# Analyze Influenza Virus Sequences Using Binary Encoding Approach

Ham Ching Lam*, Srinand Sreevatsan† and Daniel Boley‡

**Abstract**

Capturing mutation patterns of each individual influenza virus sequence is often challenging; in this paper, we demonstrated that using a binary encoding scheme coupled with dimension reduction technique, we were able to capture the intrinsic mutation pattern of the virus. Our approach looks at the variance between sequences instead of the commonly used p-distance or Hamming distance. We first convert the influenza genetic sequences to a binary strings and form a binary sequence alignment matrix and then apply Principal Component Analysis (PCA) to this matrix. PCA also provides identification power to identify reassortant virus by using data projection technique. Due to the sparsity of the binary string, we were able to analyze large volume of influenza sequence data in a very short time. For protein sequences, our scheme also allows the incorporation of biophysical properties of each amino acid. Here, we present various encouraging results from analyzing influenza nucleotide, protein and genome sequences using the proposed approach.

*Keywords*: Influenza virus, Evolution, Binary Encoding, Principal Component Analysis

*Department of Computer Science and Engineering, University of Minnesota, 4-192 Keller Hall, 200 Union Street SE, Minneapolis, MN 55455, United States of America. Email: hamching@cs.umn.edu. Tel no: 612-625-0671.

†College of Veterinary Medicine, University of Minnesota, St.Paul, Minnesota, United States of America. Email: sreev001@umn.edu

‡Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota, United States of America. Email: boley@cs.umn.edu

# 1 Introduction

The influenza A virus is a negative stranded RNA virus with eight gene segments that code for 10 proteins in its genome. It is categorized by the serology and genetics of its two surface glycoproteins hemagglutinin (HA) and neuraminidase (NA). The virus is capable of infecting about twenty five percent of the worldwide human population each year [15]. 16 HA antigenetically distinct subtypes have been isolated from mammalian and avian hosts, with the H3N2 being the most widespread and dominant circulating strain in the human population [9]. Selective pressure exists for the virus to generate immunological escape variants that are antigenetically different and to diversify because immunized hosts are resistant to infection with influenza they have been exposed to for several years [10].

Large effort and vast amount of sequence data have been used together to piece together the evolutionary history of Influenza viruses. The influenza evolutionary tree itself is one of the most popular and powerful tools we have in understanding the evolution of the virus. The continuous evolution of the virus makes it challenging to get a global picture of inter-relation among all these viruses. With an evolutionary tree, we can: (1) obtain a more complete understanding of how and where virus evolved that would help explain how certain changes ended up in certain clade along the evolutionary tree. (2) enable us to more easily decipher what's in the virus samples we already have and to make prediction on what antigenic property we'll find in newly isolated viruses. Evolution turns out to be a good structural framework for understanding influenza virus evolution dynamic [7, 9]. However, with the large number of sequence data continuously being deposited to the influenza database, the data often appears to be clouded, unclear, and even redundant. An approach that can quickly provide an overview of the virus evolution under these challenges is most valuable to influenza analysis.

Our first aim in this paper is to present an alternative sequence representation method that is capable of capturing the intrinsic patterns of mutation of the virus; a second aim is to extract these patterns through a dimension reduction technique. To show the utility

and flexibility of the encoding scheme, we performed influenza sequence analysis to expose avian-host to human-host cross-overs using both nucleotide, protein and genome sequences downloaded from NCBI Influenza database [1].

# 2 Results

In this section, we present various results from applying our encoding scheme to influenza genetic sequences using Principal Component Analysis as the processing algorithm. We illustrate the evolution trajectory of H3N2 virus obtained from using nucleotide sequences. We then provide a global view of all the subtypes of influenza viruses based on their HA surface protein. Next, we give results from integrating biophysical information of each amino acid to enhance the distinguishing feature of each virus sequence. We tested this approach on H3 and H5 subtype viruses. We present results of the predictive power of PCA based on our encoding scheme by detecting reassortant virus and mixed infection virus using complete virus genome sequence.

