k, y_k, z_k are assembled into a Markov transition matrix M as shown below. The entries in each row of M add up to 1.

$$\begin{pmatrix} 0 & 1 & & & & \\ x_1 & y_1 & z_1 & & & \\ & \ddots & \ddots & \ddots & \ddots & \\ 0 & & x_{n-1} & y_{n-1} & z_{n-1} & \\ & & & x_n & y_n & \end{pmatrix} = \frac{1}{n} \begin{pmatrix} 0 & n & & & & \\ \frac{1}{a-1} & \frac{a-2}{a-1} & n-1 & & & \\ & \ddots & \ddots & \ddots & \ddots & \\ 0 & & \frac{n-1}{a-1} & n-1\frac{a-2}{a-1} & \frac{1}{n} & \\ & & & \frac{1}{a-1} & \frac{a-2}{a-1} & \end{pmatrix}$$

Using this model, we can compute the probability q_t that a virus will have a Hamming distance at most κ from the initial source sequence after t mutations. We give the general form of how to compute the above probability. We let $v_t = (v_{t0}, v_{t1}, \ldots, v_{tn})$ be the row vector of probabilities of being in state $H_0, H_1, \ldots, H_n$, respectively, after t

mutations. At $t = 0$ we are in state H_0 consisting of just the initial sequence. This is represented by the row vector $v_0 = (1, 0, 0, ..., 0)$. Then the vector of probabilities after $t + 1$ mutations is related to the probabilities after t mutations by $v_{t+1} = v_t * M$. The probability of being at most distance κ from s_0 after t mutations is the sum of the first $k + 1$ components of v_t: $q_t(k) = \sum_{i=0}^{k} v_{ti}$.

The above analysis counts events consisting of a single mutation. The mutation rate is modeled by a Poisson process [80, 81]. This includes the possibility that no mutation or several mutations take place in a given time interval, assuming all sites undergo the same substitution rate. This assumes that the probability of a mutation in a given time interval depends only on the length of the interval but is independent of the behavior outside the time interval. If λ is the average number of mutations in a time interval of 1 year, then the probability that t mutations occur in any time interval of length Y years is given by $p_t(Y) = \frac{(Y\lambda)^t}{t!} e^{-Y\lambda}$. The Poisson process models when mutations occur, and the Markov model models the nature of the mutations. Combining these two models yields the probability $P_\kappa(Y)$ that after Y years a sequence would appear with a genetic distance κ from s_0 of κ, namely $P_\kappa(Y) = \sum_{t=0}^{\infty} p_t(Y) \cdot q_t(\kappa)$.

3.3 Hemagglutinin Sequence Data

The HA protein is the major surface antigen of the influenza virus. Its role is to bind to host cell receptors promoting fusion between the virion envelope and the host cell [82]. Influenza A virus HA genes have been classified into 16 subtypes (H1-H16) according to

their antigenic properties. This HA protein is cleaved into two peptide chains HA1 and HA2 respectively when matured [22]. The HA2 chain has been found to vary less and is more conserved compared to HA1 chain [83]. The HA1 chain is 329 residues long and is the immunogenic part of HA protein. Past studies have shown that HA1 is undergoing continual diversifying change [72, 75] and is the most variable portion of the influenza genome[84].

Using the NCBI Influenza database available online, we have collected 3439 influenza virus type A protein sequences deposited before December, 2007 (excluding identical sequences and lab strains/NIAID FLU project). This collection of protein sequences contains isolates from around the globe and from a diverse range of hosts. We used protein sequences because they were known to give more reliable results than nucleotide sequences when constructing evolutionary history [22]. Each of the 3439 sequences has a unique annotation which contains the host organism, the strain number, the year of isolation, subtype, and protein name. We aligned all sequences to a consensus sequence using the NCBI alignment tool. According to the study presented by [84], a uniform consensus strain tends to circulate for some time, since the mutations that occur during replication do not become fixed in the early stages of circulating. The aligned sequence data were then used with a genetic distance function to determine the pairwise genetic distance (including gaps) of the sequences.

The genetic distance between two sequences can be thought of as the **edit distance**, which is the number of single letter changes needed to transform one sequence to the

other. This yields a simple scoring function assigning a zero to a matching amino acid base and a one to a mismatch. The sum of all mismatches is usually called the Hamming distance (k) or Hamming score for the pairwise sequence comparison. For comparison of very similar biological sequences, this Hamming distance can be used under the assumption that the observed difference between a pair of sites represents one mutation [85]. The present study could also be carried out using BLAST or any alignment algorithm, but as considerably greater expense. In [74], Hamming distance was successfully used to find interesting clusters of IV HA sequences and to predict vaccine strains with good results. Hamming distance as genetic distance between viruses has also been used effectively in modeling influenza viruses [86]. In our study, we compute the Hamming distance based on a consensus alignment to account for the small number of insertions and deletions. We then store the pairwise Hamming distance scores of HA gene in a pairwise affinity matrix and identify virus sequence pairs sharing high sequence similarity (at least 90 percent) but separated by a long time gap.

3.4 Results

Strain	H	Y	EG	$\mathcal{P}$-value
AAD17229: A/South Carolina/1/1918	0	0	0	source sequence
AAA91616: A/swine/St-Hyacinthe/148/1990(H1N1)	20	72	47.3	6.3499e-06

Table 3.1: H1N1 subtype long time gap strains (Rate: 2×10^{-3} per site per year). H = Hamming distance, Y = Year, EG = Expected number of mutations.

Figure 3.1: Seasonal human influenza H1N1 virus pairwise Hamming distance computed using the oldest strain (A/North Carolina/1918) as the source to every other isolates in the dataset. Toward year 2000's, most of the H1N1 have evolved away from the source strain.

We first identified viruses having very close genetic distance but with large time gap. Figure 3.1 shows the H1 subtype HA1 domain pairwise sequence genetic distance plotted against time of isolation in year. The genetic distance corresponds to the Hamming distance including gaps. Tables 1 and 2 show viruses sharing very high sequence similarity but with large time gap. We used the amino acid substitution rate of $r = 2{\times}10^{-3}$ per site per year for H1 and H2 subtype viruses, estimated using the entire region of the HA gene and assuming that the molecular clock is followed [22] throughout evolutionary history. This yields an annual mutation rate of $\lambda = nr = 329{\cdot}2{\times}10^{-3} = 0.658$. We give two examples of unlikely similarities over long time gaps in table 3.1 and 3.4. Each table includes the accession number "Accession", strain name "Strain", the Hamming distance "H"

Strain	H	Y	EG	$\mathcal{P}$-value
AAY28987: A/Human/ Canada/720/2005(H2N2)	0	0	0	source sequence
AAA64365: A/RI/5+/ 1957(H2N2)	6	48	31.5	7.807e-09
AAA64363: A/RI/5-/ 1957(H2N2)	3	48	31.5	1.206e-11
AAA64366: A/Singapore /1/1957(H2N2)	5	48	31.5	1.155e-09
AAA43185:A/Human/ Japan/305/1957(H2N2)	5	48	31.5	1.155e-09

Table 3.2: H2 subtype long time gap strains

(calculated from the first strain), expected number of mutations "EG", the year difference "Y", and the $\mathcal{P}$-value, the probability that this Hamming distance (or less) would be observed after the given time interval as predicted by our model. Using the pandemic strain A/South Carolina/1/1918 and A/swine/St-Hyacinthe/148/1990(H1N1) from Table 3.1, the interpretation of the result is that after 72 years, the expected number of mutations is 47.3 and the probability of being within a Hamming distance of 20 of the original source sequence is 6.35×10^{-6}. A very recent published research study [79] employing the state-of-the-art Bayesian Markov chain Monte Carlo [87] which allows for substitution rate variation and maximum likelihood phylogenetic methods indicates that this A/swine/St-Hyacinthe/148/1990(H1N1) virus is a contaminant from the A/swine/1930 strain. The genetic distance of the pandemic strain to the A/swine/1930 strain is 22. The genetic distance of A/swine/1930 to A/swine/St-Hyacinthe/148/1990(H1N1) is only 3 indicating that these two strains are virtually identical. From table 2, we see that A/Human/Canada/720/2005(H2N2) strain isolated in 2005 is exceptionally similar to the two asian pandemic strains A/Singapore/1/1957(H2N2) and A/Human/Japan/305/

1957(H2N2) in terms of the genetic distance. These two pandemic strains were human transmissible and currently no influenza vaccines contained the H2N2 virus [88]. This reappearance of the highly pathogenic H2N2 virus could cause a potential pandemic as current population is not immunized against this strain of virus. The origin of the A/Human/Canada/720/2005(H2N2) strain was traced back to human error at a laboratory distributing virus samples for training purposes and the distributed strains were quickly destroyed at all receiving laboratories [88].

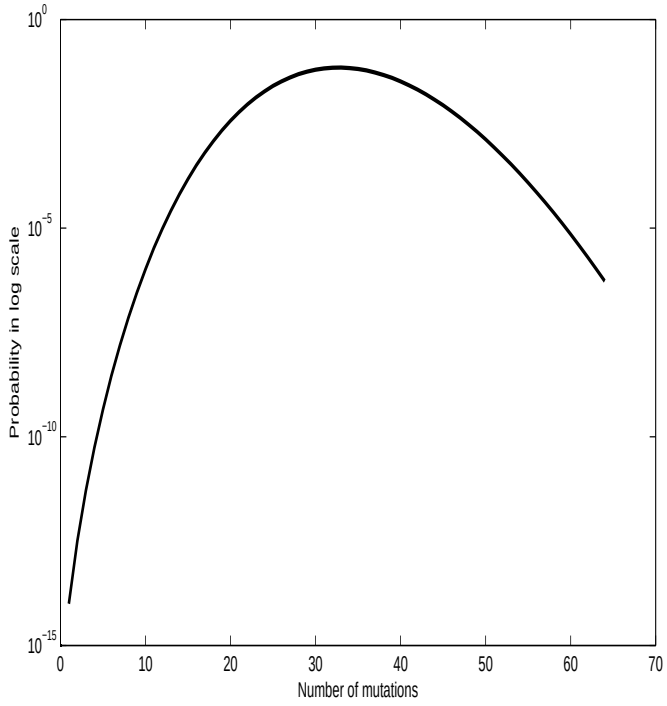

Figure 3.2: Poisson process distribution plot

To check how our model matches the data, we show the predicted distribution of

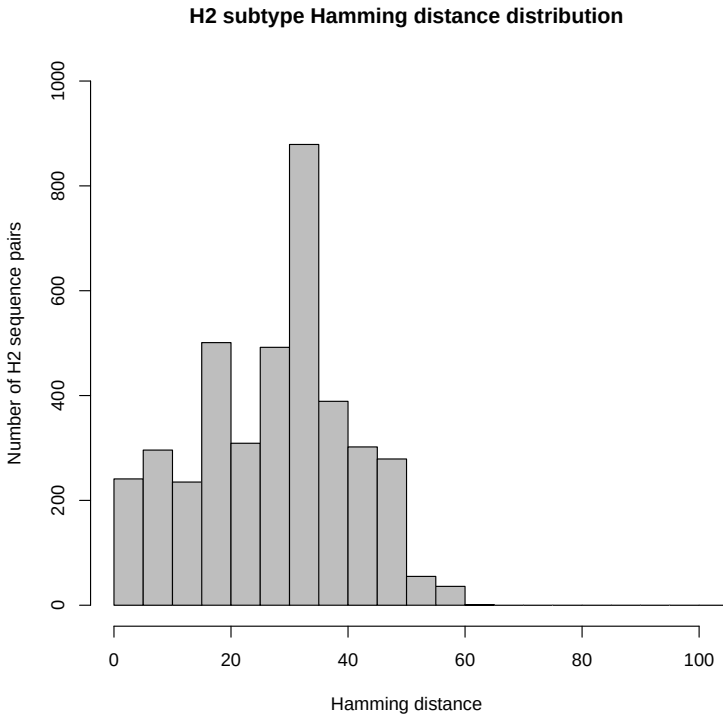

Figure 3.3: H2 subtype histogram plot

Hamming distances in Figure 3.2 based on a time interval of $Y = 49$ and annual mutation rate of $nr = 0.658$ for the H2 subtype. The peak of the curve indicates that with high probability, roughly 30-40 mutation events would have taken place. This tells us that we should expect to see the majority of H2 sequence pairs with Hamming distances in the vicinity of 40 given the length of time interval equals 49 years base on Poisson process assumption. We compare this to the actual distribution of Hamming distances found in the H2 subtype data shown in Figure 3.3 over the range of data available (from 1957 through 2006 or a span of 49 years). Figure 3.3 shows that the majority of the H2 sequence pairs have Hamming distances around 35, which matches the Poisson process prediction. Figure 3.5 illustrates how the probability values of 3 H2 strains in Table 3.2 are rapidly dropping against the expected number of mutations from the Markov model calculation. Figure 3.4 shows the predicted distribution within the time interval of 70-85 years from the combined Poisson process and Markov chain model using H1 subtype HA1 sequences. The curve shows that with high probability most sequences should be in states H_{60} to H_{70}. This reflects what is observed in figure 3.1 and figure 3.6 where most sequences have Hamming distance around 60-70. This suggests that our model is able to capture the overall evolutionary behavior of the influenza virus according to a molecular clock, leading to a natural increase in the genetic distance as time passes, consistent with [77].

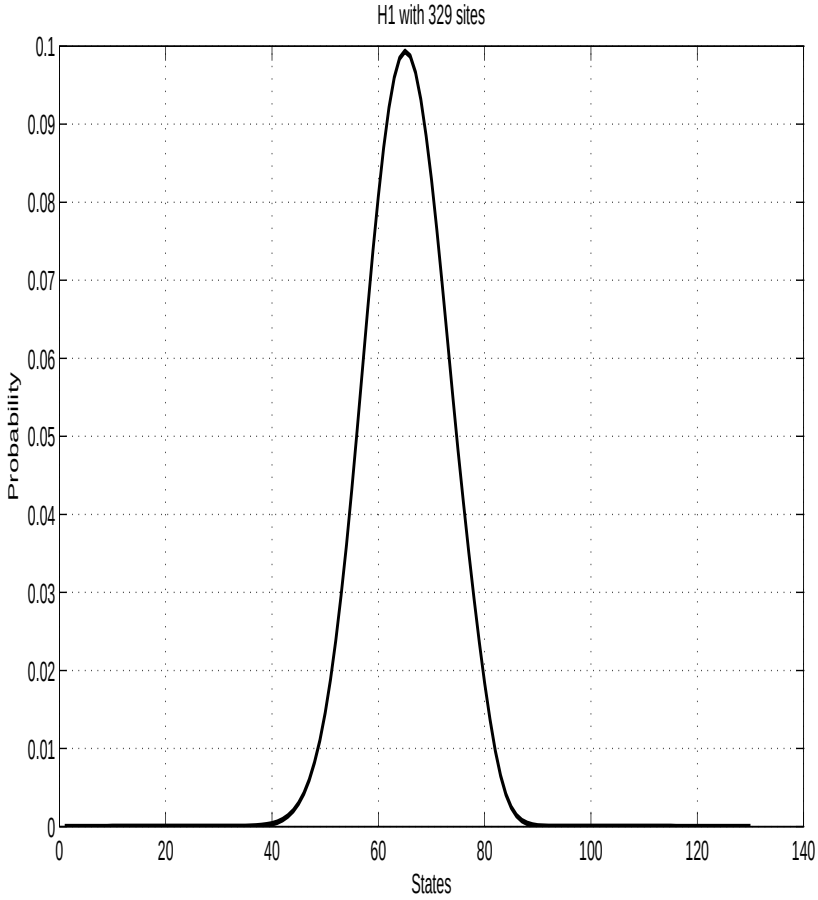

Figure 3.4: Model prediction plot

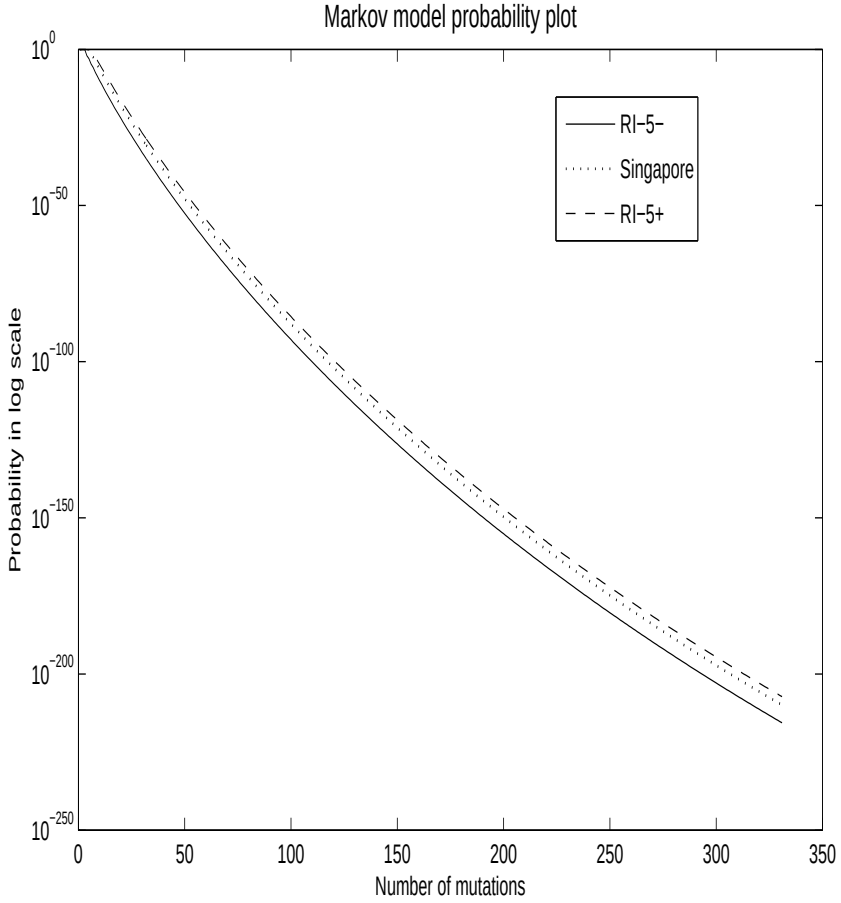

Figure 3.5: H2 strains probability plot

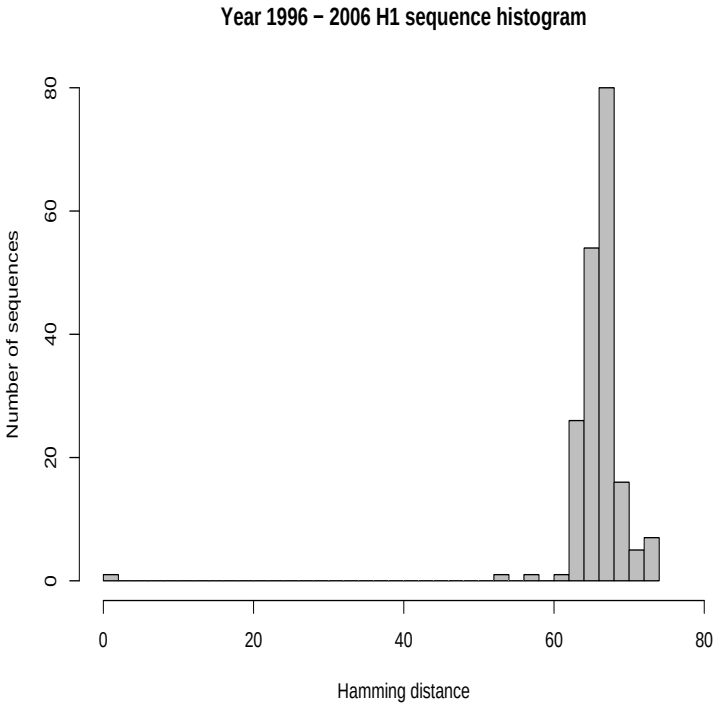

Figure 3.6: Histogram of H1 from 1996-2006

3.5 Conclusions

The extensive genetic diversity of influenza A viruses through genetic drift and reassortment in the past century has resulted in many new strains being produced. However, H1, H2, and H3 subtypes strains have displayed cyclic behavior resulting in influenza pandemics [89]. In the present study, we applied neutral evolution theory to influenza virus HA protein sequences to investigate the evolutionary dynamics of the virus. We did not include the other mutational changes (compensatory and transition/transversion) as our aim is to model the influenza strictly under a neutral evolution process. The model can be extended to include other types of mutational changes but it will require to model each site individually. Using the combination of a Poisson model with a novel Markov model, we were able to calculate the probability values of finding a very similar sequence composition separated by a large time gap. We have so far been able to identify several anomalies due to laboratory artifacts or human error. This finding is promising since we have yet to apply it in a full scale comprehensive analysis of all 16 subtypes of the virus. However, judging by the extremely low probability values obtained for some observed sample strains, we conclude that there may be one or more sources of various strains of the virus in which they are preserved over long time periods. The existence of reservoirs preserving viruses for decades cannot be completely eliminated.

Chapter 4

Influenza Evolution Analysis with Binary Encoding Approach

4.1 Introduction

In this chapter, we present the computational approach that is used to study the evolution of influenza viruses. This approach is based on encoding nucleotides and amino acids using a fixed length binary code. The application of this approach to difference influenza virus datasets is presented in this chapter and subsequent chapters in this thesis.

4.2 Background

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 [90, 81, 91, 92, 93, 94]. 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 that can help to generate new hypothesis. In this chapter, 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 influenza virus in the past century can be visualized. The unsupervised machine learning approach consists of three steps: (1) genetic sequence conversion by binary encoding, (2) dimensional reduction and scatter matrix computation, and (3) visualization. Genetic sequence conversion is a data preprocessing step where biological sequence data is converted into numerical values for analysis. Dimensional reduction is carried out by the unsupervised machine learning method called Principal Component Analysis (PCA). The results from PCA are directly used in the scatter matrix computation in order to quantify the evolution paths of the influenza virus. The final visualization step is the visualization of the projected data on the leading components of the PCA. This step offers us insight and alternative perspective into this rapid evolving

antigen in difference hosts and environments.