## 2.1 H3N2 evolution trajectory

Multi-Dimensional Scaling (MDS) was used as a dimension reduction technique by [17] to project genetic and antigenic influenza data to visualize the relationship between strains on a two dimensional plane. MDS must first compute the pairwise distance between strains and then proceed to optimize an objective function to preserve the pairwise distance between strains as best as possible. MDS is often used to provide visualization of influenza clusters to gain a first hand understanding of their evolution trajectory. On the other hand, the same objective can be achieved by using PCA where strains' pairwise distance computations are not needed. To achieve this objective, PCA uses the covariance between each strain and find the new and reduced dimensions to visualize the data (please see Materials and Methods section for more detail on PCA). The results from using the proposed encoding scheme on

nucleotide sequences show that the evolution trajectory of the H3N2 virus produced from Principal Component Analysis (PCA) is the same as that produced from Multi-Dimensional Scaling algorithm when the Euclidean metric was used for pairwise distance calculation between strains. In the PCA case, two dimensions are usually sufficient to explain most of the variability of the data. In Figure 1 we show that it produced a distinctive H3N2 evolution trajectory using H3N2 hemagglutinin (HA) gene nucleotide sequences. We shaded the vaccine strains in black in the figure and also listed them in table 1. Each vaccine strain follows nicely in a chronological manner in the curved pattern (from lower left to lower right) among all other H3N2 strains. This trajectory indicates that H3N2 virus is evolving away from its earliest 1968 isolated strain.

## 2.2 Incorporating amino acid biophysical information

The proposed 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 [14] carried out a study to determine the most suitable biophysical properties to use with unsupervised classifiers [14] 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 nucleotide sequences. We obtained the hydrophobicity values for all the amino acids published from the study conducted by Ray and Kepler [14]. After appending each H-value to the binary string of each amino acid and converting all the protein H3 and H5 sequences into binary strings, PCA was used to provide visualization (Figure 2 and 3) 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 2. 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.

## 2.3 Complete view of all subtypes of influenza viruses

The diversity and distribution of the influenza virus has been studied by [3, 7] by building a panorama of phylogenetic trees. Here, we decided to apply our encoding scheme to all 16 subtypes of the influenza virus hemagglutinin nucleotide sequences totalling 16993 to produce a two dimensional whole view of all subtypes. After converting the hemagglutinin nucleotide sequences to binary strings, we used PCA to project all the subtypes (H1 to H16), obtaining a global view of the virus. In Figure 4, we see a tripod shape with H1N1, H3N2, and H5 each occupying a tripod leg (each of the black triangle designates the earliest of each isolate subtype). All the other subtypes remain in the center of the tripod, showing very little change. This indicates that the three subtypes H1N1, H3N2, and H5 are evolving faster than the other subtypes. Among the 16 subtypes, H13 and H16 are very close to each other. This is in agreement with [7]. On the other hand, H2, H4, H9, H10, and H15 appear to be close to each other. Subtypes H2 and H9 are very close to each other, but phylogenetic analysis indicates that these two subtypes were derived from different lineages. One explanation is that there is small synonymous differences (mutation at nucleotide level but does not change the encoded amino acid) exist between these two subtypes based on sequence level analysis. The lineage different can come from viruses evolving within the same host type (e.g. Human H1N1 and Human H3N2) but with different antigenic property for each lineage. Subtypes H4, H10, and H15 are clustered together in the plot, and phylogenetic analysis from [7] showed that they were derived from the same lineage.

## 2.4 Detecting reassortants

Due to the segmented nature of influenza virus genome (8 individual segments of single stranded RNA that encodes 2 surface proteins and 8 internal proteins), reassortment between influenza viruses are common and can lead to the generation of novel strains of the virus [8]. In fact, pandemic strains have been found to carry gene segments originating from multiple hosts within their genome [13]. Here, we desire to test the predictive power of PCA coupled with our binary encoding scheme with hydrophobicity information incorporated. We wish to identify influenza viruses originating from a single host but carrying gene segments belonging to multiple hosts. Our objective is to see whether PCA is able to identify virus's surface proteins that have gone through reassortment process.

For the first test, we built an artificial reassortant virus (RV) dataset consisting of viruses with surface proteins HA and NA from avian hosts but internal proteins originating from a human host. Each RV genome is constructed by replacing the flu virus's (FV1) human-host HA and NA proteins with avian-host HA and NA proteins. We first pre-computed the principal components using flu virus (FV1) genome sequences whose genes all originated from human host only. Then we projected the reassortant virus (RV) genome sequences containing avian HA and NA genes onto these pre-computed FV1 principal components. From Figure 5, we see that reassortant virus (RV) with proteins originating from human host (circle) are closely "attached" to the human proteins (triangle) of the flu virus(FV1). On the other hand, its surface proteins (Avian HA and Avian NA) are clearly isolated from the surface proteins of human-host origin (Human HA and Human NA).

We performed a second analysis test using a real reassortant virus H3N2 A/SW/CO/77 genome sequence identified in [5] to test the predictive power of our approach. We selected this isolate because its genetic characterization by [5] using phylogenetic trees indicated that 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 6), we first computed the principal

components using field isolates of human origin flu viruses (see Materials and Methods for human virus genomes used) and then projected the A/Swine/CO/77 genome onto these precomputed principal components. We see that the HA and NA proteins of 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 control test (result shown in Figure 7), we selected the H3N2 A/swine/Wisconsin/2/1970 swine virus as the control genome because 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 [5]. This is also to assure that the control strain contains only gene segments from a single host type of swine origin. Based on phylogenetic analysis, A/swine/Wisconsin/2/1970 does not contain foreign host type gene. In this control test, we precomputed 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 (labeled Sw HA) and NA (labeled Sw NA) proteins are clearly distantly apart from the swine origin counterparts. 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.

## 2.5 Mixed infection detection

Using whole genome phylogenetic analysis, a mixed infection was detected in a bald eagle isolate, A/bald eagle/Virginia/Sg-00154/2008, by Ramakrishnan [12] that carries avian H1 and H2 subtypes genetic footprint. Here, we use our approach to analyze the bald eagle strain by first computing the principal components of the H1 and H2 viruses gnomes and then projecting the bald eagle genome onto the precomputed principal components. In Figure 8, black crossess represent the proteins from the bald eagle genome, the triangles are

the avian H1N1 proteins and the plus signs are the avian H2N1 proteins. From the figure, we see that PB1, PB2, and PA genes are separated from the main gene cluster that includes all other genes (HA, NA, M, NP, NS genes). The three viral polymerase (P) subunits: PB1, PB2, and PA genes are essential for RNA and viral replication [11]. They carry distinct genetic footprints that distinguish them from other genes and is shown through the PCA projection analysis in this figure. Besides the separation observed for the P genes, the other key observation is that genes from the eagle isolate are very closely attached to both avian H1 and H2 subtypes. This suggests that the bald eagle strain indeed shares high genetic similarity from the avian H1 and H2 subtypes. To further confirm the finding of mixed infection, we performed a single gene analysis of the bald eagle strain using its NA gene as the target. We collected NA protein sequences from avian H1 and H2 subtypes along with the NA gene of the bald eagle strain to form an input matrix and fed it to the PCA algorithm. From Figure 9, the bald eagle's NA gene lands on the path of the H1 and H2 subtype. This suggests that bald eagle strain's NA gene carries two lineages of N1. This result is in agreement with the finding from [12].

# 3 Materials and Methods

## 3.1 Data

All influenza virus nucleotide, protein, and genome sequences used in this study were downloaded from NCBI Influenza Virus Database [1] as of February 2011. 239 H3N2 HA1 nucleotide sequences were used for the trajectory analysis. H3N2 and H5N1 subtypes HA protein sequences totalling 5708 were used in the analysis presented in section 2.2. 16,993 hemagglutinin nucleotide sequences representing all subtypes of the flu virus were used to obtain the whole view plot of the virus. The majority of sequences were from H1 with 6632 sequences, H3 with 4071 sequences, and H5 with 3088 sequences. For reassortant detection, we selected human host flu genome sequences isolated from early 1970s to 1980s for the experiment test.

This test set consists of genome sequences of strain Port Chalmers: A/Port Chalmers/1/1973, Udorn: A/udorn/1972, and Memphis: A/Memphis/15/1988 (accession numbers available upon request). For the control test, we selected A/swine/Wisconsin/2/1970 genome from NCBI flu genome database. For the mixed infection analysis, we used 88 (11 genome sets) gene sequences from avian H1N1 subtype and 48 (6 genome sets) gene sequences from H2N1 subtype. Both the avian H1N1 and H2N1 subtype sequences were isolated from year 1980 to 1990. The neuraminidase (NA) protein sequence set contains 149 sequences from avian H1 subtype and 27 sequences from H2 subtype. Each influenza virus genome is named by its subtype, host, geographic location, strain number and year. The strain name refers to the virus genome which consists of 8 segments that codes for 10 proteins. In our study, we use the term genome to refer to a collection of 10 protein sequences that belong to one influenza strain. The term "sequence" is used to refer to a biological sequence of either nucleotide or amino acid of each individual protein within a genome. All the sequence accession numbers are available upon request.