4.3 Genetic Sequence Conversion

For nucleotide sequences, we encode Adenine (A) to "0001", Guanine (G) to "0100", Cytosine (C) to "0010" and Thymine (T) to "0001" [95, 96]. Each nucleotide base is uniquely represented by a 4 digits binary string. For example, to encode a nucleotide sequence of "AGA" and another of "ACA", AGA is transformed to 0 0 0 1 0 1 0 0 0 0 0 1 and ACA is transformed to 0 0 0 1 0 0 1 0 0 0 0 1. When these two sequences are compared, the mutation in the second position is captured by the different between 0100 and 0010. This encoding scheme allows for direct capture of mutation information between sequences and facilitate direct subsequent computational analysis. For protein sequences, we convert each amino acid to a binary string of length twenty and each string is different by only one bit. For example, Alanine is coded as "1 0 0 0...0 0 0" and Cysteine is coded as "0100...000". In addition, the biophysical properties data of each amino acid can be directly append to the end of the twenty bits string. For example, the hydrophobicity value of Alanine is 1.8 and the binary string of Alanine becomes "1 0 0 0 ... 0 0 0 1.8" which further distinguishes the differences between each amino acid. Even though the length of the nucleotide sequence has been increased by a factor of 4 and protein sequence by a factor of 20, the sparse data can be stored efficiently (low memory requirements) and there are algorithms to process them efficiently.

4.3.1 Incorporating amino acid biophysical information

The binary encoding scheme with the inclusion of amino acids' biophysical properties leads to substantially better results in distinguishing different subtype when protein sequences are used. The biophysical property we have used in this study is the hydrophobicity property of amino acids. Ray [97] carried out a study to determine the most suitable biophysical properties to use with unsupervised classifiers [97] and found that three properties: Volume, Hydrophobicity, and Isoelectric property are best suited for classification purposes. In our study, we have tried all three of the said properties and found that hydrophobicity is best suited for influenza sequences. We demonstrate this result by applying our coding scheme combined with hydrophobicity values (H-value) on H3 and H5 subtypes genetic sequences. We obtained the hydrophobicity values for all the amino acids published from the study conducted by Ray and Kepler [97]. After appending each H-value to the binary string of each amino acid and converted all the protein H3 and H5 sequences into binary strings, PCA (see 4.4.2) was used to provide visualization (Figure 4.1 and Figure 4.2) between the two subtypes on two dimensional plane. For comparison purpose, we produced a projection of H3 and H5 sequence without using the H-value, as shown in Figure 4.1. Although we see data separation in both cases, the projection result with H-value applied clearly explained more variance (at 70 percent) than the one without (at upper 30 percent). The separation between H3 and H5 also has become more pronounced with less overlapping strains from each subtype.

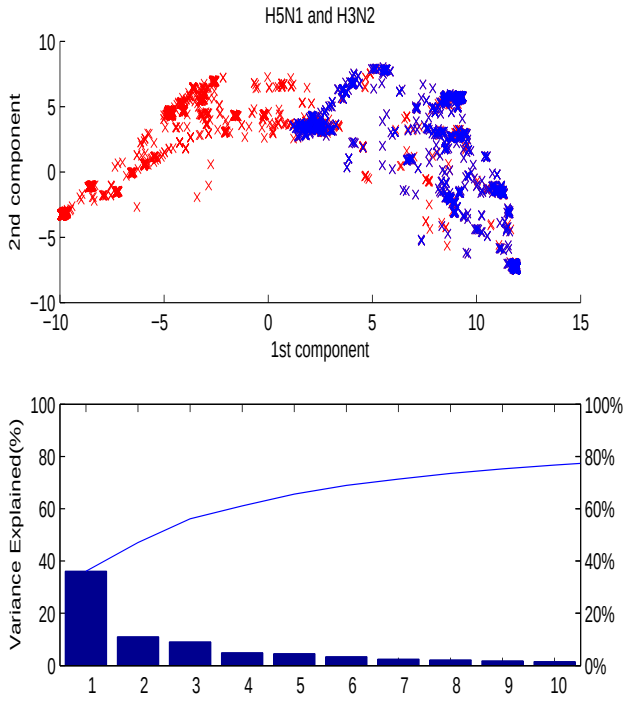

Figure 4.1: Top: PCA projection of H5N1 (red) and H3N2 (blue) protein sequences without applying hydrophobicity information. Bottom: The variance captured by the projection method without applying hydrophobicity information.

4.4 Principal Component Analysis

The dimension reduction step mentioned in section 4.1 is performed by using Principal Component Analysis. Principal Component Analysis (PCA) was described in [98, 99] as finding "the best fitting straight line to a points coincides in direction with the maximum axis of the correlation". It is quite 'old' [100] but remains one of the most used techniques today in data analysis. PCA is a statistical analytical tool that is widely used in data exploration and pattern recognition. The general idea of PCA is that it

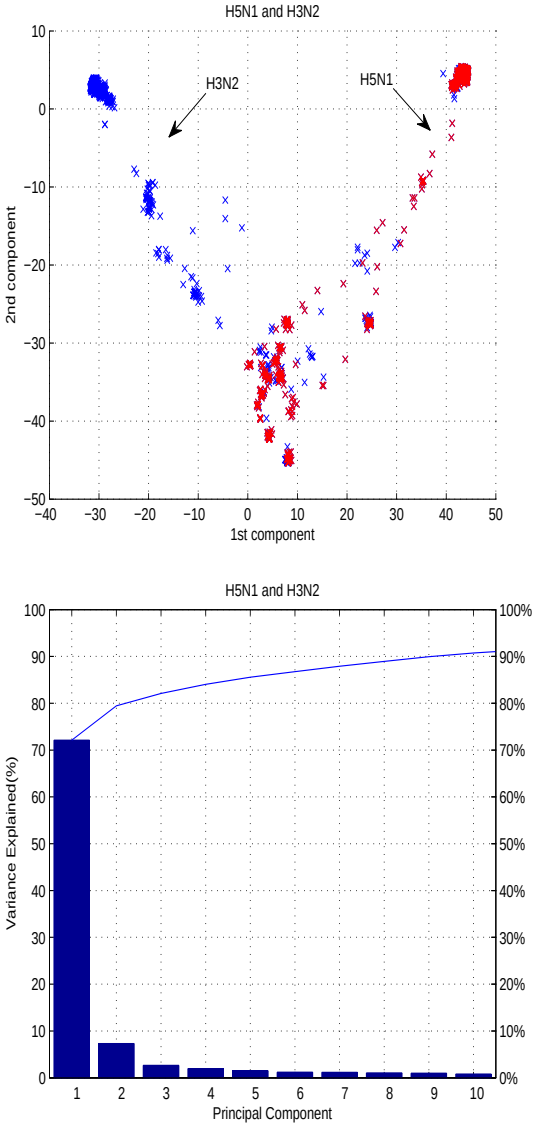

Figure 4.2: PCA projection of H3 (blue) and H5 (red) protein sequences with hydrophobicity information incorporated.

transforms large number of correlated variables into a smaller number of uncorrelated variables while retaining maximal amount of variation. This transformation process

can identify patterns in data that highlight their similarities and differences, thus making PCA a very powerful tool in data analysis. PCA has been used extensively for dimensional reduction in analyzing high dimensional dataset. [52, 101, 102, 103]. The dimension reduction is a linear projection technique that projects high dimensional input vectors into low dimensional ones whose components are uncorrelated [104, 105]. The linearity refers to the (1) change of basis when performing projection as the new basis are linear independent of each other and (2) the 'new' variables (PC's) are the linear combination of the original variables. This projection is a global orthogonal projection on the complete dataset. PCA can also be viewed as an unsupervised machine learning method because it ignores all the data labels and only relies on the attributes of the data. The objective of PCA is to minimize loss of information in the data while representing each data sample with a reduced set of attribute values. When applying PCA to analyze a dataset, it is often possible to capture a large percentage of the total variance with only a few principal components. This is because each principal component captures the maximum proportion of the total variance successively.

4.4.1 Limitations of PCA

Principal Component Analysis has its limitations [106] and these limitations are not restricted to specific data being analyzed using PCA. The limitations can be outlined as follows:

- The interpretation of each principal component (PC's) is very difficult. This is

because each PC is a linear combination of the original variables.

- PCA only performs orthogonal transformations of the original variables. Some data might require a nonlinear transformation or mappings in order to reveal special structure in the data.
- The directions with largest variance are assumed to be of most interest.
- If the original variables were uncorrelated, PCA can only order them according to their variance.
- Discarded PCs are not always 'useless' and it is not clear on how many PCs to retain/discard when performing dimension reduction. For visualization purposes, the leading two or three PCs are used.
- PCA can be sensitive to outliers since it is based on the sample covariance matrix which is sensitive to outliers.
- PCA is not scale invariant as eigenvectors are not scale invariant.
- When all measurements are positively correlated, the first principal component is often some kind of average of the measurements [106].

4.4.2 Application of PCA to Influenza Data

Principal Component Analysis (PCA) is used in all forms of analysis from gene expression data analysis [107, 108, 109] to computer vision. It is a simple non-parametric

method of extracting relevant information from unstructured data sets. The extraction can be viewed as dimensional reduction where a complex high dimension data set is reduced to a lower dimension in order to reveal hidden, simplified structure buried within the data. In order to find the best lower dimension to capture the structure of the high dimensional data, PCA proceeds by diagonalizing the covariance matrix of the data set, consistent with the goal to maximize the variance captured in the projected data onto the lower dimensions. One property is that PCA requires the directions of projection be orthogonal to each other and the variance associated with each direction be maximized. The orthogonal requirement makes PCA solvable with highly efficient linear algebra decomposition techniques. Here, we introduce the working mechanic of PCA from a linear algebra perspective. Consider a data matrix $X_{m,n}$ with dimensions of m by n with m being the number of strains and n being the number of sites. Each row of X corresponds to a strain of virus and each column of X corresponds to a particular site. We first need to center the rows of the data matrix X (i.e. replace X with $X - \frac{1}{m}ee^T X$, where e is a column vector of all ones) and then obtain the covariance matrix C from X by $C = \frac{1}{(m-1)}X^T X$. C is a square symmetric $m \times m$ matrix whose diagonal entries are the variances of the individual strains across sites and the off-diagonal terms are the covariances between different strains. If one wishes to reduce the row dimensions, one can simply apply this entire computation to the transpose of the data matrix. The goal of PCA is to find a set of orthonormal axes that diagonalizes matrix C. The diagonalization of C is computed by finding its eigenvectors. Since C is symmetric and square,

its eigenvectors are the orthonormal principal directions, and its eigenvalues correspond to the variances of the data along those principal directions. The eigenvectors of C are now the new basis for the data X. The projection of the data matrix X onto this new basis gives the alternative "PCA view" of the data with mean zero and variance maximized along each principal component direction. A quick decomposition technique to obtain the orthonormal basis is using the Singular Value Decomposition (SVD), see 4.4.3 and 4.4.4. One can center the matrix, calculate the C matrix, and then applying SVD to C. SVD of C gives $C = U\Sigma V^T$ where the matrix V contains the orthonormal basis we sought. We can project the data to the new basis with $X * V$; The matrix Σ is a diagonal matrix that contains the eigenvalues of C which are the variances of the orthonormal basis/principal components.

Once the transformation is made,we the select the leading two or three components for visualization of the genetic sequence data. In order to better understand the distance relationship as 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: $\|s-t\|^2_{proj} \leq \|s-t\|^2_{bin\,2} = \|s-t\|_{bin\,1} = 2\|s-t\|_H$.

To illustrate this distance relationship, we computed the pairwise distance of the oldest strain (A/Hong Kong/68 1968) to every other strains in the A/H3N2 dataset in both the reduced PCA 2 dimensional space and in full sequence space (Figure 4.3 and Figure 4.4). The pairwise distance in the reduced 2 dimensional space is in high agreement with the pairwise distance computed in full sequence space as indicated by the Pearson correlation coefficient of 0.9792.

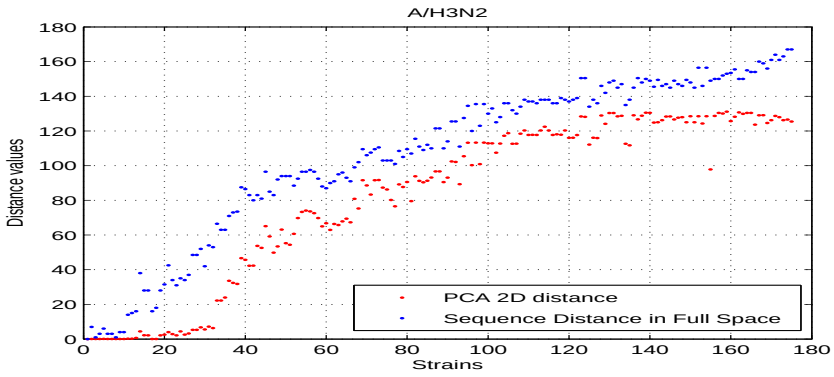

Figure 4.3: Pairwise distance comparison of influenza A/H3N2 virus in PCA 2 dimensional space and full sequence space. The pairwise distance is measured using the oldest strain (A/Hong Kong/68 1968) as the source to every other strains in the dataset. X-axis represents A/H3N2 strains and y-axis represents distance values.

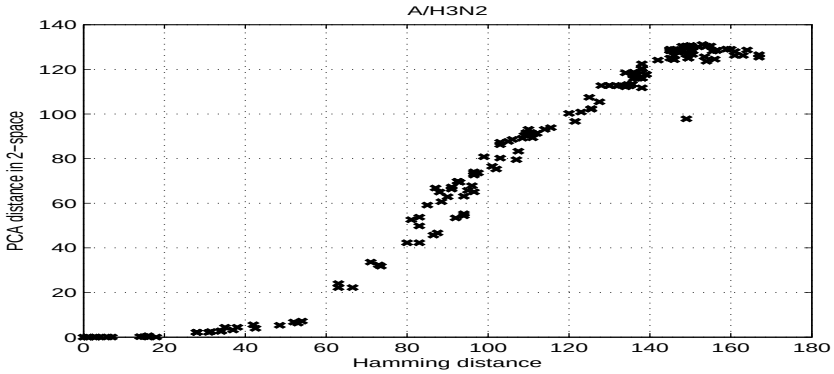

Figure 4.4: Pairwise distance comparison of influenza A/H3N2 virus in PCA 2 dimensional space and full sequence space. X-axis represents Hamming distance and y-axis represents PCA 2D space distance value. Pairwise distance is measured by using the oldest strain (A/Hong Kong/68 1968) as the source to every other strains in the dataset. The correlation coefficient between the pairwise Hamming distance and the pairwise PCA 2D distance is 0.9792.

4.4.3 Singular Value Decomposition

Given $A \in R^{m \times n}$, in rank k, a singular value decomposition (SVD) of A (Figure 4.25) is a factorization $A = U\Sigma V^T$ where $U \in R^{m \times m}$ is orthogonal, $V \in R^{n \times n}$ is orthogonal, and $\Sigma \in R^{m \times n}$ is diagonal. The diagonal elements in Σ must be non-negative and ordered in decreasing order; that is $\sigma_1 \geq \sigma_2 \geq \sigma_3 \cdots \geq 0$. These σ values are called the singular values of A and the number of non-zero σ indicates the rank k of A. The dimension of Σ is the same as A even when A is not a square matrix. The matrix U has the left singular vectors of A and matrix V has the right singular vectors of A. The singular decomposition of A can be computed by finding the eigenvectors of $A^T A$ and AA^T. $A^T A = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T \Sigma V^T$. The $U^T U = I$ and $\Sigma^T \Sigma$ gives $\sigma_1^2, \sigma_2^2 \cdots$ which are the eigenvalues of the symmetric matrix $A^T A$. The

singular values can be found by taking the square root of $\sigma_1^2, \sigma_2^2 \cdots$. Next, we can compute the eigenvector of $A^T A$ to find $v's$ and make them the unit vectors. The $v's$ are perpendicular because eigenvectors of every symmetric matrix are perpendicular. To construct the matrix U, a simple way is to multiply $v's$ by A: $u_1 = Av_1$ and make it into unit vector and collect all the $u's$ into U. The arrangement of $v's$ and $u's$ should follow the order of the eigenvalues [110, 111].

Mathematically speaking, the above mentioned method can become unstable when it is used to compute the SVD of a matrix [112]. An alternative and stable method to reduce the SVD to an eigenvalue problem is to first transform A to a $2m \times 2m$ hermitian matrix H as in $H = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$ and then compute its eigenvalue decomposition [112].

4.4.4 Low-Rank Approximations

The SVD approach gives an optimal low rank approximation to data matrix A [110]. The best rank one approximation to A is the matrix $\sigma_1 u_1 v_1^T$. It is the largest singular value σ_1 and the left and right singular vectors u_1 and v_1. Matrix A can be represented as a sum of rank-one matrices: $A = \sum_j^r \sigma_j u_j v_j^T$. The partial sum captures as much information of the matrix A as possible [113, 114, 115, 116, 112, 117]. In many applications, it is very useful to approximate A with a low rank matrix and the connection with PCA is directly realized by the SVD decomposition of the symmetric covariance matrix constructed from A. As mentioned above, the eigenvectors of AA^T are simply

the left singular vectors of A and the eigenvectors of A^TA are the right singular vectors of A. The SVD provides a complete approach as it gives both sets of eigenvectors when analyzing data in terms of either rows or columns depending on applications. One can 'zero out' some small singular values to produce a low rank approximation of the original data matrix. For visualization purposes, the singular values give the 'explained variance' of the matrix A. The explained variance is the ratio to the total amount of variance in the projected data to that in the original [110]. In other words, it is a comparison of the variance in the approximation to that in the original data [118]. It can also be interpreted as $\frac{\|\hat{A}\|_F^2}{\|A\|_F^2} = \frac{\sigma_1^2+\cdots+\sigma_r^2}{\sigma_1^2+\cdots+\sigma_n^2}$ where $\hat{A}$ is the low rank approximation of A [119].

4.5 Clumpiness measure

The 'clumpiness' or separateness between each clusters observed in the PCA plots is quantified by computing a class separateness value λ. The λ value is computed for all cases (vaccinated samples and nonvaccinated samples) in order to determine the cohesiveness of strains within each year as observed in the vaccinated samples. Although this clumpiness measure does not account for the directionality observed in the PCA plots, it is a first step in quantifying the difference between vaccinated and nonvaccinated samples. The clumpiness measure is a two step process involving a scatter matrix computation and a class separateness measure by simulation of class label randomization.

4.5.1 Multi-class scatter computation

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 **C**. These are just the scalar scatter values: sum of squared distances between points and their respective centers (Figure 4.5). The 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 [120].

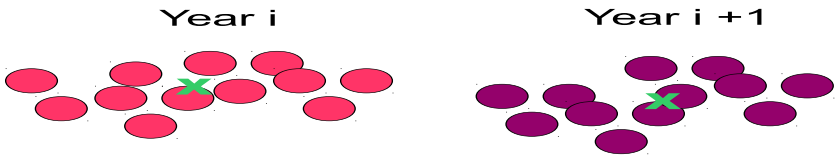

Figure 4.5: Class scatter visualization in PCA 2D space. Green 'x' is the center of the cluster. Red cluster contains viruses from year i and maroon cluster contains viruses from year $i + 1$.

4.5.2 Class separateness measure

From the visualization results of the vaccinated samples, strains tend to cluster with the vaccine seed strain and that each cluster contains viruses isolated within the same year. Follow this observation, we perform a class labels (virus isolation year) randomization in order to determine if the cohesiveness of viruses in a vaccinated sample could have been happened by chance. Under the neutral evolution assumption, the class labels observed could have been generated by chance since the occurrence of mutations is random under this assumption. Once the λ_o is computed for all cases (vaccinated samples and nonvaccinated samples), we randomized the class labels and recomputed the class separateness measure value λ in order to find the probability of observing the observed λ_o. This computation was carried out using the Algorithm I (Box 2). The $K1$ and $K2$ parameters are set to 10,000 and 1000 respectively. The observed λ_o for vaccinated samples was below rounding error of 10^{-16} which made the computation of p-value not possible. Therefore, we resorted to reporting the distance of observed λ_o from the mean $\bar{D}$ in the form of $\bar{D} + / - \hat{D}$.

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.

Box 2:

Algorithm I

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 L

    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 observed λ_o from the mean in the form of $\bar{D} + / - \hat{D}$

4.6 Materials: influenza genetic sequence data

The human-host seasonal influenza A/H1N1, A/H3N2, influenza B virus and avian H5 (from Mexico) sequences are the vaccine controlled sample data and the avian influenza H5 and human H5N1 sequences are the non-vaccine controlled or wild type samples in the study. The avian H5 sequences from Mexico are from vaccinated chickens in Mexico. The chickens were vaccinated against the avian H5 from 1994 to 2002 [45]. Table 4.1 lists the human and avian samples with year range for the viruses in each sample.

The genetic evolution of vaccine controlled seasonal influenza virus was quantified by distance in terms of standard deviation and visualized from their introduction into humans and compared to the genetic evolution of non-vaccine controlled influenza viruses evolving in the wild. We used avian H5 and Human H5N1 viruses because they are currently not being vaccinated against and can act as the 'control' in this present study. We included the human H5N1 virus as the 'control' since this subtype is not currently being vaccinated against but is under active research due to its high mortality rate in infected humans. All influenza hemagglutinin (HA) nucleotide sequences were downloaded from the NCBI Influenza Database [6]. For each dataset (A/H1N1, A/H3N2, type B, avian H5 (Mexico), and avian H5 (non-vaccine controlled) and human H5N1), sequences with gaps or wildcard characters were removed before the analysis. The number of sequences in each dataset is listed in Table 4.1. We have focused on the HA1 domain of the HA protein because it is the most variable region of the entire HA and this HA1 domain (987 nucleotides) also contains antibody binding sites (epitopes)[39]. Although data sample bias does exist in flu sequence databases due to curation and lack of common agreement in sequence upload standard. The viruses selected for HA1 sequencing have been shown to be a representative subset of the total sample including both the dominant variants circulating during the flu season and the outliers [121].

Table 4.1: **Human and avian datasets**

Samples	Year	Seqs
Human A/H1N1	1918-13	2140
Human A/H3N2	1968-09	175(235)
Human Type B (Vic/Yam)	1970-13	818
Human H5N1	1997-12	127(128)
Avian H5 (Mexico)	1994-02	32
Avian H5 (China)	1997-02	32

4.7 Results

In this section, we present the results by using genetic sequences alone of the vaccine controlled seasonal human influenza A/H1N1, A/H3N2, influenza type B viruses, and avian H5 samples and non-vaccine controlled avian influenza H5, and human H5N1 virus samples.

4.7.1 Seasonal human influenza H3N2 virus

Seasonal influenza A/H3N2 which has the highest number of vaccine updates among the three vaccine controlled influenza viruses, we observed that A/H3N2 viruses clustered around vaccine seed strains chronologically since their introduction into humans in 1968 (Figure 4.6). As the time progresses, genetic distance between early strains and late strains has been gradually increasing and the later isolated viruses appeared to have evolved away from older strains (as shown in Figure 4.6 and Figure 4.7). In Figure 4.6, each vaccine strain is marked by an arrow along the evolutionary path. In Figure 4.7, the z-axis represents pairwise Hamming distance (in full sequence space) computed using the oldest strain (A/Aichi/2/1968) as the source strain to every other strains in

the dataset. The gradual genetic drift from the earliest strain to the contemporary strains formed a 'horse shoes' shape in the plot. One can see cluster separation between antigenically distinct clusters across the evolutionary path. This is because each cluster contains a vaccine strain and that each vaccine strain is used against a specific antigenic variant of the virus. The narrow band of the evolution path also suggests that the mutations on the HA gene are most likely concentrated on the antigenic region of the HA protein. It is believed that the mutations within the antigenic binding regions of the HA gene changes the shape of the HA protein in order for the virus to escape antibodies binding. If the accumulated mutations were to occur in random positions on the HA gene since 1968, a narrow and directionally restricted evolutionary path would not have been captured by the PCA algorithm, rather, the PCA plot would have shown multiple large clusters scattered within the plot. The observed evolutionary path formed by tight chronological clusters suggests that the fast turnover of the genetic diversity of the virus.

The Class-separateness analysis of the A/H3N2 indicated that the observed class separateness value λ_o is at 30.5 which is far from the mean of the distribution generated by using Alg I. The distance in terms of standard deviation is at $978.3 \pm .031$ as shown in Figure 4.8. Figure 4.9 shows the 'zoom-in' of the blue line at the left end of Figure 4.8. This distribution of $\lambda_{i=1...K1}$ is only one instant from the inner loop of Alg I. The $K1$ parameter was set at 10000 and K2 was set to 1000.

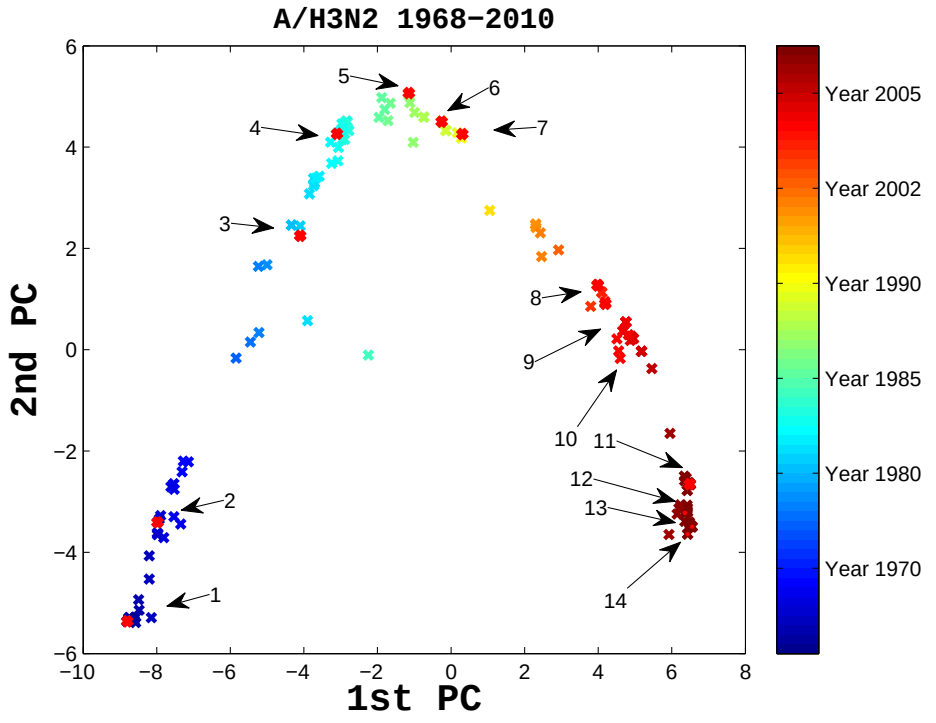

Figure 4.6: Two dimensional visualization of the evolution of seasonal human influenza A/H3N2 virus evolution. Each arrow points to a vaccine strain (red dot) and each vaccine strain corresponds to each vaccine update. The horizontal and vertical axes represent the first and second principal component respectively. The color bar indicates the isolation year of the virus from year 1968 (blue) to year 2010 (red).

4.7.2 Seasonal human Type B influenza virus

Seasonal influenza B virus was first isolated in 1940 and has diverged into two antigenically and genetically distinct lineages since the 1980s [122]. This virus has been actively evolving since the divergence as seen by the number of vaccine updates (16) since its introduction to the human population. The two lineages are represented by the reference strains B/Victoria/2/87 and B/Yamagata/16/88 respectively [123, 124].

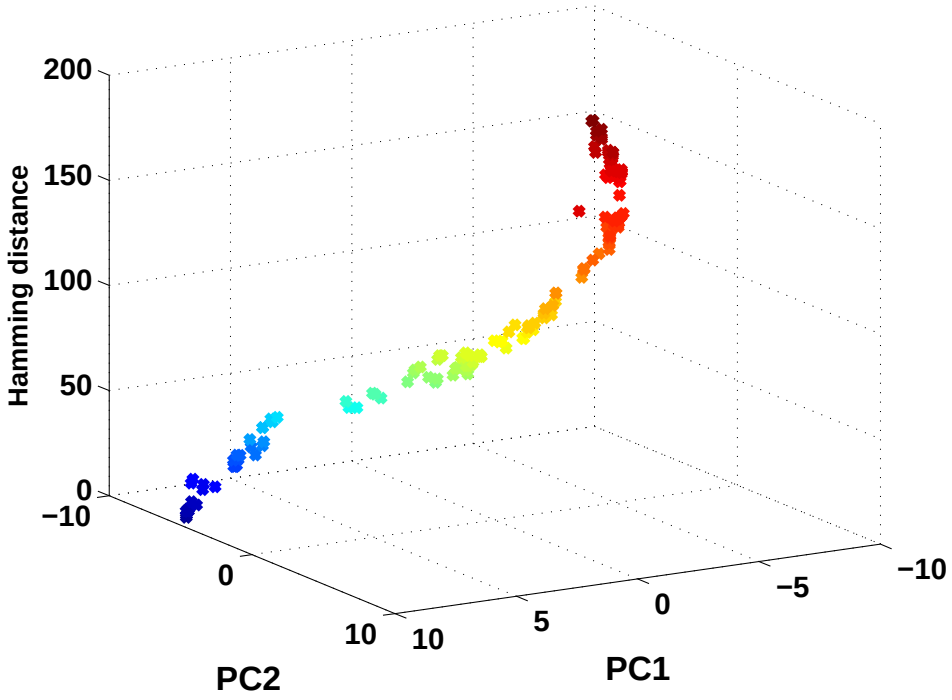

Figure 4.7: A three dimensional visualization of seasonal human influenza A/H3N2 virus evolution. Z-axis represents pairwise Hamming distance (in full space) computed using the oldest strain (A/Hong Kong/68 1968) as the source to every other strains in the dataset. A gradual increase in the genetic distance suggests that new strains are evolving away from the older ones as time passes.

The B/Victoria lineage predominated during the 1980s while the B/Yamagata lineage predominated in the most part of the world during the 1990s [125]. Since then, these two lineages have co-circulated in human population ever since. The visualization of the evolution of Type B virus using HA gene sequences from 1970 to 2012 are shown in Figure 4.10 and Figure 4.11 respectively. The very first observation from these two figures is that the evolution pattern of Type B virus is very similar to the seasonal human influenza A/H3N2 virus. Specifically, there are distinct directional evolution trends

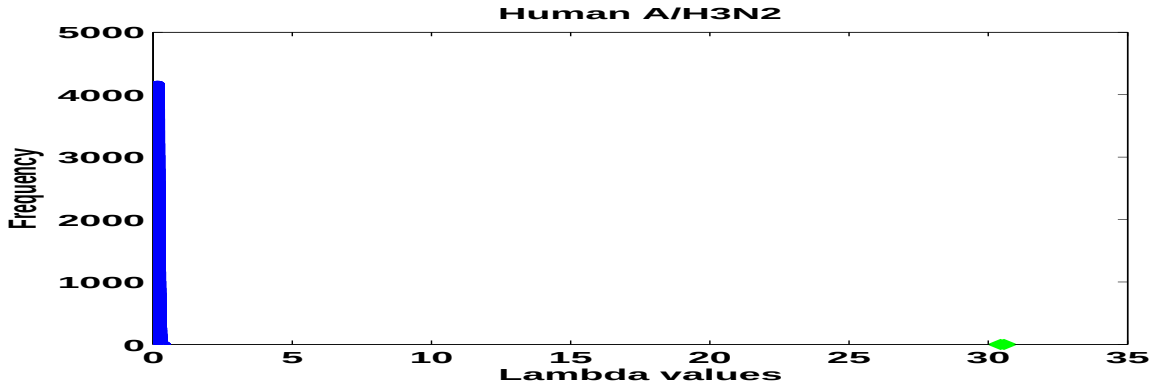

Figure 4.8: Human A/H3N2 class label randomization simulation result. The green dot represents the λ_o value (observed class separateness value). The blue line on the far left of this figure represent the distribution of $\lambda_{i=1...K1}$.

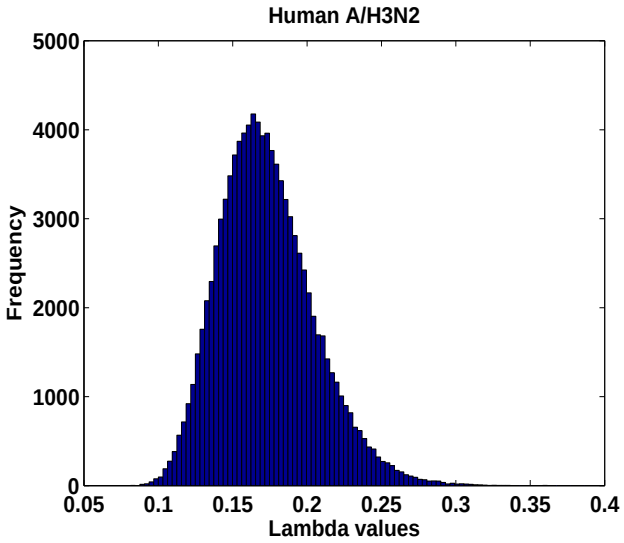

Figure 4.9: The 'zoom in' of the blue line at the left end of Fig. 4.8 which shows the distribution of $\lambda_{i=1...K1}$ from one instant of the $k1$ loop within the Alg I. The histogram is generated using 100 bins.

for each lineage originating from the time when lineage split. Viruses in each lineage gradually evolving away from the oldest strain. The chronological patterns can also be

marked by using the vaccine seed strains as in the case of seasonal human A/H3N2 virus. In Figure 4.11, the z-axis represents pairwise Hamming distance (in full space) computed using the oldest strain (B/Osaka/70) as the source to every other strains in the dataset. The Yamagata lineage (red) and Victoria lineage (blue) with vaccine updates as marked by the virus isolation year and the magenta color represents vaccine updates before the lineage split.

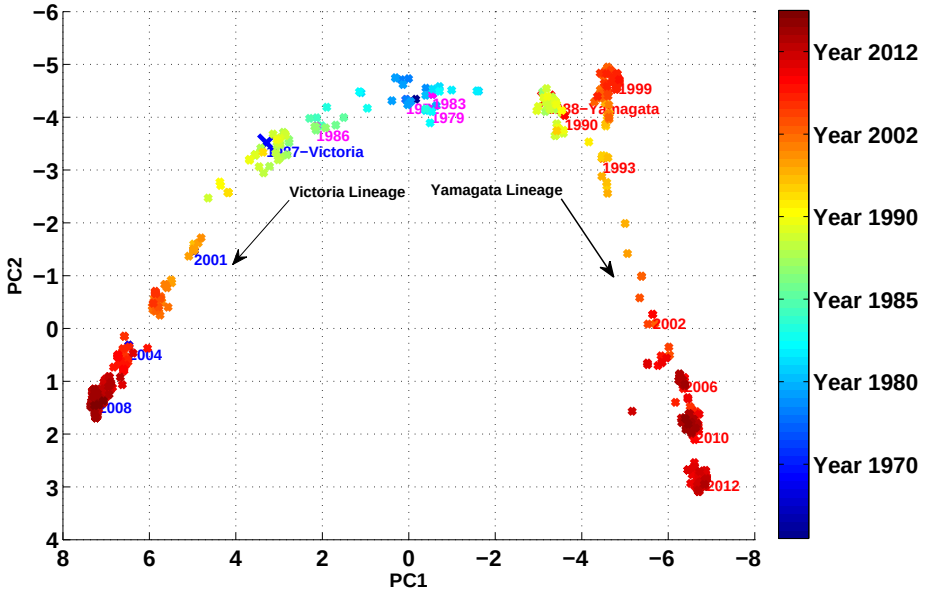

Figure 4.10: Two dimensional visualization of the seasonal Human Type B influenza virus evolution. Two lineages Yamagata (red) and Victoria (blue) have been co-circulating in humans since 1986. Each vaccine update is represented by the 'isolation year' of the virus from each lineage and it is marked on the evolutionary paths for both lineages. Magenta color represents vaccine updates before the lineage split. The horizontal and vertical axes represent the first and second principal component respectively.

The Class-separateness analysis of the Type B influenza indicated that the observed

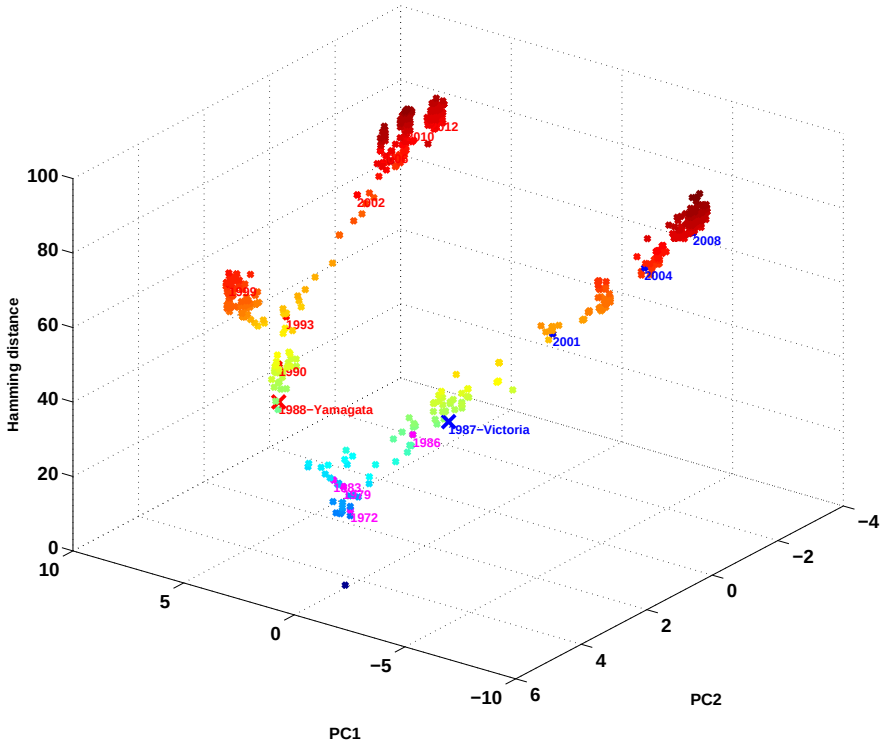

Figure 4.11: Three dimensional visualization of the seasonal human Type B influenza virus evolution. Z-axis represents pairwise Hamming distance (in full space) computed using the oldest strain (B/Osaka/70) as the source to every other strains in the dataset. The Yamagata lineage (red) and Victoria lineage (blue) with vaccine updates are marked by the virus isolation year. Magenta color represents vaccine updates before the lineage split.

class separateness value λ_o is at 26.3 for Victoria lineage and 25.3 for Yamagata lineage (Figures 4.14 and 4.12). They are both far from the mean of the distribution generated by using Alg I. The distance in terms of standard deviation is at $1310 \pm .02$ for Victoria lineage and $1327.8 \pm .019$ for the Yamagata lineage. Figure 4.13 and Figure 4.15 shows the 'zoom in' of the distribution represents by the thick blue line in Figures 4.14 and

4.12 respectively. This distribution of $\lambda_{i=1...K1}$ is only one instant from the inner loop of Alg I. The $K1$ parameter was set at 10000 and K2 was set to 1000.

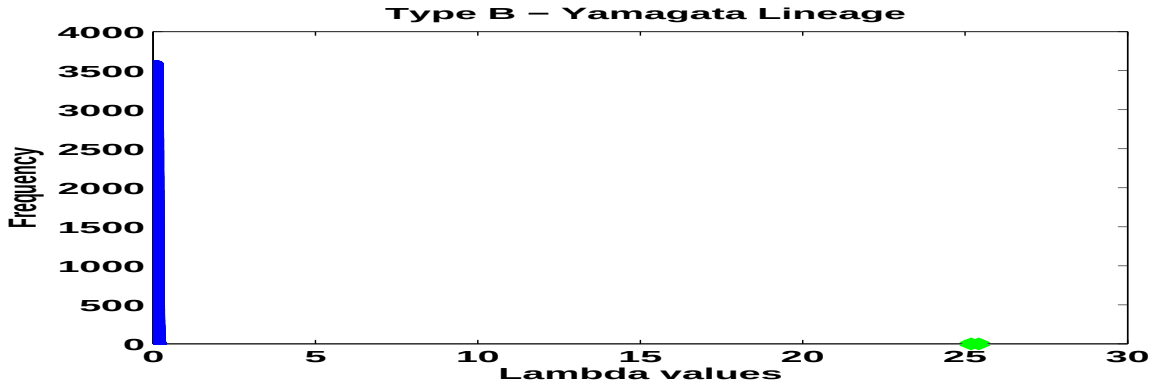

Figure 4.12: Type B (Yamagata) class label randomization simulation result. The green dot represents the λ_o value (observed class separateness value). The blue line on the far left of this figure represent the distribution of $\lambda_{i=1...K1}$.

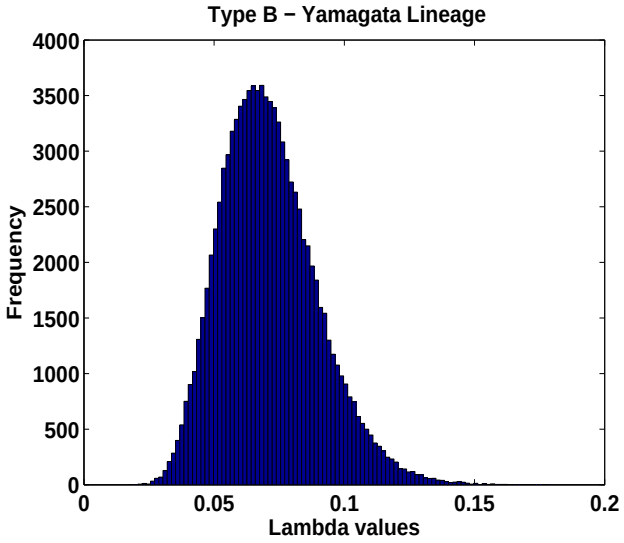

Figure 4.13: The 'zoom in' of the 'blue line' at the left end of Fig.4.12 which shows the distribution of $\lambda_{i=1...K1}$ from one instant of the $k1$ loop within the Alg I. Histogram is generated using 100 bins.

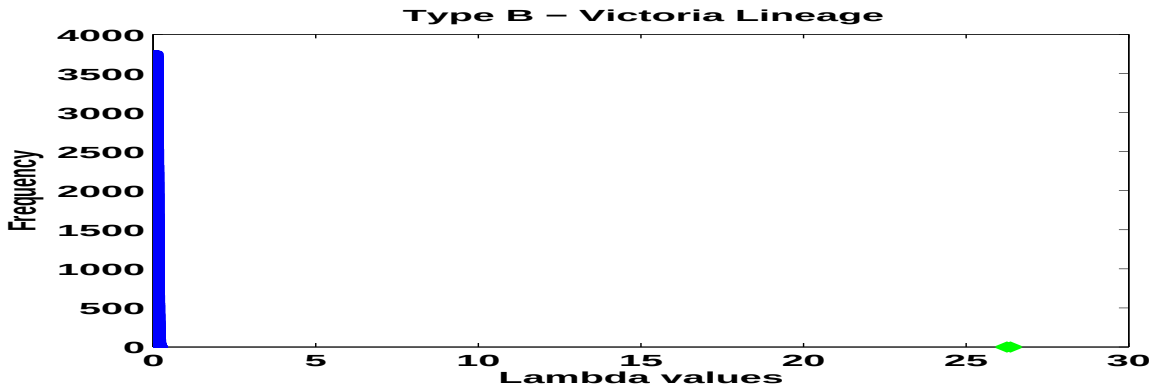

Figure 4.14: Type B (Victoria) class label randomization simulation result. The green dot represents the λ_o value (observed class separateness value). The blue line on the far left of this figure represent the distribution of $\lambda_{i=1...K1}$.

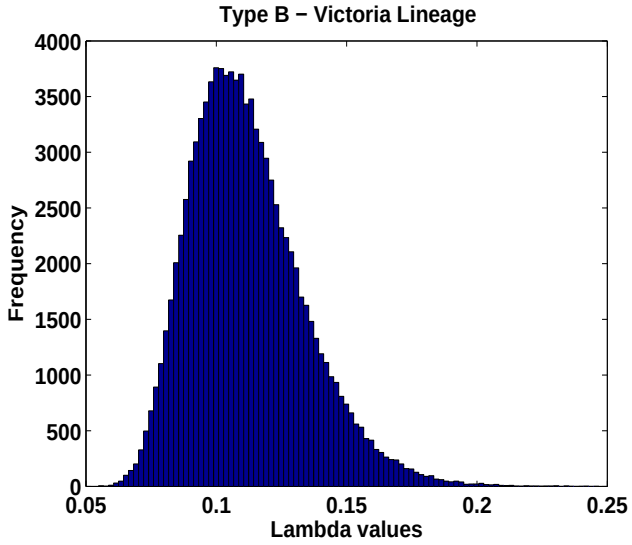

Figure 4.15: The 'zoom in' of the 'blue line' at the left end of Fig.4.14 which shows the distribution of $\lambda_{i=1...K1}$ from one instant of the $k1$ loop within the Alg I. Histogram is generated using 100 bins.