## 3.2 Binary encoding

Transforming nucleotide or protein sequence to a feature vector that captures the mutation pattern is the key in determining the evolution trajectory of the influenza virus. Binary encoding approach [16] is simple and has the ability to capture the mutation pattern of the virus. The feature vector is a string of zeros and ones that represents a biological sequence directly. This encoding is an embedding in high-dimensional Euclidean space with the property that the distance between each different "letter", or "nucleotide" or "amino acid" is the same. It also allows one to add almost arbitrary weightings to account for biological effects like hydrophobic vs. hydrophylic amino acids. Using the usual ASCII representation encoding would introduce a biologically meaningless ordering to the individual letters. In addition, if protein sequences are used, our approach allows the incorporation of biophysical properties of each amino acid into each protein sequence which further enhances

the differences between each amino acid. For nucleotide sequences, we encode Adenine (A) to "10000", Guanine (G) to "01000", Cytosine (C) to "00100" and Thymine (T) to "00010" and gap character (-) to "00001". Each nucleotide base is uniquely represented by a 5 digits binary string. For example, to encode a nucleotide sequence of "AGA" and another of "ACA", AGA is encoded as 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0 and ACA is transformed to 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0. 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 one (including gap character) 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 one 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. Once the sequences are converted to binary, these strings are collected into a matrix which can be viewed as an alignment matrix and is fed to the PCA algorithm. Even though the length of the nucleotide sequence has been increased by a factor of 5 and protein sequence by a factor of 22, the sparsity of the representation does not incur a high computational overhead.

## 3.3 Principal Component Analysis

Principal Component Analysis (PCA) is used in all forms of analysis from bioinformatics to computer vision. It is a simple non-parametric method of extracting relevant information from unstructured data sets. The extraction can be viewed as linear 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 based on the sum of squares error criterion. 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 restriction 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 briefly introduce the working mechanism 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 $C = \frac{1}{(m-1)}XX^T$. $C$ 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) [6]. 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 then project the data to these orthonormal basis with $X * V$; The matrix $\Sigma$ is a diagonal matrix that contains the eigenvalues of $C$ which are the variances captured by the orthonormal basis/principal components.

For the H3N2 evolution trajectory analysis, the H3N2 HA nucleotide sequences of the same length were converted to binary strings which yielded a data matrix that can be directly used with PCA algorithm. The first two principal components corresponding to the two largest eigenvalues were then plotted to obtain the trajectory. In section 2.2, H3 and H5 HA protein sequences of the same length were used and converted to binary strings. In section 2.3, all HA nucleotide sequences with the same length were used and converted to binary strings.

For both sections 2.2 and 2.3, PCA were then directly applied to the converted binary strings and the first two principal components were selected for plotting and visualization purposes. In section 2.4, influenza genome (consisted of 10 protein sequences) was converted to binary strings with H-value incorporated. PCA algorithm was then used to find the first two principal components for the training data set (the FV1 genome, human flu virus genome, and A/Swine/Wisconsin/72 genome) and then projecting the testing dataset (RV genome, and A/Swine/CO/77 genome) onto the two principal components.

Here, we illustrate the reassortant identification process using the first test case from section 2.4. The training data FV1 genome was first converted to binary format (to make equal length of all protein sequences, zero are padded at the end) and formed the input matrix. We computed the principal components of this training data matrix using Matlab's princomp function with the svds function as the core matrix solver. Once we obtained the precomputed principal components, the testing data was projected and plotted to visualize the result. A step by step procedure is as follows:

1. Convert each protein gene sequence (training and testing data) to binary format with H-value incorporated.

2. Each binary string is made equal length with zeros padded at the end.

3. Collect converted training data and form the input matrix.

4. Compute principal components of the training data matrix with princomp function.

5. Project the training data onto 2 leading principal components.

6. Project combinded training and testing data onto computed principal components.

7. Plot the projected testing data onto the same plot.

Below is a short matlab code segment illustrates the steps of the reassortant detection process.

```
TR = dlmread('FV1_genome'); % Read in training data
U = dlmread('RV_genome');   % Read in test data
[coeff,score,latent] = princomp(TR); % Get PCA using svds
X = U * coeff; % Project TestData onto precomputed PC coordinate
plot(b(:,1),b(:,2), 'ro'); hold on;% Plot 1st two PCs of TrainData
plot(X(:,1),X(:,2), 'kx'); % Visualize projection result of the TestData
```

To detect potential mixed-infection, a bald eagle genome was used to compared with virus genomes from avian H1N1 and H2N1 subtypes. We converted all the protein gene sequences to binary with H-value incorporated. We first computed the principal components using the avian H1 and H2 subtypes, then projected the bald eagle genome onto the precomputed components. Figure 8, a three dimensional plot was used to better reveal the genetic relationship between the 2 subtypes and the bald eagle strain. For the NA gene analysis, protein sequences were converted to binary with H-value added. PCA algorithm was then applied to the complete input matrix.

We perform all the computation using Matlab 7.6 version software. The PCA results were generated by the princomp function with svds function as the core solver from Matlab's Stats toolbox. We timed the run time of generating the principal components in section 2.3 because large number of sequences were used in this analysis. The dimension of the input matrix is 16,996 by 3612. The run time was 4.8 minutes on a desktop computer equipped with two 3.0 GHz processors and 6GB of memory. If one was to use the NCBI phylogenetic

tree construction tool [1] to build a complete view from all the sequences, it would take much longer time and be restricted to the input data size of 1000 sequences at a time.

# 4 Discussion

In this paper, we have shown that using a flexible encoding scheme to convert influenza virus's nucleotide or protein sequence can enable us to automatically extract unique mutation patterns that carries evolution information of the virus. We have highlighted some analysis results using our approach that are important in the field of influenza sequence analysis. For example, a hidden difficulty when analyzing sequences from each flu season is that we do not know which strains in the data evolved from which other strains in the data. There is no indication or extra information showing the relationship between strains [2]. A pairwise comparison with this uncertainty can give results that could be biased because pairwise comparison implicitly assumes that one virus of the pair is the progenitor of the other [2].

The encoding approach proposed here still involves pairwise comparisons as part of the covariance calculation in PCA, but the encoding scheme introduced here allows PCA to automatically capture the locations of the mutation patterns. This is to say that the location of mutation along the sequence is more important than the pairwise distance information. We have demonstrated this with the plots of H3N2 evolution trajectory using PCA (Figure 1). With PCA, we can quickly examine the variances associated with strains instead of relying on pairwise Hamming or P distance between strains. Usually only a small number of components is needed to capture a large fraction of the total variance. The largest $K$ variances are associated with the first $K$ principal components, and there is usually a precipitous drop off in the variances after the $K$-th as seen in Figure 2 and Figure 3. Therefore, the most interesting dynamic of the data can be captured in the first $K$ dimensions. With influenza virus sequences showing a very high genetic similarity characteristic within subtypes [18], this means that most of the sites carry redundant information and only a portion of the sequence contains vital genetic variation signal. This underlying phenomenon seems to be

tailor-made for PCA. We have shown that after converting the sequences to binary strings, PCA has no problem in capturing the intrinsic pattern of the virus sequence data. Although PCA and MDS yield approximately the same trajectory results, an advantage of using PCA is that PCA carries prediction capability. The prediction power of PCA comes from the fact that one can pre-compute a set of principal components with existing data (or training data) and then project a set of new data (test data) onto the pre-computed principal components. This simple procedure can highlight the differences or similarity between the two data sets. We illustrated this by using it to detect reassortant viruses. To detect reassortant, we precomputed principal components from existing virus dataset that do not contain any mixed-host proteins within its genome. We then project new virus genome dataset suspected to contain reassortant proteins onto the precomputed principal components to detect any outliner or abnormally. Here, we have shown that PCA can quickly identify the mixing of human and avian genes in a virus genome. This aspect of prediction power from PCA is far more useful than using multidimensional scaling approach. On the other hand, using the same projection method, one can reveal the high degree of genetic similarity between different strains; we show this in the mixed infection detection analysis. From these two analyses, we show that it is important and valuable to use complete genome sequences of the viruses to understand the evolution dynamic of the viruses.