4.7.3 Seasonal human influenza A/H1N1 virus

Seasonal human influenza A/H1N1 virus has been circulating in humans since 1918 after the Spanish flu pandemic outbreak [32, 33]. A new A/H1N1pdm09 pandemic strain appeared in 2009 and subsequently began to circulate in humans and replaced the classical A/H1N1 strain as the dominant circulating seasonal A/H1N1 subtype. The evolution visualization of the classical A/H1N1 with A/H1N1pdm09 since its introduction to humans is presented in Figure 4.16. HA gene sequences from 1918 to 2013 were used to produce the PCA plot. First, there is a separation between viruses before and after the 2009 pandemic. Also, the classical A/H1N1 viruses appeared to be following a narrow directional evolutionary path. Second, the emergence of two evolutionary paths after the 2009 pandemic can be observed. This is very similar to the Type B influenza virus in which a lineage is split into two lineages. However, there has not been a vaccine update since 2009 on the A/H1N1 component ever since the pandemic A/H1N1pdm09 strain overtook the classical A/H1N1 as the dominant strain. This suggests that this emergence of separated paths observed in the figure is most likely due to the genetic diversity within the the post-2009 strains [126]. In addition, study by Klein et al. has shown that a fourfold increase of the mean hamming distance in the coding region of the HA gene between strains from 2009 to 2013 [126] and a twofold increase in the non-coding region of the HA gene. This is consistent of what is observed in the Figure 4.16 where evolutionary paths are progressively moving outward from the 2009 cluster, suggesting the virus is actively evolving. A probable reason to support this observation

is that the emergence of antigenic drift variants in the future is likely. Besides these distinct observations, one can see that the evolution of A/H1N1 is similar to other seasonal human influenza viruses in that there are: (1) narrow band of evolutionary paths and (2) the newer strains are gradually evolving away from the older strains, and (3) chronological patterns can be observed.

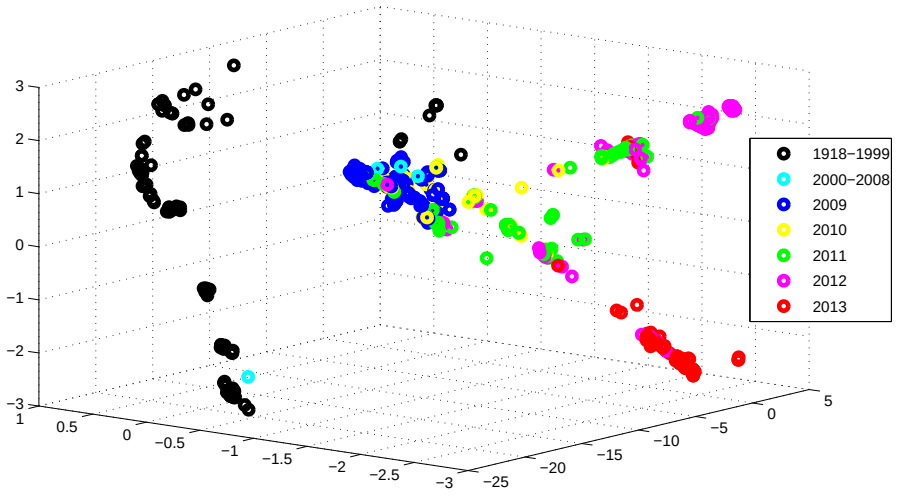

Figure 4.16: Three dimensional visualization of influenza A/H1N1 evolution. The x, y, and z axises are the 1st, 2nd, and 3rd principal components respectively.

The Class-separateness analysis of the A/H1N1 indicated that the observed class separateness value λ_o is at 24.7 which is far from the mean of the distribution generated by using Alg I. The distance in terms of standard deviation is at $617.2 \pm .04$ as shown in Figure 4.17. Figure 4.18 shows the 'zoom in' of the distribution represents by the thick blue line in Figure 4.17. This distribution of $\lambda_{i=1...K1}$ is only one instant from the

inner loop of Alg I. The $K1$ parameter was set at 10000 and K2 was set to 1000.

Figure 4.17: A/H1N1 class label randomization simulation result. The green dot represents the λ_o value (observed class separateness value). The blue line on the far left of this figure represent the distribution of $\lambda_{i=1...K1}$.

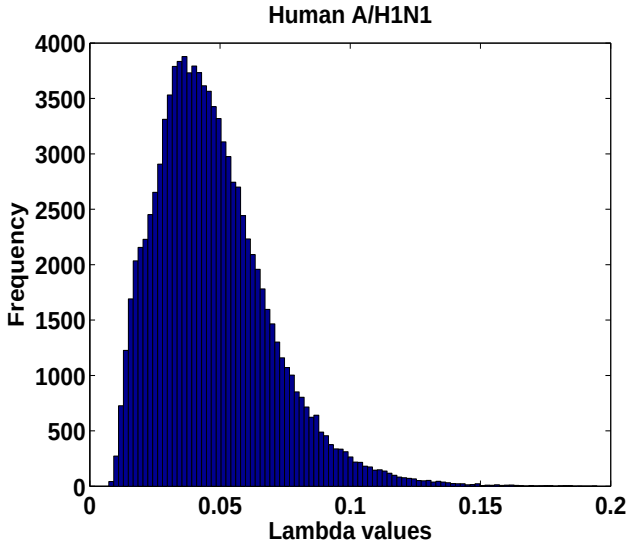

Figure 4.18: The 'zoom in' of the 'blue line' at the left end of Fig. 4.17 which shows the distribution of $\lambda_{i=1...K1}$ from one instant of the $k1$ loop within the Alg I. Histogram is generated using 100 bins.

4.7.4 Human influenza H5N1 virus

The sporadic infections caused by the avian H5N1 in humans has raised concern of a potential pandemic outbreak of the H5N1 influenza virus. The avian H5N1 subtype is thought to originated from wild bird species since its emergence in 1996 [127, 128]. Since then, this subtype has been classified into many clades depending on its pathogenicity. The Highly Pathogenic Avian Influenza A (HPAI) group of H5N1 can cause fatal infection in humans and the Low Pathogenic Avian Influenza Virus (LPAI) group does not cause fatal infection. The HPAI group has been endemic in bird species and its ecology and antigenic properties have led to a higher diver virus strains in endemic areas [129]. As of now, vaccination strategy has not been applied to fight against this subtype in human population. Figure 4.19 presents the evolution of this virus from 1997 to 2012. Figure 4.19 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 [130].

The Class-separateness analysis of the non-vaccine controlled H5N1 indicated that the observed class separateness value λ_o is at 1.01 which is not far from the mean of the distribution generated by using Alg I. The distance in terms of standard deviation is at $34.8 \pm .029$ as shown in Figure 4.20. The distribution of $\lambda_{i=1\ldots K1}$ is from one instant of the inner loop of Alg I. The $K1$ parameter was set at 10000 and K2 was set to 1000.

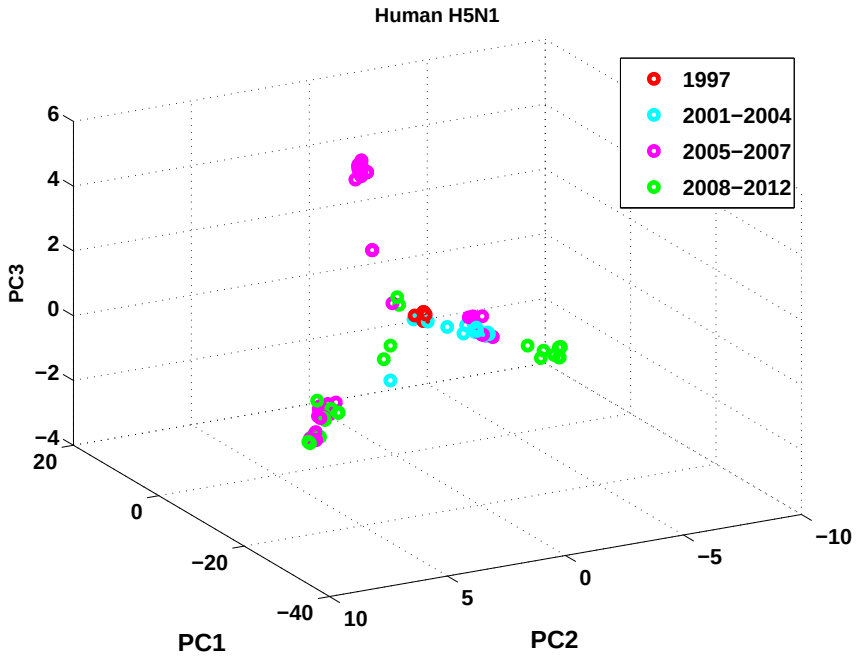

Figure 4.19: Three dimensional visualization of human H5N1 influenza virus evolution. The PC1, PC2, and PC3 represent the first, second, and third principal component respectively.

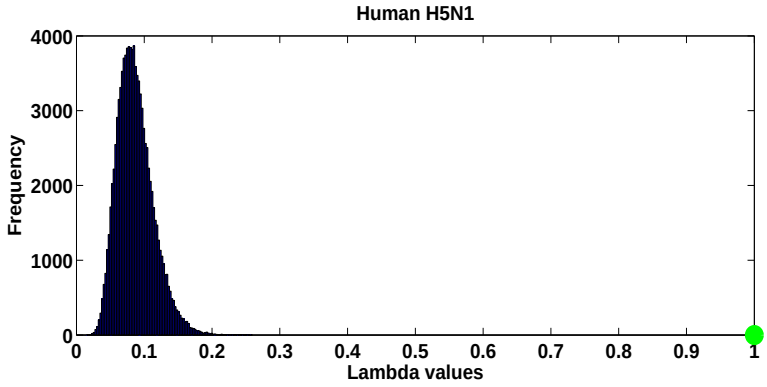

Figure 4.20: Human H5N1 class label randomization simulation result. The green dot represents the λ_o value (observed class separateness value). The blue line on the far left of this figure represent the distribution of $\lambda_{i=1...K1}$.

4.7.5 Avian H5 influenza viruses

Avian influenza A viruses differ from human viruses by recognition of the receptor binding of α-2,3 sialic acid instead of the α-2,6 [131]. There are 16 HA subtypes circulating in the avian population and that they are lesser studied then the human subtypes. The visualization of the evolution of avian H5 is presented in Figure 4.21. The overall observation that arises from our analysis is that rather than forming a restricted directional trend, the evolution of avian influenza virus is characterized by a collection of clusters scattered on the PCA 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 human seasonal influenza viruses can be observed in the plots. The increased genetic diversity since 2000 has been observed in [132] and is captured in Figure 4.21 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 these two avian subtypes evolve much slower than seasonal human influenza viruses.

The Class-separateness analysis of the avian H5 (non-vaccine controlled) indicated that the observed class separateness value λ_o is at 0.268 which is not far from the mean of the distribution generated by using Alg I. The distance in terms of standard deviation is at 3.16 ± 0.06 as shown in Figure 4.22. The distribution of $\lambda_{i=1...K1}$ is from one instant of the inner loop of Alg I. The $K1$ parameter was set at 10000 and K2 was set to 1000.

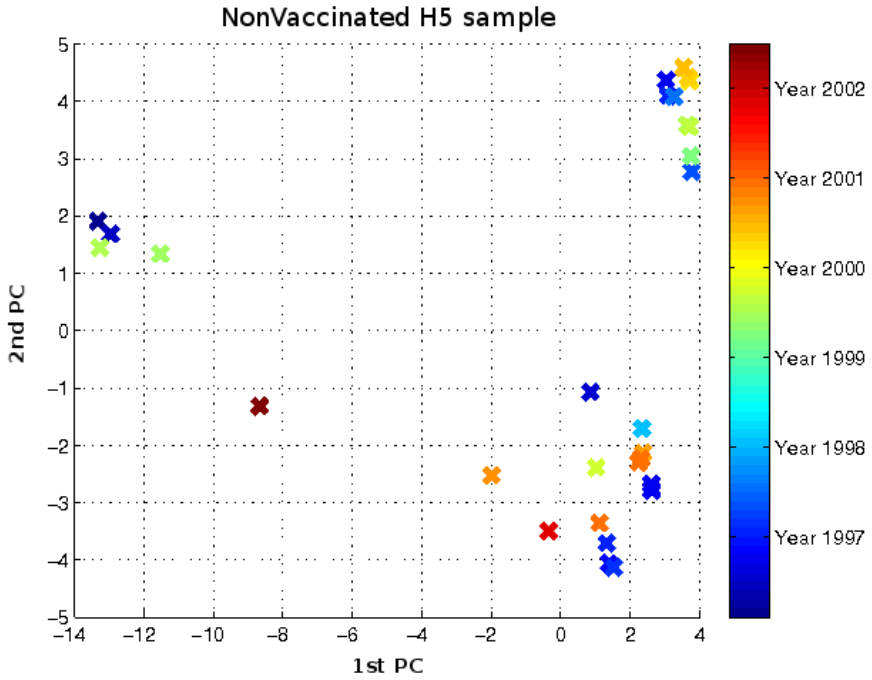

Figure 4.21: Two dimensional visualization of avian H5 influenza virus evolution. The old and new clusters are overlapped and there is no distinct chronological directional trend.

In late 1993, an outbreak of H5N2 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 [45, 4]. The vaccination program was in effect for over 13 years but an increase in respiratory signs of disease was observed in vaccinated chickens [4]. The suspected cause of this increase signs of disease was the antigenic drifts occurred in the influenza viruses [86]. 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 early 1994 and has been in

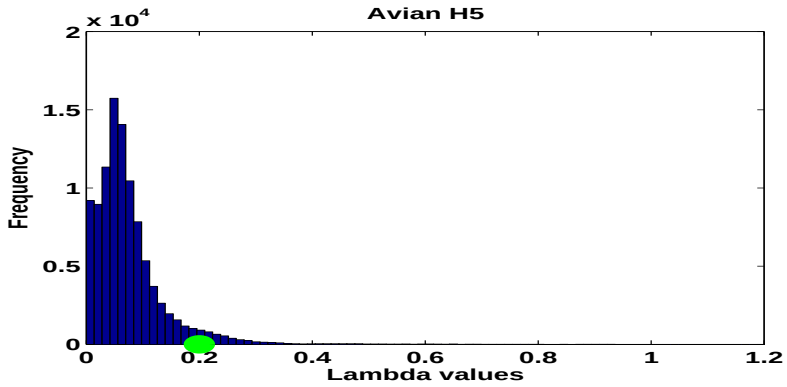

Figure 4.22: Avian H5 class label randomization simulation result. The green dot represents the λ_o value (observed class separateness value). The blue line on the far left of this figure represent the distribution of $\lambda_{i=1...K1}$.

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 4.23) to show the evolution of the field isolates from 1994 to 2002. The first observation from Figure 4.23 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 [45, 4] based on phylogenetic analyses conducted on the same sequence sample. The conclusions drawn from [45, 4] was that cumulative genetic drifts in the HA gene provided an advantage in viral evolution and that the diverged isolates were distinct from the vaccine strain used in the vaccination program.

The Class-separateness analysis of the vaccine controlled avian H5 (Mexico) sample

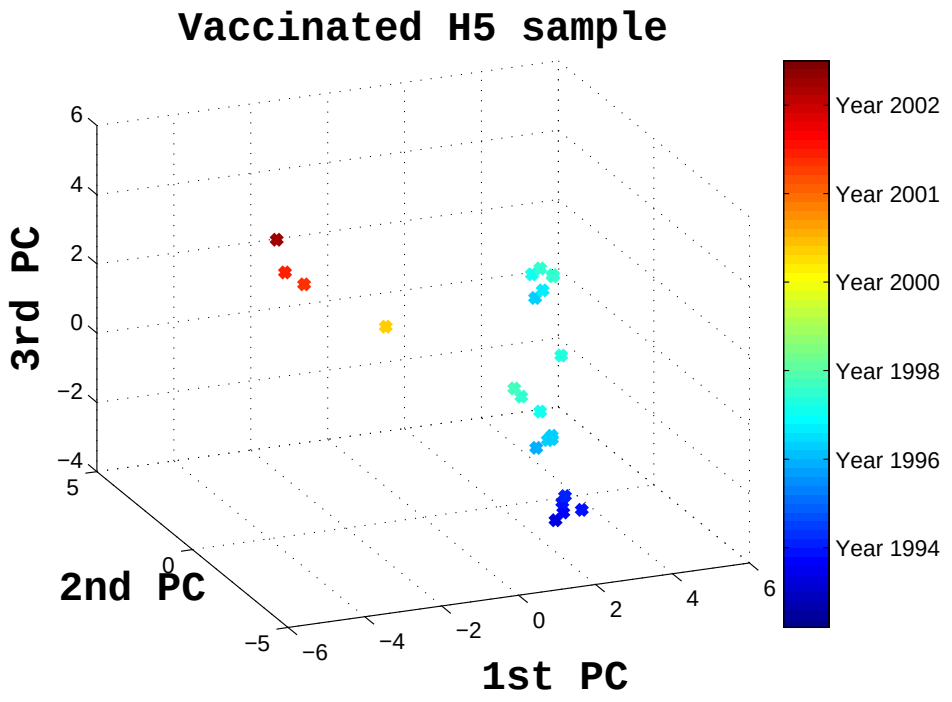

Figure 4.23: Avian H5 influenza virus (vaccinated sample) evolution from 1994-2002.

indicated that the observed class separateness value λ_o is at 1.7 which is quite far from the mean of the distribution generated by using Alg I. compares to the λ_o value of the non-vaccine controlled avian H5 sample. The distance in terms of standard deviation is at 12.23 ± 0.11 as shown in Figure 4.24. The distribution of $\lambda_{i=1...K1}$ is only one instant from the inner loop of Alg I. The $K1$ parameter was set at 10000 and K2 was set to 1000.

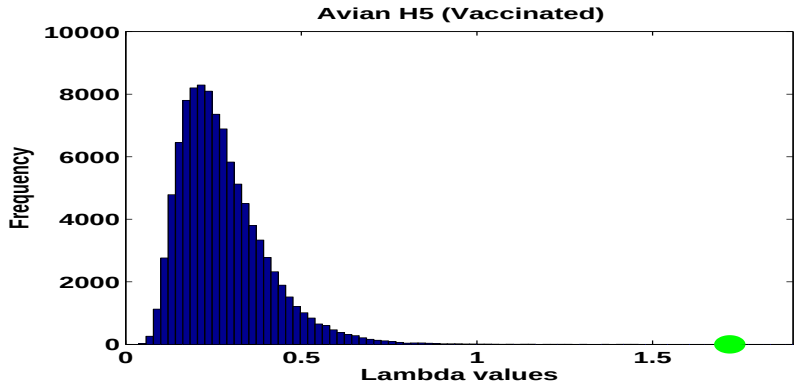

Figure 4.24: Avian H5 (vaccine controlled) class label randomization simulation result. The green dot represents the λ_o value (observed class separateness value). The blue line on the far left of this figure represent the distribution of $\lambda_{i=1...K1}$.

4.8 Discussion and Conclusions

From a visualization point of view, we can summarize the observations between vaccine controlled influenza virus and nonvaccine controlled influenza virus as follows:

- Vaccine controlled seasonal human influenza
  - Restricted directional evolution trend.
  - Evolution follows a chronological pattern.
  - Viruses cluster around vaccine strains.
  - Clear separation between clusters.
  - Narrow band of clusters along the directional evolution path.
  - Clusters contain viruses with the same isolation year.
- Non-vaccine controlled influenza

- Wider and overlapped clusters.
- Late clusters seem to be 'scattered' around early clusters.
- No obvious chronological ordering of clusters.
- Clusters contain viruses span longer time period.

Apart from the genetic distance changes, each 'clump' on the evolutionary path of the virus can be identified by using the isolation year of the virus. This suggests that the isolation year of the virus can be used as a class cluster label when computing the cluster scatter for vaccine controlled and nonvaccine controlled influenza virus. The hypothesis is that vaccine controlled viruses tended to cluster toward the yearly vaccine seed strain for each season, thus leading to clusters contain viruses with the same isolation year as observed in the PCA plot. Following this observation, we computed the class or clusters separateness of seasonal A/H1N1, A/H3N2, influenza B Victoria and Yamagata lineages, avian H5 (Mexico), avian H5, and human H5N1 influenza viruses using the multi-class scatter matrix computation method for both the before and after class labels randomization process. We performed 10000 runs of inner loop "K1" from Alg. I on both influenza sample groups (vaccine controlled and non-vaccine controlled). The results from Alg I are shown in Table 4.2 and in Table 4.3 respectively. The observed separateness measure λ_o of humans A/H1N1, A/H3N2, and Type B influenza are consistently at a large distance from the mean of the empirical distribution generated from Alg I. On the other hand, the observed separateness measure λ_o values of avian H5

Sample	λ_o	Distance
A/H1N1	24.7	$617.2 \pm .04$
A/H3N2	30.52	$978.3 \pm .031$
Influenza B (Vic)	26.31	$1310 \pm .02$
Influenza B (Yam)	25.38	$1327.8 \pm .019$
Human H5N1	1.01	$34.8 \pm .029$

Table 4.2: Multi-class scatter measurement results: For each human sample, we measured the 'clumpiness' before and after randomization procedure. λ_o indicates the clumpiness measurement before randomization. # of seqs is the number of sequences in the dataset. Distance is the number of standard deviations λ_o is away from the mean of the empirical distribution obtained from running Alg I.

Sample	λ_o	Distance
Avian H5(Mexico)	1.7	$12.23 \pm .11$
Avian H5	0.268	$3.16 \pm .06$

Table 4.3: Multi-class scatter measurement results: For each avian sample, we measured the 'clumpiness' before and after randomization procedure. λ_o indicates the clumpiness measurement before randomization. # of seqs is the number of sequences in the dataset. Distance is the number of standard deviations λ_o is away from the mean of the empirical distribution obtained from running Alg I.

and human H5N1 influenza viruses (non-vaccine controlled) are close to the mean of the empirical distribution. The area under the tail of the distributions beyond the observed separateness values was below rounding error of 10^{-16} which made the computation of p-value not possible. The large margin of difference in observed λ values between vaccine controlled and non-vaccine controlled samples indicated that there is a difference in evolution between these two groups of influenza viruses. The class separateness analysis results indicated that the vaccine controlled viruses showed much higher 'cohesiveness' in each year than the viruses from non-vaccine controlled samples.

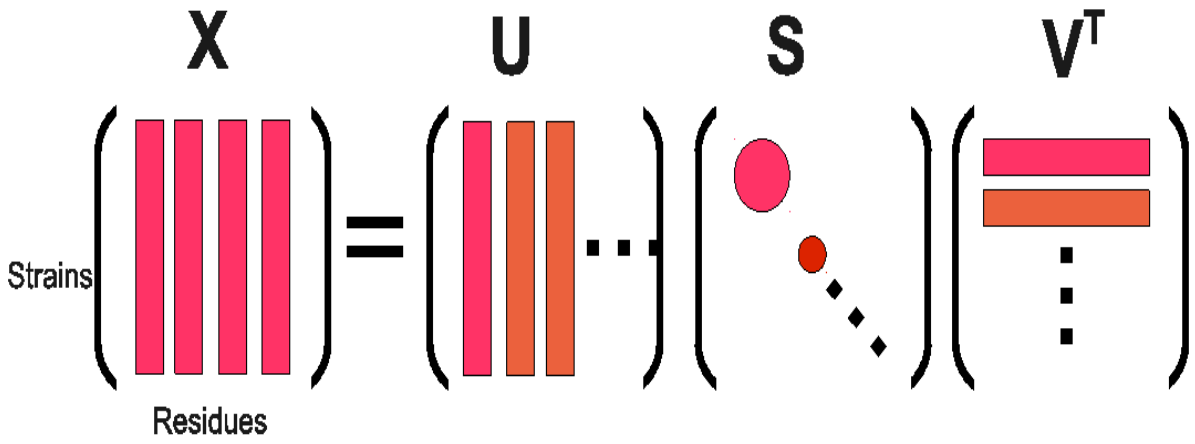

Figure 4.25: Singular Value Decomposition. X is the data matrix with rows as strains and columns as residues.

Vaccination is the principal measure for preventing influenza and reducing its impact [133, 134]. Almost a century ago after the isolation of the first influenza virus, influenza vaccines have been persistent and have evolved to respond to the evolution of the influenza viruses evolving in humans [135, 136]. 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 [65, 45] 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 seasonal human influenza viruses and

nonvaccine controlled influenza viruses in avian and human in order to provide a wider picture of the evolution dynamics of the virus in difference environments. The restricted directional evolutionary trends and clusters formation around the vaccine strains along the evolutionary paths exhibited by the vaccine controlled influenza viruses are in sharp contrast to the nonvaccine controlled influenza viruses. Apart from this distinction, the naturally emerged chronological ordering of vaccine controlled influenza viruses in the PCA plots is much more noticeable than the nonvaccine controlled viruses. This natural chronological ordering reflects the active adaptation of the viruses to their changing environment. The 'clumpiness measure' exposes the fact that vaccine controlled influenza viruses that share the same isolation year have the tendency to cluster tightly together or form narrow bands. The narrow bands indicate the gradual changes of the genetic composition of the functional sites of the HA surface protein. In contrast, nonvaccine 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 nonvaccine controlled influenza viruses has clearly captured the fact that these groups of 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 almost every year. This is clearly reflected in the very high λ_o value for vaccine controlled influenza viruses. In addition, the cohesiveness of strains in each year in vaccinated samples are much higher than the strains from nonvaccinated samples as suggested by the Alg I results. 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 nonvaccine 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 protected and wild environments. 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 [90]. 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.

Chapter 5

Influenza Reassortant Prediction

5.1 Introduction

In this chapter, we propose a method to find influenza reassortant candidates among a large collection of strains represented by their genetic sequences. The proposed method is based on visual inspection of all eight genome segments of the virus simultaneously on a single scatter plot.

In April 2009, a novel swine reassortant caused a global pandemic [137]. The segmental content of the genome was traced to be originated from different lineages. Trifonov, V.[138, 139] confirmed that this pandemic reassortant virus consisted of six internal gene segments (PB2, PB1, PA, HA, NP and NS) from the triple-reassortant swine H3N2 North American swine lineage and M and NA gene segments descending from a Eurasian lineage of swine influenza virus. The exchange of gene segments between

two or more influenza viruses resulting in the production of reassortant viruses is called the reassortment process and is termed antigenic shift (see section 1.6.1) when the surface HA and NA genes are swapped that resulting in a new antigenic subtype [11] of influenza. A new subtype of influenza can lead to a potential pandemic outbreak if transmitted effectively among humans since humans have no prior immuno capability to fight the new subtype.

The detection of reassortant virus is difficult [140] as many virus lineages co-exist among mammalian and avian hosts. Yet, the understanding of the influenza virus evolution at the genomic scale is vital to shed light on the inter-relationship between different lineages that generate reassortment events. Therefore, in order to successfully identify reassortant virus, the problem becomes to identify the lineage origin of emerging novel strains. Given the genetic diversity (16 subtypes) and the numerous virus lineages exist and are co-circulating in human, swine, and avian populations; it is a daunting task and a huge challenge to pinpoint the lineage origin of a flu virus. On a crude level, one can first use BLAST [141] to search the NCBI Influenza Virus [6] sequence database to identify the most recent ancestors of the query sequence. Then one can use a phylogenetic analysis to perform a one to one gene segment comparison to resolve small sequence polymorphism in order to assign a specific lineage for the query sequence [142, 143]. There are two immediate challenges one faces in tracing the lineage origin of a virus. One, the complete genome sequence for historical strains and recent circulating strains are limited, and often not all their eight segments have been completely sequenced and

made available in the database. Two, if phylogenetic analysis is to be utilized, separate tree needs to be built for each gene segment. Even though this is the most accepted method, it depends on the underlying tree building algorithm. In addition, correct reference sequences need to be used when constructing trees to identify reassortant virus. This reference sequence set usually can be collected by using previously reported strains in the literature. Presently, there is no gold standard reference influenza genomic dataset available specifically for influenza reassortant detection.

5.2 Background

The three most recent pandemics (1957, 1968, and 2009) have been caused by reassortant influenza A strains [144]. Given the segmented genomic structure of the influenza virus, a progeny virus can inherit gene segments from two different subtypes of the influenza viruses during co-infection and replication [145, 146, 147]. Historically, the main focus has been on the two segments (4 and 6) which encode the HA and NA surface proteins when detecting reassortment events between subtypes. This is because the antibodies from human immunity system cannot recognize the new surface proteins from a reassorted virus immediately, and this can result in a pandemic outbreak. Therefore, identification of new reassortants is important in understanding the evolution of the virus and can help set early warning sign when reassortants are detected.

Phylogenetic tree analysis [148, 143, 149, 150] has been used as the main computational method to detect reassortant influenza viruses. In order to detect reassortant influenza virus using phylogenetic trees, the assumption has been to look for the tree topology discrepancy between trees built for each influenza genome segment. This is because reassortment event has caused the topology to be different. However, the topology difference can also be caused by phylogenetic construction errors. Alternative computational approaches have been proposed to detect influenza reassortants [151, 147, 152, 153]. In [151], full influenza genome nucleotide sequence was utilized with each segment's closest neighbors calculated using Phylip software with the Jukes Cantor evolution model. The calculation of 'closest' neighbor gives the genetic distance matrix for all pairs of strains and segments. A comparison analysis of the neighbors for each segment is then utilized to identify the number of common neighbors for each segment. The higher the number of common neighbor strains for each segment, the more likely that the test strain is not a reassortant virus. If there are very few common neighbors for the segment, then the test strain is likely to be a reassortant virus.

5.3 Method

In this section, we outline the proposed reassortant detection method and demonstrate its usefulness by presenting the results from analyzing full genome sequences of influenza viruses.

5.3.1 Data Processing

Influenza genome sequences were downloaded from NCBI influenza database [6]. Each downloaded full genome nucleotide sequence contains eight segments (PB2, PB1, PA, HA, NP, NA, M, NS) which are different in length (refer to chapter 1 for the length of each segment). Each segment was pre-processed by converting to a binary string [95]. After the binary conversion, each segment is collected into a matrix A with the longest string as the column dimension of A. 0 is appended to the shorter binary strings so that all eight binary strings will have the same length.

5.3.2 Reassortants Detection

A special feature of the Principal Component Analysis (PCA) is the capability to project high dimensional data points onto a low dimension space. This is special because the computed principal components from the data points can be 'reused' for future data points. We take advantage of this 'reusability' of the PCA components to help in identifying influenza reassortants. The computation of the principal components can be carried out by using Singular Value Decomposition (SVD) algorithm as described in Chapter 4.4.2.

The influenza reassortant method developed here can be viewed as a three-step process. First, we compute the principal components V of the training data points in the binary data matrix A. Second, we apply a pre-processing step to the new testing data points in binary matrix X. The pre-processing is needed in order to make sure

that the training and testing data are consistent in scale. Three, the projection of the pre-processed new data points $\hat{X}$ is accomplished through $\hat{X} \times V$. For example, given the new data points in matrix X of size $\hat{m}$ by n, one must first center the new data matrix by subtracting the column mean from each column of X. Here, the column mean to be subtracted is the column mean from the training data matrix A. This is to ensure that the new data points undergo the same 'pre-processing' as the old data points before projection. Once the new data points are centered, the projection can be performed simply by $\hat{X} * V$ and the resultant matrix has the dimension of $\hat{m}$ by 2. Once the projection of the new/testing data is completed, if no reassorted gene segment is present in the testing virus, one sees eight pairs of gene segments (ie. HA segment of reference virus pairs with HA segment of testing virus) on the scatter plot.

5.3.3 Automated detection

Although the visualization of the genome segments in a two dimensional space gives a clear view of the relationship between the test virus and the reference virus's genome segments, one can improve this approach by providing a simple automatic method in with simple pairwise distance measure can be used to detect reassorted genome segments. The assumption here is that if two gene segments are from the same lineage, their genetic composition should have a high degree of similarity and should exhibit a very small distance between each other in the 2D space graph. On the other hand, if two gene segments are from two different lineages, then their genetic composition

should show a low degree of similarity and produce a large distance between each other in the plot. The simple method is to compute the pairwise distance between each genome segment of the test and reference viruses using the PCA coordinates of each genome segment. An $8-by-8$ pairwise distance matrix can be easily obtained and can then extract the diagonal elements of this matrix to inspect the distance between each genome segment. The diagonal elements of this matrix gives the distance between the pairings of the gene segments in the PCA plot. This is because we are only interested in the pairings of the same genome segment between the test and reference virus genomes. One factor that needs to be addressed is a cutoff value for the pairwise distance. This cutoff value indicates whether the pair of gene segment is 'close' to each other or far away from each other. If the gene segments are below this cutoff value, then we say that their pairwise distance is small and that they are similar. On the other hand, if the gene segments are above the cutoff value, then we say that this pair is far away from each other and thus are from different parental strains or lineages. The cutoff value acts as a threshold value or cutoff for the minimum distance between each pair of gene segment that is needed in order to capture the separation between two gene segments on the PCA 2D space. The selection of the cutoff value can be determined by the mean of pairwise distance matrix. Given that the genetic composition of two viruses from the same lineage can have small variation due to synonymous substitutions, using the mean pairwise distance as cutoff should allow for some flexibility to account for this type of difference between two virus genomes.

5.4 Results

5.4.1 Swine influenza H3N2 Reassortant Virus

In 1998, a double reassortant influenza virus A/Swine/North Carolina/35922/98(H3N2) was isolated in North Carolina swine farm. This double reassortant virus [154] was found to have gene segments (PB2, PA, NP, M, NS) from classical swine H1N1 lineage combined with gene segments (PB1, HA, NA) from a human seasonal H3N2 influenza virus [155]. This North American swine H3N2 strain carried gene segments from two different lineages. The classical swine H1N1 lineage is closely related to the 1918 H1N1 Spanish flu virus and other human influenza viruses isolated in the 1930s [68]. The surface genes HA and NA (segment 4 and segment 6) and internal gene segment 2 (PB1) came from the seasonal human H3N2 lineage which had been circulating in humans since 1968. The rest of the gene segments came from classical swine influenza virus [68]. The 'pairing' of these gene segments is shown in Figure 5.1.

A reassortant virus H3N2 A/SW/CO/77 genome sequence identified in [156] to test the predictive power of our approach. We selected this isolate because its genetic characterization by [156] using phylogenetic trees indicated that A/SW/CO/77 pig isolate's HA and NA proteins are closely related to the human influenza virus. In this second analysis, we conducted two tests: an experiment test and a control test. For the experiment test (result shown in figure 5.2), we first computed the principal components using field isolates of human origin flu viruses (see Materials and Methods for human

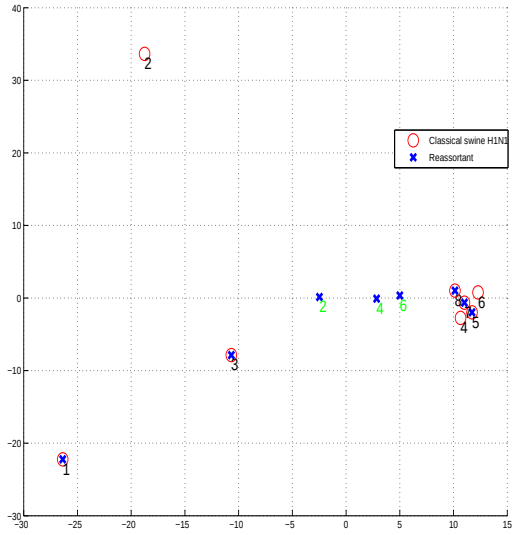

Figure 5.1: A swine H3N2 double reassortant virus (blue) isolated in North Carolina in 1988. The red circles are the classical swine H1N1 gene segments and the blue 'crosses' are the gene segments of the swine H3N2 double reassortant virus. Segments 2 (PB1), 4 (HA), and 6 (NA) of this reassortant virus were from human seasonal H3N2 influenza virus and the rest of its gene segments were from classical swine lineage. Each gene segment is numbered from 1 to 8. Segment 2, 4, and 6 of the reassortant virus do not overlap with the same gene segments from the classical swine H1N1 virus.

virus genomes used) and then projected the A/Swine/CO/77 genome onto these pre-computed principal components. We see that the HA and NA proteins of A/SW/CO/77 are closely "attached" to the human HA and NA counterparts, which suggests that these two surface proteins were originated from a human-host type virus during reassortment event.

For the reference/control virus (result shown in Figure 5.2 and Figure 5.3), we selected the H3N2 A/swine/Wisconsin/2/1970 swine virus as the control genome because

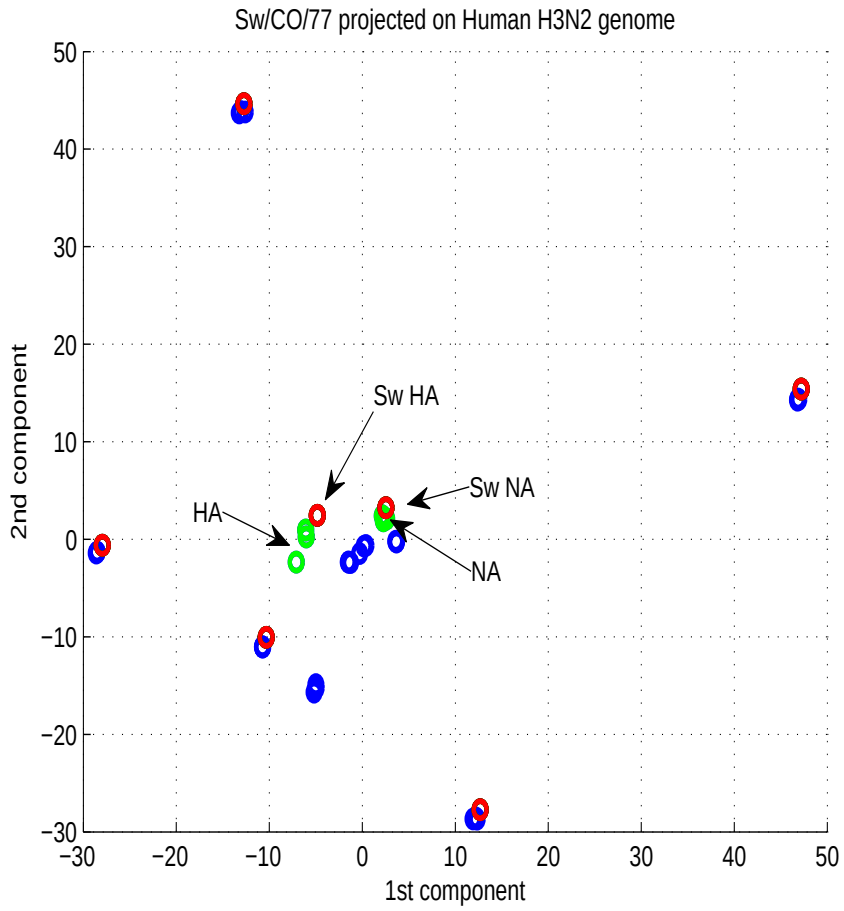

Figure 5.2: SW/CO/77 genome projected onto principal components computed using human origin flu viruses genomes. Green dots represent the Human HA and NA surface genes, red dots are the SW/CO/77 genes, and blue dots are the internal genes from human host genome.

A/SW/CO/77 was isolated in 1977. The reason for selecting a 1977 strain as a control is that the swine flu virus lineage at that time had not diverged into multiple lineages that carried gene segments with mixed host type [156]. This is also to assure that the control strain contains only gene segments from a single host type of swine origin. Based

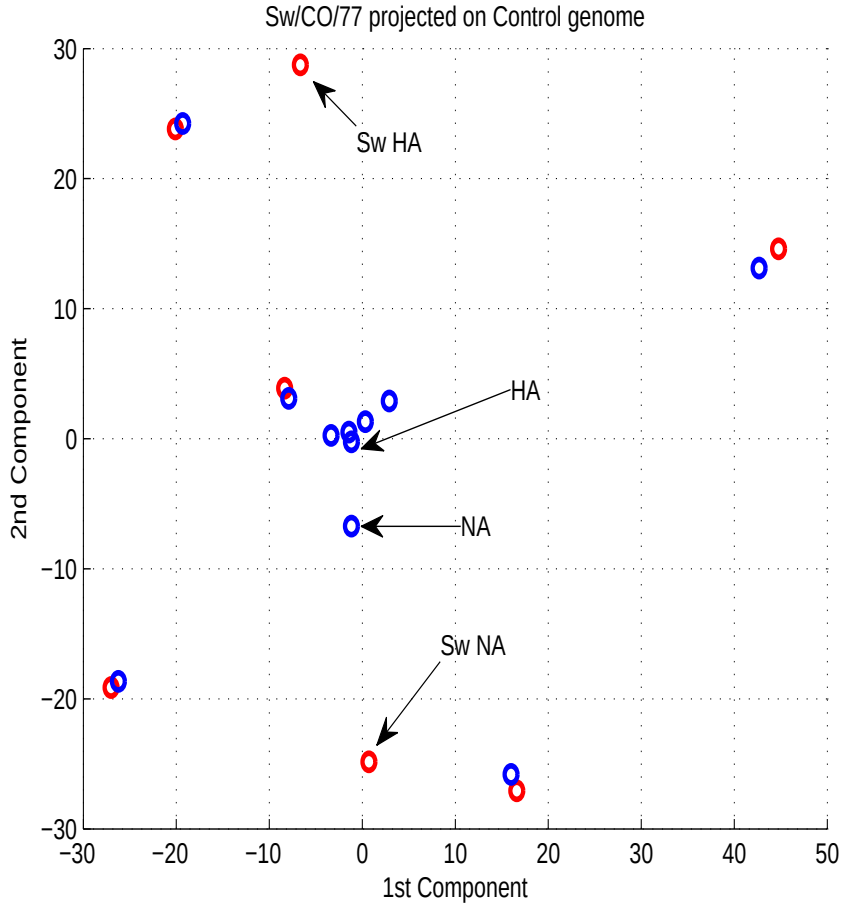

Figure 5.3: SW/CO/77 genome projected onto principal components computed using swine virus genome as reference. Red dots represent A/SW/CO/77 genes and blue dots represent the reference virus A/Swine/Wisconsin/ genes.

on phylogenetic analysis, A/swine/Wisconsin/2/1970 does not contain foreign host type gene. We pre-computed the principal components using the control genome sequence and then projected the A/SW/CO/77 genome onto the first two components. Clearly, we can see that A/SW/CO/77 strain's HA and NA proteins (red dots) are clearly distantly apart from the swine origin counterparts (blue dots). From the results of these

two reassortant detection tests, we can see that there is an unique feature or a signature pattern that represent each specific host type. With the right feature representation, PCA can quickly isolate and identify these type of attributes in the dataset.

5.5 Discussion and Conclusions

Identification of reassortment events is neither trivial nor straightforward. The identification process can usually be divided into a three step procedure. One, multiple sequence alignment of reference sequences with test sequence is usually performed. Two, phylogenetic tree construction for all eight segments using the aligned sequences. Three, identify conflicting tree topologies between gene segments as the potential reassortment. Selecting reference sequences is not a straightforward task in itself. The choice of reference sequences can have significant impact on the result of the reassortment identification. A reassortment event is more easily detected if the two strains involved in producing the reassortant are sufficiently divergent in their sequences. Reassortment between very similar strains is likely to go undetected by most, if not all, methods.

The influenza reassortants detection method presented in this chapter is based on visual inspection of eight influenza gene segments simultaneously on a single scatter plot. The assumptions we made in this approach is that if two influenza viruses have the same evolution history, their gene segments should have high degree of genetic similarity and that the segment length should be the same. These assumptions are consistent with the approach taken by Rabadan et al.,[157] in that they suggested that the evolutionary

rate of gene segments should remain the same if two viruses share the same evolution history.

The proposed method avoids using multiple sequence alignment and phylogenetic tree approaches because the objective of reassortant detection or identification is to discover 'immediate genetic changes' within the influenza virus genome when gene segments comparisons are made to any existing circulating strains. The heuristic results from multiple sequence alignment (MSA) and phylogenetic tree construction are often based on underlying evolutionary assumptions. Previous studies have shown most of the available MSA methods only have accuracy of less than 70 percent, and even as low as 40 percent [158]. Because the mechanisms behind the preferential reassortments are still not completely understood, it is our believe methods to detect reassortment events can benefit from relying on the signal in the data instead of underlying evolutionary process assumptions.

Chapter 6

North American Swine H3N2 Influenza

6.1 Introduction

In this chapter, we present the results using method developed in chapter 4 to study the North American swine H3N2 influenza viruses. Based on our computational analysis of the genetic information of the hemagglutinin gene, we identified 2 hemagglutinin amino acid residues positions 142 and 144 have led to the formation of the distinct pre- and post-2009 clusters and could potentially act as a cluster signature to identify future swine A/H3N2 virus. Recent H3N2v exclusively clustered into a post-2009 cluster which consists of swine influenza viruses from Ohio agricultural fairs.

6.2 Background

Three main subtypes of influenza virus (H1N1, H3N2, and H1N2) are circulating in domesticated swine populations around the world. In 1998, H3N2 virus was isolated in pigs in North Carolina, Minnesota, Iowa, and Texas [155] with gene segments similar to those of human (HA, NA, and PB1) and classic swine (NS, NP, and M), and avian (PB2 and PA) lineages. This triple reassorted virus has been circulating in the US swine population since 1998, and in 2010 a new variant that contained a matrix gene of the 2009 A(H1N1)pdm09 virus was first identified in U.S. pigs [159]. Swine influenza surveillance has intensified since the pandemic of 2009 caused by swine origin H1N1 leading to multiple sequence submissions made available for study. Previous studies [160, 161, 18] have demonstrated that different swine A/H3N2 influenza virus clusters were emerging in the U.S. swine populations.

The emergence of these post-2009 swine A/H3N2 influenza clusters can potentially lead to the steady increase of the swine A/H3N2 influenza virus diversity in North America. It is therefore important to identify any potential cluster signature that can quickly determine the cluster origin of the virus and to allow for more effective monitoring of its evolution. We therefore hypothesize that the genetic information on the hemagglutinin surface gene is sufficient to define the cluster diversity of the HA gene segments. Here, we show that using computational techniques, variation in two amino acid residues of the HA gene is sufficient to explain the existing cluster diversity among swine influenza A/H3N2 virus in North America.

6.3 MATERIALS AND METHODS

6.3.1 Sequence data

A total of 816 U.S. swine influenza A/H3N2 virus hemagglutinin nucleotide and protein sequences from 1999 to 2013 available from NCBI influenza database [6] were downloaded for our study. Majority of the sequences represent collections from Illinois, Minnesota, Indiana, Iowa, Michigan, Nebraska, North Carolina, Ohio, Pennsylvania, Texas, Wisconsin, and Wyoming over a period representing the years of 2009 to 2013. We also included representative strains of the classic swine clusters I,II,III, and IV in our dataset. Eight H3N2v virus hemagglutinin sequences from Indiana, Iowa, Maine, Pennsylvania, and West Virginia isolated in year 2011 were also included in the analysis.

6.3.2 Methods

6.3.3 Multiple sequence alignment and conversion

Multiple sequence alignment on all sequences were performed using NCBI multiple sequence alignment tool [6] using default parameters. All sequences were digitized to 0 and 1 values that directly enables a binary sequence analysis as detailed in Chapter 4 of this thesis. The processed sequence data was then visualized using the high dimensional reduction method found in Chapter 4 and in [95].

6.3.4 Weight assignment

In order to identify the positions on the HA gene that contribute the most to the clustering, we also computed a weighted PCA with each site weighted by its variability. The variability of each position is shown in Figure 6.1. Given the aligned HA protein sequences, the weight ω which measures the variability of each HA position is obtained by computing its columnwise Shannon entropy $H(Y)$ [162] where Y is a discrete random variable with alphabet Λ of twenty amino acids. The probability P_{ij} is estimated as the observed fraction of each amino acid in position j that equals $P_{ij} = \frac{1}{m}\sum_{l=1}^{m} I(Y_{lj} = y_i)$ where Y_{ij} is amino acid in strain l position j, an $I(\dot{})$ is the indicator function. The Shannon entropy at j is $H_j(Y) = -\sum_{i=1}^{20} P_j(y_i)\log_2 P_j(y_i)$ for $j = 1, ..., n$. The weight ω_j is then assigned to the diagonal element of a diagonal weight matrix W of size n by n. A high weight or high variability position may indicate an antibody binding site that is under immune pressure. A low weight or low variability position may correlate to structurally conserved site that is responsible for maintaining the core functionality of the protein. Each column of X corresponding to position j was multiplied by $\omega_j, j = 1, ..n$.

6.3.5 Phylogenetic analysis

Phylogenetic analysis of the amino acid sequences of the swine influenza A/H3N2 virus HA gene using the Mega 5.0 software [163, 164] to generate neighbor-joining phylogenetic tree (Figure 6.2) from the dataset with 1000 bootstrap replications to verify the tree

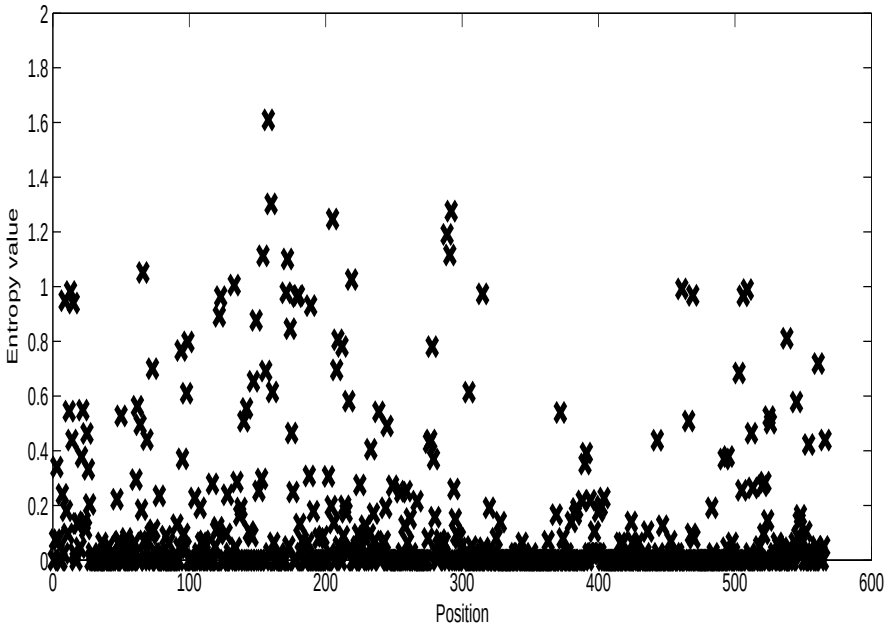

Figure 6.1: HA weight distribution computed using Shannon entropy formulation. Position 142 and 144 (H3 numbering) have entropy value of 1.609 and 1.302 respectively.

topology was performed. The model used was the maximum composition likelihood model with gamma distribution parameter set to 1.

6.4 RESULTS

6.4.1 Clustering analysis

Five distinct clusters of the U.S. swine influenza A/H3N2 virus were produced using the K-means algorithm [52] on the 3D projection data generated based on PCA algorithm. These five distinct clusters are well separated in the PCA 3D space (Figure 6.3). Figures

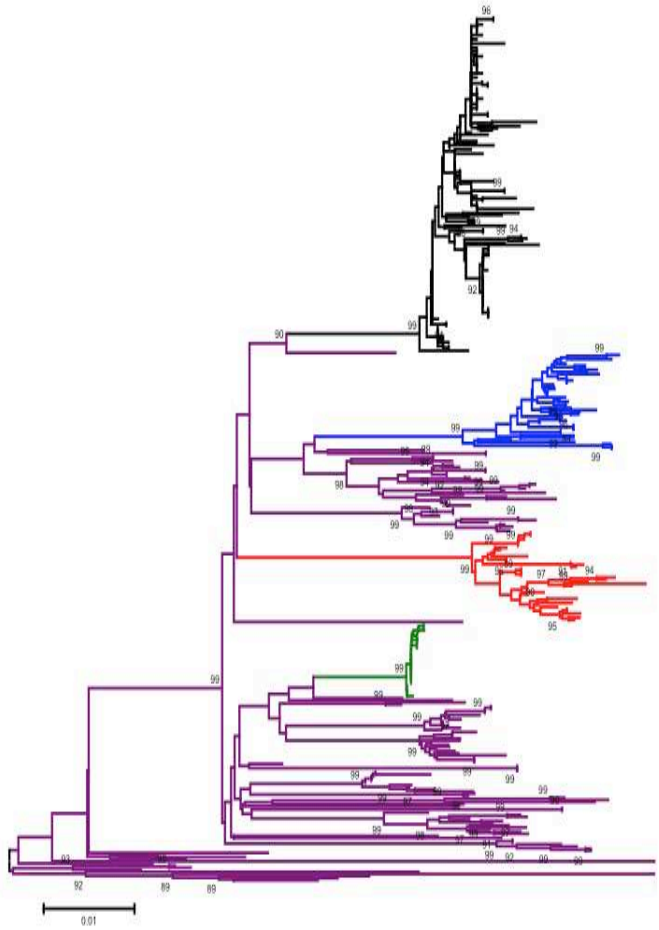

Figure 6.2: North American swine influenza H3N2 virus phylogenetic tree. The color coded branches correspond to the clusters observed in Figure 6.3

6.4 and 6.5 illustrated the clusters by virus isolation year and isolation location. The four-post-2009 clusters are the A, B, C, and E which contain contemporary (2009 - 2013) swine A/H3N2 virus [18]. Cluster D consists of pre-2009 swine influenza A/H3N2 viruses that have been circulating in U.S. between 1998 and 2008 and the Canada originated swine influenza "cluster IV" virus [165]. Viruses in cluster E show the highest genetic

similarity with maximum 6 amino acids difference. The most diverse cluster is the pre-2009 cluster D with maximum amino acid difference of 144. This is not surprising as this cluster contains viruses from the historic swine H3N2 viruses in the U.S. that emerged from the three clusters of [165] derived from three distinct human seasonal H3N2 virus from 1995, 1997 and 1996 and the "cluster IV" viruses from 2006 respectively [154, 166]. In addition, HA gene of the four post-2009 clusters A, B, C, and E exhibited far less diversity compared with cluster D judged by the viruses pairwise within cluster distance. Cluster A consists of all the isolates from Ohio agricultural fairs from years 2010, 2011, and 2012. On the other hand, cluster E consists of all 2009 Ohio agricultural fairs isolates. This finding is consistent with study on recent North American swine influenza A/H3N2 viruses performed by Feng et. al 2013 [161] in which they had shown that 2009 Ohio agricultural fairs isolates were antigenically different than the 2010 and 2011 fair isolates. Cluster C (Red) contains isolates from Iowa, Minnesota, Illinois, Nebraska, Texas and South Dakota from year 2011 and 2012. Cluster B (Blue) contains isolates from Illinois, Iowa, Minnesota, and South Dakota from year 2012. This suggests that neither geographical locations nor year of virus isolation played a role in virus diversity. The most likely explanation is that each cluster reflects a group of antigenically distinct group of viruses that are emerging in the U.S. swine populations. It is not uncommon to find multiple antigenic groups within the same geographical location within the same year.

In July 2011, an influenza H3N2v virus that normally circulates in pigs were found

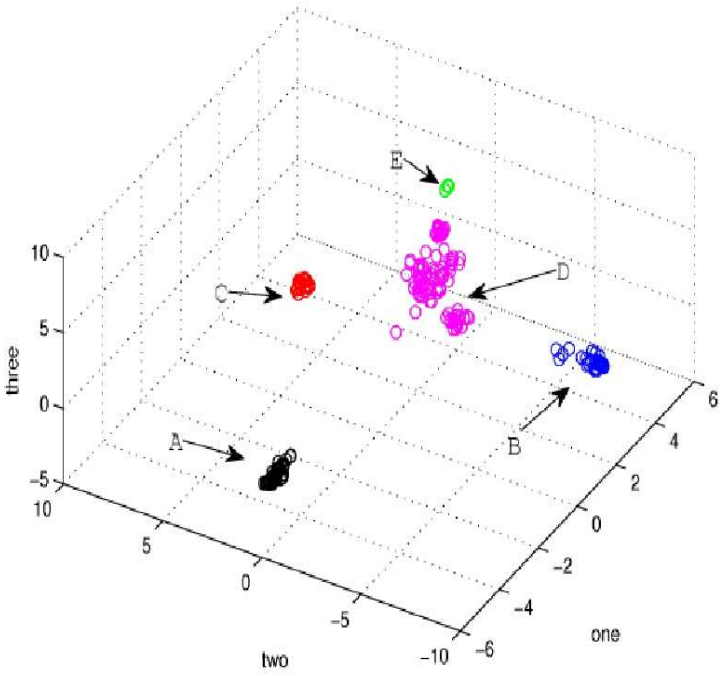

Figure 6.3: North American swine influenza H3N2 virus clusters.

in humans in Indiana, Iowa, Maine, Pennsylvania, and West Virginia. The virus carries the matrix (M) gene from the 2009 H1N1 pandemic virus. In 2012, 307 cases of H3N2v infection were detected in U.S. from 11 states [167, 168]. Using the projection from the unweighted PCA, all available H3N2v isolates from Indiana were projected into different clusters (Figure 6.6). These H3N2v virus and the clustered viruses share the same cluster signature residues at position 142 and 144 respectively.

In order to determine all the changes that led to the clustering pattern produced, we

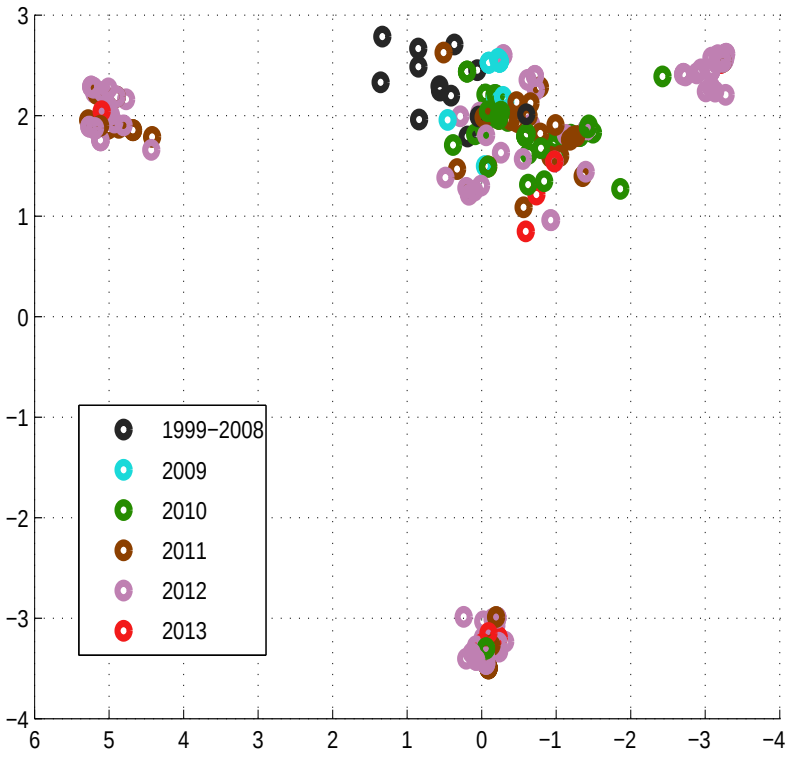

Figure 6.4: PCA projection of 2013 swine H3N2 influenza virus (blue). Viruses are colored by isolation years.

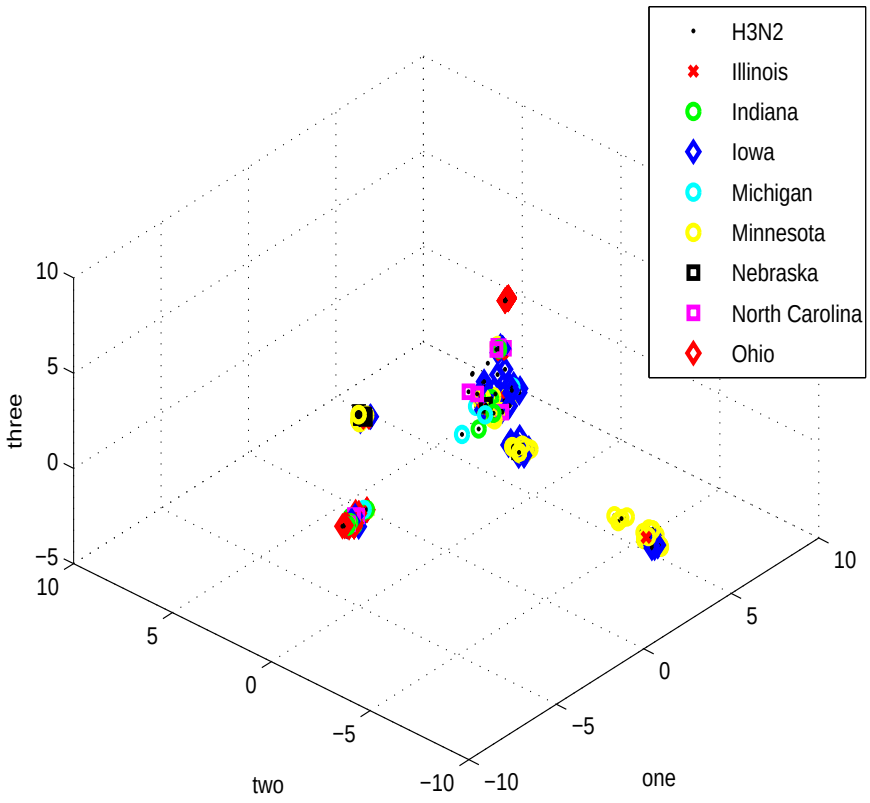

Figure 6.5: North American swine H3N2 influenza viruses by geographical locations.

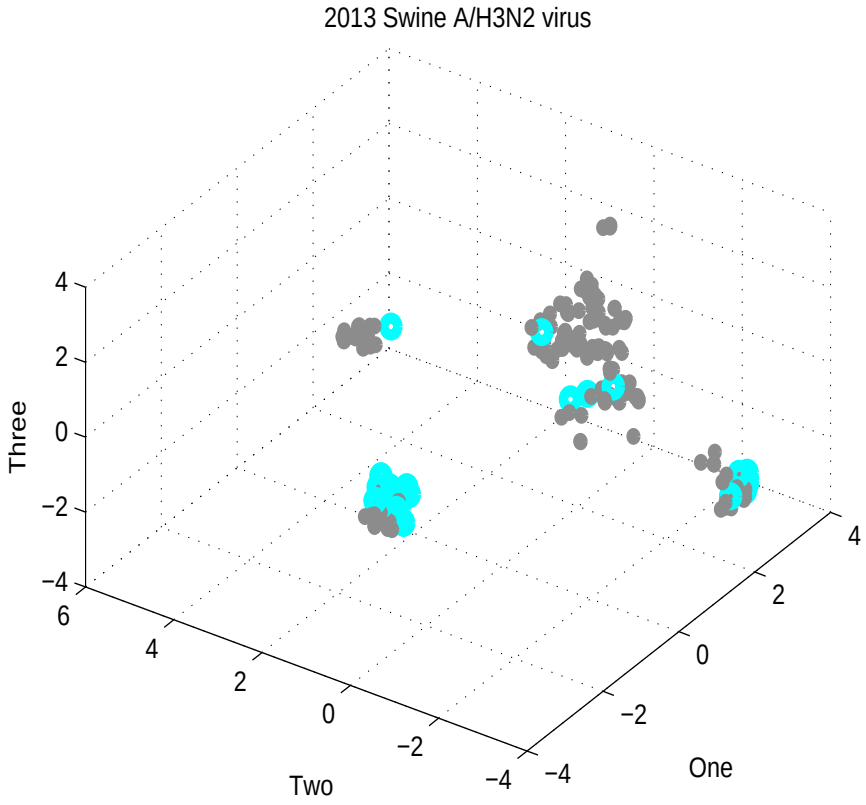

Figure 6.6: PCA projection of 2013 swine H3N2v influenza virus.

used the Shannon Entropy weighting scheme to yield a weighted PCA. The weights are shown in Figure 6.1 and weighted PCA in Figure 6.7. The same clustering pattern was produced when using the phenotype information of the influenza A/H3N2 virus with weighting for each HA position. This indicated that variability of the positions played a key role in forming the five distinct clusters observed. We then proceeded to identify the positions responsible to the formation of these five distinctive clusters. Based on the weighting value of each position, we selected position 142 (H3 numbering) within the

antigenic site B which shows the highest variability and reconstructed the data matrix using only phenotype information from this position and computed the weighted PCA. Using this position alone, we were able to reproduce the same clustering pattern (Figure 4) observed in Figure 1. Further, when both the weight values of residue position 142 and 144 (second highest entropy value) were used, we also were able to reproduce the same clustering pattern seen in Figure 1. This suggested that the variability present in these two positions is more than sufficient to classify the virus into difference clusters. At HA residue position 142 (antigenic site B), cluster A has glycine (Gly), cluster B has glutamic (Glu), cluster C has asparagine (Asn), cluster D has arginine (Arg) or serine (Ser), and cluster E has lysine (Lys) respectively. At residue position 144, cluster A has valine (Val), cluster B has aspartic (Asp), cluster C has Valine (Val), cluster D has isoleucine (Ile), or aspartic (Asp), or phenylalanine (Phe), and cluster E has glycine (Gly) respectively. The combination of the residues 142 and 144 can form a useful cluster signature or biological marker to designate each individual cluster and can be used to quickly classify swine A/H3N2 virus. For example, cluster A has the Gly-Val signature and cluster B has the Glu-Asp signature. Residue 142 is located on the loop of the HA near the receptor-binding pocket. A mutation at this position may contribute to a shift in receptor binding specificity [169, 170]. Residue 144 is located on the solvent exposed surface of the HA globular head [65, 171] and has been shown to influence the generation of escape mutants due to its glycosylation property [171].

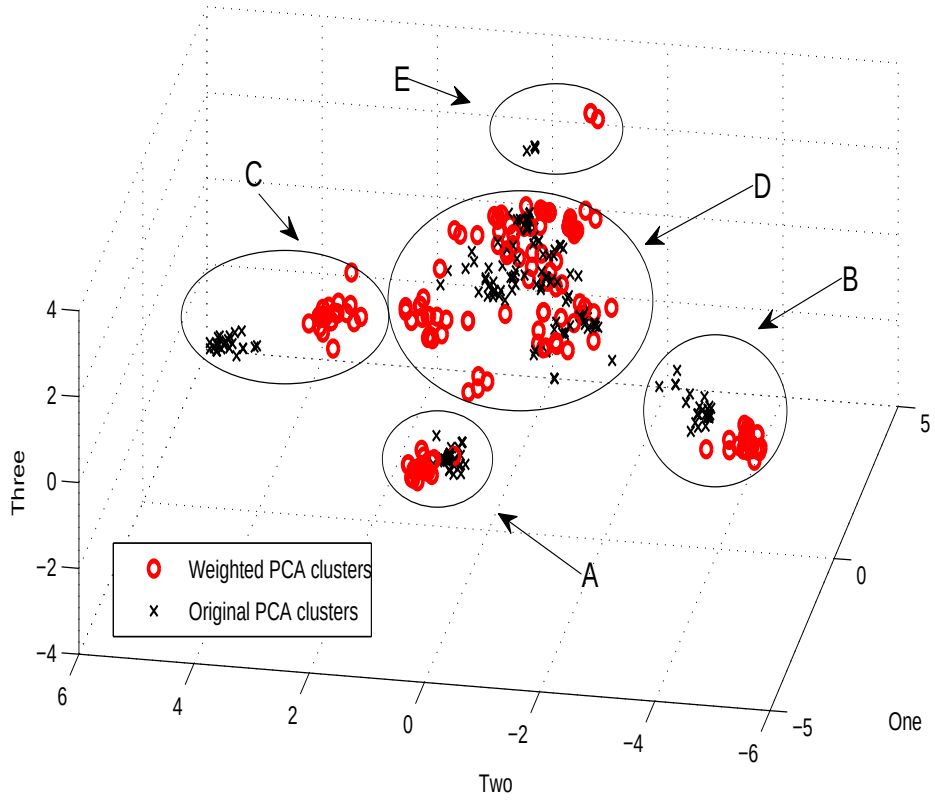

Figure 6.7: Weighted PCA clusters and unweighted PCA clusters.

6.5 DISCUSSION

The influenza A virus can represent a great health concern for the public if not closely monitored. We have shown that the North America swine influenza A/H3N2 virus can be clustered into distinct pre- and post-2009 clusters. Given that pigs are thought to be the mixing vessel for the influenza A virus, reassortant virus originated from pigs with potential highly infectious and human transmissible characteristic can emerge without

warning. It is therefore necessary and important to closely monitor the evolution of the virus. Using computational techniques, we have shown that four post-2009 distinct clusters of swine A/H3N2 virus are emerging in the U.S. swine populations and that the combination of hemagglutinin residues 142 and 144 (H3 numbering) led to the formation of these clusters. There is a possibility that mutations on sites 142 and 144 are related or compensatory to each other. In order to provide such scenario, one would have to use a mutual information approach to determine if these two sites are dependent sites.

Phylogenetic analysis confirmed that these four post-2009 clusters are most likely derived from the historic cluster (cluster D) of the swine A/H3N2 influenza viruses that have been circulating in the U.S. since 1998. H3N2v swine isolated from humans in Indiana clustered with cluster A in which both showed the same cluster signature of Gly-Val combination at residues 142 and 144. Phylogenetic analysis of the HA surface gene confirmed that the U.S. swine A/H3N2 influenza can be grouped into five separate clusters with the four post-2009 clusters originated from the historical swine influenza virus cluster (cluster D). Cluster E is restricted to only 2009 Ohio agricultural fairs isolates and no other cluster contains this 2009 genetic variants, hence it is quite possible that this E group virus is localized to Ohio exhibition swine in that year. In short, our findings highlight the need to closely monitor epitope regions of the hemagglutinin gene of influenza A viruses circulating in swine and the genetic sequence information of the influenza virus can be of great importance to determine the cause of cluster formation.

Chapter 7

Conclusions

7.1 Summary of contributions

The preceding chapters of this thesis describe the computational analysis of the evolution of influenza viruses. A major focus of this thesis has been the study of the evolution of influenza virus under vaccine pressure. Other computational analysis techniques have also been developed to study different aspects of the influenza virus evolution. The vast number of influenza genetic sequences drives the need to develop tools that can allow quick and easy interpretation of the sequenced isolates so that a better understanding of the evolution of the virus can be achieved. The computational techniques developed in this thesis can be viewed as a high throughput analysis approach to analyze genetic sequences that are rapidly growing in size and quantity. The major contributions of this thesis are listed below:

- A compact markov model.
- Influenza virus and vaccine.
- Influenza reassortant detection based on PCA projection.
- Cluster determinant analysis of the North American swine influenza virus.

7.2 Compact markov model

A two-layered statistical constructed from a Markov Chain model and a Poisson process is presented in chapter 3. The Markov model models the single point mutations and the Poisson process models the occurrence of the mutations. The Markov model is compact because we model the state transition based on the Hamming distance of the genetic sequence. The number of states only depend on the length of the genetic sequence. This is a much less complex Markov chain model than any current Markov chain models used in molecular evolution simulation.

7.3 Influenza virus and vaccine

As cost effective sequencing technologies propel the growth of influenza sequence data in quality and quantity, much can be revealed about the evolution dynamics of the influenza viruses base on the analysis of their genetic sequences. As we continue to make progress toward the understanding of the evolution of influenza virus through analyzing sequence samples from around the world, a more detailed picture of how the influenza

virus evolves under difference selective pressures will eventually emerge. At the present time, our primary focus has been on the understanding of the evolution of influenza virus under vaccine pressure. Vaccination programs are widely used around the world in fighting seasonal human influenza virus in north and south hemispheres. However, in order to provide protection, vaccine needs to be updated annually as influenza virus's surface protein HA undergoes constant antigenic drift that leads to immune escape variants. This constant vaccine update in order to track the evolution of the virus has given rise to the development of the universal vaccine [172, 173, 174, 175, 176] which targets the internal proteins of the virus. Although vaccination is commonly used in human population and its efficacy and effectiveness have been examined from time to time, the effects of vaccination on the virus itself has not been actively studied. The results presented in this dissertation research show that influenza virus evolves differently when under vaccine pressure. The most striking result is the evolution of the seasonal human influenza AH3N2 virus. This is also the most frequent updated vaccine strain in the trivalent vaccine that is administered every year. On the other hand, influenza virus evolving in the wild does not show the same evolution pattern as the vaccine controlled influenza viruses. The observational differences of the evolution patterns between vaccine controlled and nonvaccine controlled influenza viruses are further analyzed using scatter matrix computational approach. The results from the analysis indicated that their evolutionary trends or patterns are not the same and that the yearly cohesiveness of viruses from vaccinated samples are much higher than viruses

nonvaccinated samples. The observational differences can be generally summarized as follows:

- Vaccine controlled seasonal human influenza
  - Restricted directional evolution trend.
  - Evolution follows a chronological pattern.
  - Viruses cluster around vaccine strains.
  - Clear separation between clusters.
  - Narrow band of clusters along the directional evolution path.
- Non-vaccine controlled influenza
  - Wider and overlapped clusters.
  - Late clusters seem to be 'scattered' around early clusters.
  - No obvious chronological ordering of clusters.

Influenza vaccination is the most used method in protecting us from the virus but it is also a problem that appears over and over again as the virus evolves continuously. The evolution of influenza virus in a vaccinated environment is rarely studied and how does the virus successfully escape vaccine induced immunity continuously have continue to cast doubt on the effect and 'side effect' of vaccination. In order to fully resolve this problem, one must combine the understanding of the effects of vaccination have on the virus and on its host respectively.

7.4 Influenza reassortant detection method

The influenza reassortant detection method developed in this thesis is based on a visual inspection approach to identify reassortant virus. The method is quick and simple because no phylogenetic trees construction for all the influenza gene segments is needed. The detection method relies on the projection technique from Principal Component Analysis in that the pre-computed principle components of a reference strain genome is used as the 'reference'. The unknown or test virus genome is projected onto the pre-computed principle components and the unknown/test virus projection is then visualized on the same PCA plot along with the reference genome. This method allows for the visualization of all the influenza genome segments at once in a single plot.

7.5 Cluster analysis of North American swine influenza virus

A computational analysis based on the methods developed in this thesis was undertaken to study the genetic diversity of the North American swine influenza H3N2 virus using the HA gene sequence data. At the time of study, five major clusters were found and that each cluster has its own cluster signature that can be represented by using two positions (142, 144) on the HA gene. The significant of these two positions suggests that future isolates can be grouped into clusters based on the genetic bases at these two positions. In other words, it can provide molecular signatures for monitoring future

genetic diversity of the North American swine influenza H3N2 virus.

7.6 Future directions

In this part of the chapter, we provide some future directions that are meant to carry this thesis research further.

7.6.1 Investigate other genome segments

Influenza A has 8 genome segments and we have so far studied segment 4 in this thesis. The rest of the genome segments code for different proteins and have different functional role. Method developed in this thesis can be easily extended to study their evolution trend or trajectory.

7.6.2 Vaccine strain/New antigenic variant prediction

As the human seasonal influenza trivalent vaccine needs to be updated every year, the decision in selecting the matching vaccine seed strain to the next season dominant circulating strain can be difficult. As of now, the selection decision is based on results from antigenic characterization and phylogenetic analysis of influenza isolates available prior to the start of a new flu season. In other words, we are predicting the new antigenic variants based on existing or older data for the new flu season. In order to successfully predict the upcoming dominant circulating strains, it is imperative to first understand the effects (see section 7.3) of vaccination has on the evolution of the virus itself. The

next step is then to develop a computational framework that captures the effects of vaccination.

We also need to consider the number of dimensions needed to give the best approximation to the original 'signal' in the data in order to successfully predict the emerging new antigenic variants. A simple way to determine the number of singular values to use is to inspect the singular values $\sigma's$ produced from the SVD algorithm (see 4.4.3). As an example, we used the seasonal human influenza A/H3N2 virus dataset and generated the first 10 singular values using the SVD algorithm. A sharp drop of the singular value from the first to second principal component indicates that the first PC most likely can capture the majority of variance in the dataset. However, from the 6th PC onward, the drop off has been relatively level and this indicates that any higher principal components retained after PC 6 are mostly noises in the dataset. In this example, the number of dimensions that can capture the original 'signal' without much noise should be 6.

7.6.3 Kernel PCA application

Standard PCA can only handle linear dimensionality reduction. However, if the given data cannot be well represented by a linear subspace due to its complex structure, then standard PCA will fail to capture any important signal from the data. In order to deal with the complex nature of the data, Kernel PCA (KPCA) [177] has been developed to allow us to perform nonlinear dimensionality reduction. To use KPCA, one needs to select a proper kernel function to handle the sequence data, for example, the string

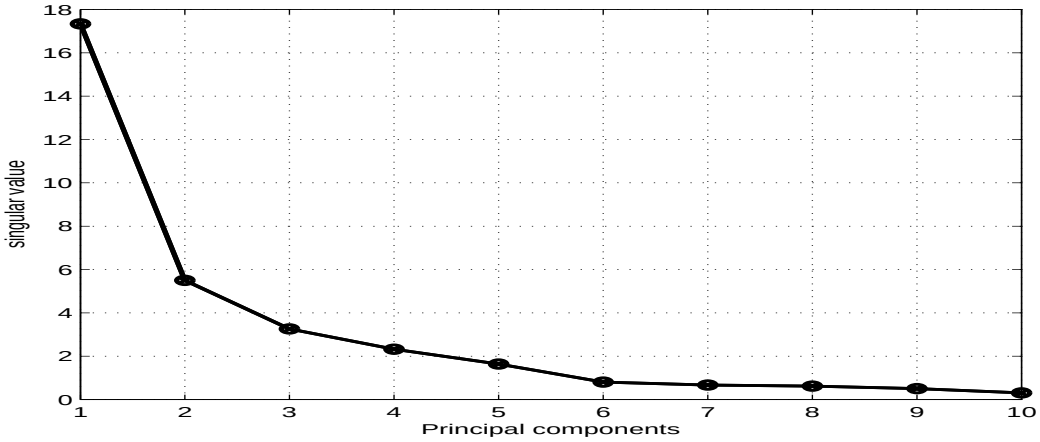

Figure 7.1: Singular values from the seasonal human influenza A/H3N2 dataset.

kernel [178, 179]. Applying kernel PCA to influenza sequence data should should help to reveal the hidden information at a finer resolution.

7.7 Final remarks

The rapid accumulation of the influenza genetic sequences in flu sequence databases afford us a great opportunity to understand the evolution of the influenza virus. By using the large number of genetic sequence data, prediction of new epidemics can be achieved and realized in the near future. In addition, new hypotheses about the evolution of influenza virus can also be developed or generated by studying these vast amount of genetic sequence data. The vaccine strain selection problem can also be greatly benefited from studying the signals hidden in the sequences. Although the future evolutionary paths of the seasonal human influenza viruses are not completely known, the continuous

surveillance of influenza virus in mammalian and in avian species should be able to bring us a step closer in seeing any emerging trends or variants in the near future.

References

[1] Robert G. Webster. The importance of animal influenza for human disease. *Vaccine*, 20 Suppl 2:S16–S20, May 2002.

[2] Don Noah and George Fidas. The global infectious disease threat and its implications for the united states. Technical report, DTIC Document, 2000.

[3] Benjamin P Blackburne, Alan J Hay, and Richard A Goldstein. Changing selective pressure during antigenic changes in human influenza h3. *PLoS pathogens*, 4(5):e1000058, 2008.

[4] 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.

[5] Ahmed S. Abdel-Moneim, Manal A. Afifi, and Magdy F. El-Kady. Genetic drift evolution under vaccination pressure among h5n1 egyptian isolates. *Virol J*, 8:283, 2011.

[6] 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.

[7] W. J. Bean, M. Schell, J. Katz, Y. Kawaoka, C. Naeve, O. Gorman, and R. G. Webster. Evolution of the h3 influenza virus hemagglutinin from human and nonhuman hosts. *J Virol*, 66(2):1129–1138, Feb 1992.

[8] Susanna C Manrubia and L. Viral evolution. *Physics of Life Reviews*, 3(2):65–92, 2006.

[9] 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.

[10] O. T. Gorman, W. J. Bean, and R. G. Webster. Evolutionary processes in influenza viruses: divergence, rapid evolution, and stasis. *Curr Top Microbiol Immunol*, 176:75–97, 1992.

[11] Yi Guan, Dhanasekaran Vijaykrishna, Justin Bahl, Huachen Zhu, Jia Wang, and Gavin J D. Smith. The emergence of pandemic influenza viruses. *Protein Cell*, 1(1):9–13, Jan 2010.

[12] David A Steinhauer. Role of hemagglutinin cleavage for the pathogenicity of influenza virus. *Virology*, 258(1):1–20, 1999.

[13] David A Steinhauer and John J Skehel. Genetics of influenza viruses. *Annu Rev Genet*, 36:305–332, 2002.

[14] S Nakajima, E Nobusawa, and K Nakajima. Variation in response among individuals to antigenic sites on the ha protein of human influenza virus may be responsible for the emergence of drift strains in the human population. *Virology*, 274(1):220–231, 2000.

[15] E Nobusawa, T Aoyama, H Kato, Y Suzuki, Y Tateno, and K Nakajima. Comparison of complete amino acid sequences and receptor-binding properties among 13 serotypes of hemagglutinins of influenza a viruses. *Virology*, 182(2):475–485, 1991.

[16] Derek J. Smith, Alan S. Lapedes, Jan C. de Jong, Theo M. Bestebroer, Guus F. Rimmelzwaan, Albert D M E. Osterhaus, and Ron A M. Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, Jul 2004.

[17] 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.

[18] Pravina Kitikoon, Martha I. Nelson, Mary Lea Killian, Tavis K. Anderson, Leo Koster, Marie R. Culhane, and Amy L. Vincent. Genotype patterns of contemporary reassorted H3N2 virus in U.S. swine. *J Gen Virol*, Mar 2013.

[19] Pravina Kitikoon, Martha I Nelson, Mary Lea Killian, Tavis K Anderson, Leo Koster, Marie R Culhane, and Amy L Vincent. Genotype patterns of contemporary reassorted h3n2 virus in us swine. *Journal of General Virology*, 2013.

[20] Amy L Vincent, Sabrina L Swenson, Kelly M Lager, Phillip C Gauger, Christina Loiacono, and Yan Zhang. Characterization of an influenza a virus isolated from pigs during an outbreak of respiratory disease in swine and people during a county fair in the united states. *Veterinary microbiology*, 137(1):51–59, 2009.

[21] Jean Lindenmann. Origin of the terms' antibody'and'antigen'. *Scandinavian journal of immunology*, 19(4):281–285, 1984.

[22] Y. Suzuki and M. Nei. Origin and evolution of influenza virus hemagglutinin genes. *Mol. Biol. Evol.*, 19(4):501–509, 2002.

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

[24] D. C. Wiley, I. A. Wilson, and J. J. Skehel. Structural identification of the antibody-binding sites of hong kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*, 289(5796):373–378, Jan 1981.

[25] Ji-Ming Chen, Ying-Xue Sun, Ji-Wang Chen, Shuo Liu, Jian-Min Yu, Chao-Jian Shen, Xiang-Dong Sun, and Dong Peng. Panorama phylogenetic diversity and distribution of type a influenza viruses based on their six internal gene sequences. *Virol J*, 6:137, 2009.

[26] Shuo Liu, Kang Ji, Jiming Chen, Di Tai, Wenming Jiang, Guangyu Hou, Jie Chen, Jinping Li, and Baoxu Huang. Panorama phylogenetic diversity and distribution of type a influenza virus. *PLoS One*, 4(3):e5022, 2009.

[27] Amy L Vincent, Wenjun Ma, Kelly M Lager, Bruce H Janke, Richard J Webby, Adolfo García-Sastre, and Jürgen A Richt. Efficacy of intranasal administration of a truncated ns1 modified live influenza virus vaccine in swine. *Vaccine*, 25(47):7999–8009, 2007.

[28] Natalie Pica, Ryan A Langlois, Florian Krammer, Irina Margine, and Peter Palese. Ns1-truncated live attenuated virus vaccine provides robust protection to aged mice from viral challenge. *Journal of virology*, 86(19):10293–10301, 2012.

[29] John Steel, Anice C Lowen, Lindomar Pena, Matthew Angel, Alicia Solórzano, Randy Albrecht, Daniel R Perez, Adolfo García-Sastre, and Peter Palese. Live attenuated influenza viruses containing ns1 truncations as vaccine candidates against h5n1 highly pathogenic avian influenza. *Journal of virology*, 83(4):1742–1753, 2009.

[30] C Scholtissek, S Ludwig, and WM Fitch. Analysis of influenza a virus nucleoproteins for the assessment of molecular genetic mechanisms leading to new phylogenetic virus lineages. *Archives of virology*, 131(3-4):237–250, 1993.

[31] TOSHIHIRO Ito, OWEN T Gorman, YOSHIHIRO Kawaoka, WILLIAM J Bean, and ROBERT G Webster. Evolutionary analysis of the influenza a virus m gene with comparison of the m1 and m2 proteins. *Journal of virology*, 65(10):5491–5498, 1991.

[32] David M Morens, Jeffery K Taubenberger, and Anthony S Fauci. The persistent legacy of the 1918 influenza virus. *New England Journal of Medicine*, 361(3):225–229, 2009.

[33] Shanta M Zimmer and Donald S Burke. Historical perspective of emergence of influenza A (H1N1) viruses. *New England Journal of Medicine*, 361(3):279–285, 2009.

[34] Rebecca J Garten, C Todd Davis, Colin A Russell, Bo Shu, Stephen Lindstrom, Amanda Balish, Wendy M Sessions, Xiyan Xu, Eugene Skepner, Varough Deyde, et al. Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans. *Science*, 325(5937):197–201, 2009.

[35] A. J. Hay, V. Gregory, A. R. Douglas, and Y. P. Lin. The evolution of human influenza viruses. *Philos Trans R Soc Lond B Biol Sci*, 356(1416):1861–1870, Dec 2001.

[36] Paul A Rota, Teresa R Wallis, Maurice W Harmon, Jennifer S Rota, Alan P Kendal, and Kuniaki Nerome. Cocirculation of two distinct evolutionary lineages of influenza type b virus since 1983. *Virology*, 175(1):59–68, 1990.

[37] Jonathan A McCullers, Takehiko Saito, and Amy R Iverson. Multiple genotypes of influenza b virus circulated between 1979 and 2003. *Journal of virology*, 78(23):12817–12828, 2004.

[38] Eri Nobusawa and Katsuhiko Sato. Comparison of the mutation rates of human influenza a and b viruses. *Journal of virology*, 80(7):3675–3678, 2006.

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

[40] World Health Organization. Avian influenza: assessing the pandemic threat. 2005.

[41] Andrew Burns, Dominique Van der Mensbrugghe, and Hans Timmer. *Evaluating the economic consequences of avian influenza*. World Bank, 2006.

[42] Chandrakant Lahariya, AK Sharma, and SK Pradhan. Avian flu and possible human pandemic. *Indian pediatrics*, 43(4):317, 2006.

[43] M Tollis and LD Trani. Recent developments in avian influenza research: epidemiology and immunoprophylaxis. *The Veterinary Journal*, 164(3):202–215, 2002.

[44] I Capua and DJ Alexander. Avian influenza infection in birds: a challenge and opportunity for the poultry veterinarian. *Poultry science*, 88(4):842–846, 2009.

[45] 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.

[46] S Marangon, M Cecchinato, and I Capua. Use of vaccination in avian influenza control and eradication. *Zoonoses and public health*, 55(1):65–72, 2008.

[47] DL Suarez, CW Lee, DE Swayne, A Schudel, M Lombard, et al. Avian influenza vaccination in north america: strategies and difficulties. In *OIE/FAO international scientific conference on avian influenza, Paris, France, 7-8 April, 2005.*, pages 117–124. S Karger AG, 2006.

[48] DE Swayne and BL Akey. Avian influenza control strategies in the united states of america. *Frontis*, 8:113–130, 2005.

[49] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Advances in knowledge discovery and data mining. 1996.

[50] David E Heckerman, David Maxwell Chickering, Usama M Fayyad, and Christopher A Meek. Method and system for visualization of clusters and classifications, April 10 2001. US Patent 6,216,134.

[51] MATLAB Users Guide. The mathworks. *Inc., Natick, MA*, 5, 1998.

[52] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[53] Zhipeng Cai, Tong Zhang, and Xiu-Feng Wan. A computational framework for influenza antigenic cartography. *PLoS computational biology*, 6(10):e1000949, 2010.

[54] Joshua B. Plotkin, Jonathan Dushoff, and Simon A. Levin. Hemagglutinin sequence clusters and the antigenic evolution of influenza a virus. *Proc Natl Acad Sci U S A*, 99(9):6263–6268, Apr 2002.

[55] Wilfred Ndifon, Jonathan Dushoff, and Simon A Levin. On the use of hemagglutination-inhibition for influenza surveillance: surveillance data are predictive of influenza vaccine effectiveness. *Vaccine*, 27(18):2447–2452, 2009.

[56] Jonas E Salk. A critique of serologic methods for the study of influenza viruses. *Archives of Virology*, 4(4):476–484, 1951.

[57] Vishal Gupta, David J. Earl, and Michael W. Deem. Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine*, 24(18):3881–3888, May 2006.

[58] Keyao Pan, Krystina C Subieta, and Michael W Deem. A novel sequence-based antigenic distance measure for h1n1, with application to vaccine effectiveness and the selection of vaccine strains. *Protein Engineering Design and Selection*, 24(3):291–299, 2011.

[59] J Lamar Barnett, Jialiang Yang, Zhipeng Cai, Tong Zhang, and Xiu-Feng Wan. Antigenmap 3d: an online antigenic cartography resource. *Bioinformatics*, 28(9):1292–1293, 2012.

[60] Zhipeng Cai, Tong Zhang, and Xiu-Feng Wan. Concepts and applications for influenza antigenic cartography. *Influenza and other respiratory viruses*, 5(Suppl 1):204, 2011.

[61] Ron AM Fouchier and Derek J Smith. Use of antigenic cartography in vaccine seed strain selection. *Avian diseases*, 54(s1):220–223, 2010.

[62] JA Mumford. Vaccines and viral antigenic diversity. *Revue scientifique et technique (International Office of Epizootics)*, 26(1):69–90, 2007.

[63] A. Lapedes and R. Farber. The geometry of shape space: Application to influenza. *Journal of Theor. Biol.*, 212(1):57–69, September 2001.

[64] Zhipeng Cai, Tong Zhang, and Xiu-Feng Wan. Antigenic distance measurements for seasonal influenza vaccine selection. *Vaccine*, 30(2):448–453, 2012.

[65] 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.

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

[67] David W Mount. Sequence and genome analysis. *Bioinformatics: Cold Spring Harbour Laboratory Press: Cold Spring Harbour*, 2, 2004.

[68] Alessio Lorusso, Amy L. Vincent, Marie E. Gramer, Kelly M. Lager, and Janice R. Ciacci-Zanella. Contemporary epidemiology of north american lineage triple reassortant influenza a viruses in pigs. *Curr Top Microbiol Immunol*, 370:113–132, 2013.

[69] Krista J Howden, Egan J Brockhoff, Francois D Caya, Laura J McLeod, Martin Lavoie, Joan D Ing, Janet M Bystrom, Soren Alexandersen, John M Pasick, Yohannes Berhane, et al. An investigation into human pandemic influenza virus (h1n1) 2009 on an alberta swine farm. *The Canadian Veterinary Journal*, 50(11):1153, 2009.

[70] SJ Flint, LW Enquist, VR Racaniello, and AM Skalka. Principles of virology, 2004.

[71] TAKASHI GoJOBORI, Etsuko N Moriyama, and MOToo KIMuRA. Molecular clock of viral evolution, and the neutral theory. *Proceedings of the National Academy of Sciences*, 87(24):10015–10018, 1990.

[72] W. M. Fitch, J. M. E. Leiter, X. Li, and R. Palese. Positive darwinian evolution in human influenza a viruses. *Proc. Natl. Acad. Sci. USA*, 88:4270–4274, 1991.

[73] M. Plass and E. Eyras. Differentiated evolutionary rates in alternative exons and the implications for splicing regulation. *BMC Evol. Biol*, 6(50), June 2006.

[74] J. B. Plotkin, J. Dushoff, and S. A. Levin. Hemagglutinin sequence clusters and the antigenic evolution of influenza a virus. *Proc Natl Acad Sci USA*, 99(9):6263–6268, April 2002.

[75] J B. Plotkin and J. Dushoff. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza a virus. *Proc Natl Acad Sci USA*, 100(12):7152–7157, June 2003.

[76] E. Zuckerkandl and L. Pauling. *Molecular disease, evolution, and genetic heterogeneity*. Academic Press, New York, 1962.

[77] R. Chen and E. C. Holmes. Avian influenza virus exhibits rapid evolutionary dynamics. *Mol. Biol. Evol.*, 23:2336–2341, 2006.

[78] T. Gojobori, E.N. Moriyama, and M. Kimura. Molecular clock of viral evolution, and the neutral theory. *Proc Natl Acad. Sci. USA*, 87:10015–10018, 1990.

[79] Michael Worobey. Phylogenetic evidence against evolutionary stasis and natural abiotic reservoirs of influenza a virus. *J of Virology*, 82(7):3769–3774, April 2008.

[80] W.M. Fitch and E. Margoliash. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as model case. *Biochemical Genetics*, 1(1):65–71, June 1967.

[81] M. Nei and S. Kumar. *Molecular evolution and phylogenetics*. Oxford University Press, Oxford, New York, 2000.

[82] R.G. Webster, W.J. Bean, O.T. Gorman, T.M. Chambers, and Y. Kawaoka. Evolution and ecology of influenza a viruses. *Microbiological Reviews*, pages 152–179, March 1992.

[83] W. Laver, G. Air, R. Webster, W. Gerhard, C. Ward, and T. Dopheide. The antigenic sites on influenza virus hemagglutinin. studies on their structure and variation in influenza virus. *Dev. Cell Biol*, 5:295–307, 1980.

[84] A. H. Reid, T. A. Janczewski, R.M. Lourens, A. J. Elliot, R.S. Daniels, C. L. Berry, J. S. Oxford, and J. K. Taubenberger. 1918 influenza pandemic caused by highly conserved viruses with two receptor-binding variants. *Emerging Infectious Diseases*, 9(10), 2003.

[85] I. Eidhammer, I. Jonassen, and W.R. Taylor. *Protein Bioinformatics: An algorithmic approach to sequence and structure analysis*. John Wiley and Sons, 2004.

[86] D. J. Smith, F. Forrest, D. H. Ackley, and A. S. Perelson. Variable efficacy of repeated annual influenza vaccination. *Proc Natl Acad Sci USA*, 96(24):14001–14006, November 1999.

[87] A.J. Drummond, G.K. Nicholls, A.G. Rodrigo, and W. Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161:1307–1320, 2002.

[88] WHO. Epidemic and pandemic alert and response (epr):international response to the distribution of a h2n2 influenza virus for laboratory testing: Risk considered low for laboratory workers and the public. April 2005.

[89] S. J. Flint, L.W. Enquist, V.R. Racaniello, and A.M. Skalka. *Principles of Virology*. ASM press, 2004.

[90] 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.

[91] Wen-Hsiung Li and Dan Graur. Fundamentals of molecular evolution. 1991.

[92] Pietro Lio and Nick Goldman. Models of molecular evolution and phylogeny. *Genome research*, 8(12):1233–1244, 1998.

[93] Roderick DM Page and Edward C Holmes. *Molecular evolution: a phylogenetic approach*. John Wiley & Sons, 2009.

[94] Li WenHsiung et al. *Molecular evolution.* Sinauer Associates Incorporated, 1997.

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

[96] J. I. Sagara, S. Shimizu, T. Kawabata, S. Nakamura, M. Ikeguchi, and K. Shimizu. The use of sequence comparison to detect 'identities' in trna genes. *Nucleic Acids Res*, 26(8):1974–1979, Apr 1998.

[97] Surajit Ray and Thomas B Kepler. Amino acid biophysical properties in the statistical prediction of peptide-mhc class i binding. *Immunome Res*, 3:9, 2007.

[98] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[99] Karl Pearson. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559, 1901.

[100] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[101] Jonathon Shlens. A tutorial on principal component analysis. *Systems Neurobiology Laboratory, University of California at San Diego*, 82, 2005.

[102] Andrew R Webb. *Statistical pattern recognition*. John Wiley & Sons, 2003.

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

[104] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern recognition*. Academic Press, 1999.

[105] Sergios Theodoridis and Konstantinos Koutroumbas. Pattern recognition and neural networks. In *Machine Learning and Its Applications*, pages 169–195, 2001.

[106] Brian S Everitt. *An R and S-PLUS® companion to multivariate analysis*. Springer, 2006.

[107] Michael E Wall, Patricia A Dyck, and Thomas S Brettin. Svdmansingular value decomposition analysis of microarray data. *Bioinformatics*, 17(6):566–568, 2001.

[108] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. *A practical approach to microarray data analysis*, 91, 2003.

[109] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. In a practical approach to microarray data analysis. In *Kluwel. chapter*. Citeseer, 2003.

[110] Gilbert Strang. The fundamental theorem of linear algebra. *American Mathematical Monthly*, pages 848–855, 1993.

[111] Gilbert Strang. *Introduction to linear algebra*. SIAM, 2003.

[112] Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

[113] Ian Cooper and Craig Lorenc. Image compression using singular value decomposition. *College of the Redwoods*, pages 1–22, 2006.

[114] James W Demmel. *Applied numerical linear algebra*. Siam, 1997.

[115] Jody S Hourigan and Lynn V McIndoo. Singular value decomposition. *Lin. Algebra-Maths-45, College of Redwoods*, 1998.

[116] HS Prasantha, HL Shashidhara, and KN Balasubramanya Murthy. Image compression using svd. In *Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on*, volume 3, pages 143–145. IEEE, 2007.

[117] Lloyd N Trefethen and David Bau. Numerical linear algebra. *SIAM, Philadelphia*, 1997.

[118] Youwei Zhang and Laurent El Ghaoui. Large-scale sparse principal component analysis with application to text data. In *NIPS*, volume 24, pages 532–539, 2011.

[119] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

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

[121] 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. The global circulation of seasonal influenza a (h3n2) viruses. *Science*, 320(5874):340–346, 2008.

[122] J-M Chen, Y-J Guo, K-Y Wu, J-F Guo, M Wang, J Dong, Y Zhang, Z Li, and Y-L Shu. Exploration of the emergence of the victoria lineage of influenza b virus. *Archives of virology*, 152(2):415–422, 2007.

[123] Y Kanegae, S Sugita, A Endo, M Ishida, S Senya, K Osako, K Nerome, and A Oya. Evolutionary pattern of the hemagglutinin gene of influenza b viruses isolated in japan: cocirculating lineages in the same epidemic season. *Journal of virology*, 64(6):2860–2865, 1990.

[124] MW Shaw, Xiyan Xu, Yan Li, S Normand, RT Ueki, GY Kunimoto, H Hall, A Klimov, NJ Cox, and K Subbarao. Reappearance and global spread of variants of influenza b. *Victoria*, 2(87):2000–2001, 2002.

[125] YP Lin, V Gregory, M Bennett, and A Hay. Recent changes among human influenza viruses. *Virus research*, 103(1):47–52, 2004.

[126] Eili Y. Klein, Adrian W R. Serohijos, Jeong-Mo Choi, Eugene I. Shakhnovich, and Andrew Pekosz. Influenza a h1n1 pandemic strain evolution–divergence and the potential for antigenic drift variants. *PLoS One*, 9(4):e93632, 2014.

[127] KS Li, Y Guan, J Wang, GJD Smith, KM Xu, L Duan, AP Rahardjo, P Puthavathana, C Buranathai, TD Nguyen, et al. Genesis of a highly pathogenic and potentially pandemic h5n1 influenza virus in eastern asia. *Nature*, 430(6996):209–213, 2004.

[128] Kennedy F Shortridge, Nan Nan Zhou, Yi Guan, Peng Gao, Toshihiro Ito, Yoshihiro Kawaoka, Shantha Kodihalli, Scott Krauss, Deborah Markwell, K Gopal Murti, et al. Characterization of avian h5n1 influenza viruses from poultry in hong kong. *Virology*, 252(2):331–342, 1998.

[129] Yohei Watanabe, Madiha S Ibrahim, Yasuo Suzuki, and Kazuyoshi Ikuta. The changing nature of avian influenza a virus (h5n1). *Trends in microbiology*, 20(1):11–20, 2012.

[130] 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.

[131] Holly Shelton, Guadalupe Ayora-Talavera, Junyuan Ren, Silvia Loureiro, Raymond J Pickles, Wendy S Barclay, and Ian M Jones. Receptor binding profiles of avian influenza virus hemagglutinin subtypes on human cells as a predictor of pandemic potential. *Journal of virology*, 85(4):1875–1880, 2011.

[132] 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.

[133] 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.

[134] 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.

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

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

[137] Margaret Chan. World now at the start of 2009 influenza pandemic, 2009.

[138] Hossein Khiabanian, Vladimir Trifonov, and Raul Rabadan. Reassortment patterns in swine influenza viruses. *PLoS Curr*, 1:RRN1008, 2009.

[139] V. Trifonov, H. Khiabanian, B. Greenbaum, and R. Rabadan. The origin of the recent swine influenza a(h1n1) virus infecting humans. *Euro Surveill*, 14(17), Apr 2009.

[140] Alexey D Neverov, Ksenia V Lezhnina, Alexey S Kondrashov, and Georgii A Bazykin. Intrasubtype reassortments cause adaptive amino acid replacements in h3n2 influenza genes. *PLoS genetics*, 10(1):e1004037, 2014.

[141] S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

[142] Stephen E Lindstrom, Yasuaki Hiromoto, Reiko Nerome, Katsuhiko Omoe, Shigeo Sugita, Yoshinao Yamazaki, Tomoko Takahashi, and Kuniaki Nerome. Phylogenetic analysis of the entire genome of influenza a (h3n2) viruses from japan: evidence for genetic reassortment of the six internal genes. *Journal of virology*, 72(10):8021–8031, 1998.

[143] Stephen E Lindstrom, Nancy J Cox, and Alexander Klimov. Genetic analysis of human h2n2 and early h3n2 influenza viruses, 1957–1972: evidence for genetic divergence and multiple reassortment events. *Virology*, 328(1):101–119, 2004.

[144] Heather L. Forrest and Robert G. Webster. Perspectives on influenza evolution and the role of research. *Anim Health Res Rev*, 11(1):3–18, Jun 2010.

[145] Gabriele A Landolt, Alexander I Karasin, Lynette Phillips, and Christopher W Olsen. Comparison of the pathogenesis of two genetically different h3n2 influenza a viruses in pigs. *Journal of clinical microbiology*, 41(5):1936–1941, 2003.

[146] Alexander I Karasin, Suzanne Carman, and Christopher W Olsen. Identification of human h1n2 and human-swine reassortant h1n2 and h1n1 influenza a viruses among pigs in ontario, canada (2003 to 2005). *Journal of clinical microbiology*, 44(3):1123–1126, 2006.

[147] Victoria Svinti, James A. Cotton, and James O. McInerney. New approaches for unravelling reassortment pathways. *BMC Evol Biol*, 13:1, 2013.

[148] Yoshihiro Kawaoka, Scott Krauss, and Robert G Webster. Avian-to-human transmission of the pb1 gene of influenza a viruses in the 1957 and 1968 pandemics. *Journal of virology*, 63(11):4603–4608, 1989.

[149] Edward C Holmes, Elodie Ghedin, Naomi Miller, Jill Taylor, Yiming Bao, Kirsten St George, Bryan T Grenfell, Steven L Salzberg, Claire M Fraser, David J Lipman, et al. Whole-genome analysis of human influenza a virus reveals multiple persistent lineages and reassortment among recent h3n2 viruses. *PLoS biology*, 3(9):e300, 2005.

[150] Tommy Tsan-Yuk Lam, Chung-Chau Hon, and Julian W Tang. Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical reviews in clinical laboratory sciences*, 47(1):5–49, 2010.

[151] U Chandimal de Silva, Hokuto Tanaka, Shota Nakamura, Naohisa Goto, and Teruo Yasunaga. A comprehensive analysis of reassortment in influenza a virus. *Biol Open*, 1(4):385–390, Apr 2012.

[152] Xiu-Feng Wan, Guorong Chen, Feng Luo, Michael Emch, and Ruben Donis. A quantitative genotype algorithm reflecting h5n1 avian influenza niches. *Bioinformatics*, 23(18):2368–2375, Sep 2007.

[153] Carl Kingsford Niranjan Nagarajan. Giraf: robust, computational identification of, 2010.

[154] Amy L Vincent, Wenjun Ma, Kelly M Lager, Bruce H Janke, and Jürgen A Richt. Swine influenza viruses: a north american perspective. *Advances in virus research*, 72:127–154, 2008.

[155] Nan Nan Zhou, Dennis A Senne, John S Landgraf, Sabrina L Swenson, Gene Erickson, Kurt Rossow, Lin Liu, Kyoung-jin Yoon, Scott Krauss, and Robert G Webster. Genetic reassortment of avian, swine, and human influenza a viruses in american pigs. *Journal of virology*, 73(10):8851–8856, 1999.

[156] A. I. Karasin, M. M. Schutten, L. A. Cooper, C. B. Smith, K. Subbarao, G. A. Anderson, S. Carman, and C. W. Olsen. Genetic characterization of h3n2 influenza viruses isolated from pigs in north america, 1977-1999: evidence for wholly human and reassortant virus genotypes. *Virus Res*, 68(1):71–85, Jun 2000.

[157] Raul Rabadan, Arnold J Levine, and Michael Krasnitz. Non-random reassortment in human influenza a viruses. *Influenza and Other Respiratory Viruses*, 2(1):9–22, 2008.

[158] Timo Lassmann and Erik L L. Sonnhammer. Quality assessment of multiple alignment programs. *FEBS Lett*, 529(1):126–130, Oct 2002.

[159] Mariette F Ducatez, Ben Hause, Evelyn Stigger-Rosser, Daniel Darnell, Cesar Corzo, Kevin Juleen, Randy Simonson, Christy Brockwell-Staats, Adam Rubrum, David Wang, et al. Multiple reassortment between pandemic (h1n1) 2009 and endemic influenza viruses in pigs, united states. *Emerging infectious diseases*, 17(9):1624, 2011.

[160] Amy L Vincent, Sabrina L Swenson, Kelly M Lager, Phillip C Gauger, Christina Loiacono, and Yan Zhang. Characterization of an influenza a virus isolated from pigs during an outbreak of respiratory disease in swine and people during a county fair in the united states. *Veterinary microbiology*, 137(1):51–59, 2009.

[161] Zhixin Feng, Janet Gomez, Andrew S. Bowman, Jianqiang Ye, Li-Ping Long, Sarah W. Nelson, Jialiang Yang, Brigitte Martin, Kun Jia, Jacqueline M. Nolting, Fred Cunningham, Carol Cardona, Jianqiang Zhang, Kyoung-Jin Yoon, Richard D. Slemons, and Xiu-Feng Wan. Antigenic characterization of h3n2 influenza a viruses from ohio agricultural fairs. *J Virol*, May 2013.

[162] Claude E Shannon and Warren Weaver. The mathematical theory of communication (urbana, il. *University of Illinois Press*, 19(7):1, 1949.

[163] Sudhir Kumar, Masatoshi Nei, Joel Dudley, and Koichiro Tamura. Mega: a biologist-centric software for evolutionary analysis of dna and protein sequences. *Briefings in bioinformatics*, 9(4):299–306, 2008.

[164] Koichiro Tamura, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei, and Sudhir Kumar. Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, 28(10):2731–2739, 2011.

[165] Sandeep R P. Kumar, Laure Deflube, Moanaro Biswas, Raghunath Shobana, and Subbiah Elankumaran. Genetic characterization of swine influenza viruses (h3n2) isolated from minnesota in 2006-2007. *Virus Genes*, 43(2):161–176, Oct 2011.

[166] Richard J Webby, Sabrina L Swenson, Scott L Krauss, Philip J Gerrish, Sagar M Goyal, and Robert G Webster. Evolution of swine h3n2 influenza viruses in the united states. *Journal of virology*, 74(18):8243–8251, 2000.

[167] Centers for Disease Control and Prevention (CDC). Influenza a (h3n2) variant virus-related hospitalizations: Ohio, 2012. *MMWR Morb Mortal Wkly Rep*, 61:764–767, Sep 2012.

[168] Centers for Disease Control and Prevention (CDC). Evaluation of rapid influenza diagnostic tests for influenza a (h3n2)v virus and updated case count–united states, 2012. *MMWR Morb Mortal Wkly Rep*, 61(32):619–621, Aug 2012.

[169] Toshihiro Ito, J Nelson SS Couceiro, Sørge Kelm, Linda G Baum, Scott Krauss, Maria R Castrucci, Isabella Donatelli, Hiroshi Kida, James C Paulson, Robert G Webster, et al. Molecular basis for the generation in pigs of influenza a viruses with pandemic potential. *Journal of virology*, 72(9):7367–7373, 1998.

[170] Yasuo Suzuki, Toshihiro Ito, Takashi Suzuki, Robert E Holland, Thomas M Chambers, Makoto Kiso, Hideharu Ishida, and Yoshihiro Kawaoka. Sialic acid species as a determinant of the host range of influenza a viruses. *Journal of virology*, 74(24):11825–11831, 2000.

[171] Suman R Das, Scott E Hensley, Alexandre David, Loren Schmidt, James S Gibbs, Pere Puigbò, William L Ince, Jack R Bennink, and Jonathan W Yewdell. Fitness costs limit influenza a virus hemagglutinin glycosylation as an immune evasion strategy. *Proceedings of the National Academy of Sciences*, 108(51):E1417–E1422, 2011.

[172] Walter Fiers, Marina De Filette, A Birkett, Sabine Neirynck, and W Min Jou. A universal human influenza a vaccine. *Virus research*, 103(1):173–176, 2004.

[173] Walter Gerhard, Krystyna Mozdzanowska, and Darya Zharikova. Prospects for universal influenza virus vaccine. *Emerging infectious diseases*, 12(4):569, 2006.

[174] Sabine Neirynck, Tom Deroo, Xavier Saelens, Peter Vanlandschoot, Willy Min Jou, and Walter Fiers. A universal influenza a vaccine based on the extracellular domain of the m2 protein. *Nature medicine*, 5(10):1157–1163, 1999.

[175] Peter Palese. Making better influenza virus vaccines? *Emerg Infect Dis*, 12(1):61–65, Jan 2006.

[176] Kanta Subbarao and Tomy Joseph. Scientific barriers to developing vaccines against avian influenza viruses. *Nature Reviews Immunology*, 7(4):267–278, 2007.

[177] Bernhard Scholkopf, Sebastian Mika, Chris JC Burges, Philipp Knirsch, K Muller, Gunnar Ratsch, and Alexander J Smola. Input space versus feature space in kernel-based methods. *Neural Networks, IEEE Transactions on*, 10(5):1000–1017, 1999.

[178] Christina S Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Pacific symposium on biocomputing*, volume 7, pages 566–575, 2002.

[179] Christina S Leslie, Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, 2004.