Feature representation schemes for amino acids usually employ a simple categorical representation where each amino acid is grouped together according to its pre-defined characteristic. Commonly found groups are charge group, polarity group and structure group. Each amino acid within each group is implicitly regarded as being equidistant from every other amino acid. Only the category of each amino acid is used, while the specifics for each individual amino acid is discarded. To overcome this distance bias introduced by the grouping strategy, we elected to directly incorporate each individual amino acid's property, including the individual identities. In our case, we have shown examples using the hydrophobicity property of amino acids as an extra information since it is one of the key properties relating

to protein binding[4]. The extra information allows for a more accurate representation for each amino acid. Through using PCA, the results are encouraging as only two principal components were enough to capture the hidden pattern of the data.

With the Next-Generation Sequencing (NGS) promises of sequencing DNA at unprecedented speed and production of massive quantity of data, it is imperative that new technique needs to be developed to provide quick and reliable analysis of any sequence data. Here, we believe our approach can be used at the upstream stage of sequence data analysis pipeline to gain insight as to which direction should be continued on in analyzing the available data.

# References


[1] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. The influenza virus resource at the national center for biotechnology information. *J Virol*, 82(2):596–601, Jan 2008.

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

[3] J.-M. Chen, Y.-X. Sun, J.-W. Chen, S. Liu, J.-M. Yu, C.-J. Shen, X.-D. Sun, and D. Peng. Panorama phylogenetic diversity and distribution of type a influenza viruses based on their six internal gene sequences. *Virol J*, 6:137, 2009.

[4] T. Hopp. Computer prediction of protein surface features and antigenic determinants. *Prog Clin Biol Res*, 172B:367–377, 1985.

[5] 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.

[6] V. Klema and A. Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176, 1980.

[7] S. Liu, K. Ji, J. Chen, D. Tai, W. Jiang, G. Hou, J. Chen, J. Li, and B. Huang. Panorama phylogenetic diversity and distribution of type a influenza virus. *PLoS One*, 4(3):e5022, 2009.

[8] C. J. Luke and K. Subbarao. Vaccines for pandemic influenza. *Emerg Infect Dis*, 12(1):66–72, Jan 2006.

[9] M. I. Nelson and E. C. Holmes. The evolution of epidemic influenza. *Nat Rev Genet*, 8(3):196–205, Mar 2007.

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

[11] D. R. Perez and R. O. Donis. Functional analysis of pa binding by influenza a virus pb1: effects on polymerase activity and viral infectivity. *J Virol*, 75(17):8127–8136, Sep 2001.

[12] M. A. Ramakrishnan, Z. J. Tu, S. Singh, A. K. Chockalingam, M. R. Gramer, P. Wang, S. M. Goyal, M. Yang, D. A. Halvorson, and S. Sreevatsan. The feasibility of using high resolution genome sequencing of influenza a viruses to detect mixed infections and quasispecies. *PLoS One*, 4(9):e7105, 2009.

[13] A. Rambaut, O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger, and E. C. Holmes. The genomic and epidemiological dynamics of human influenza a virus. *Nature*, 453(7195):615–619, May 2008.

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

[15] C. A. Russell, T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten, V. Gregory, I. D. Gust, A. W. Hampson, A. J. Hay, A. C. Hurt, J. C. de Jong, A. Kelso, A. I. Klimov, T. Kageyama, N. Komadina, A. S. Lapedes, Y. P. Lin, A. Mosterin, M. Obuchi, T. Odagiri, A. D. M. E. Osterhaus, G. F. Rimmelzwaan, M. W. Shaw, E. Skepner, K. Stohr, M. Tashiro, R. A. M. Fouchier, and D. J. Smith. The global circulation of seasonal influenza a (h3n2) viruses. *Science*, 320(5874):340–346, Apr 2008.

[16] 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.

[17] D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and R. A. M. Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, Jul 2004.

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

Table 1:Vaccine strains shown in black in figure 1.

| Number | Vaccine strain |
|---|---|
| 1 | A/Aichi/1968 |
| 2 | A/Port Chalmers/1/1973 |
| 3 | A/Philippines/2/1982 |
| 4 | A/leningrad/360/1986 |
| 5 | A/Shanghai/11/1987 |
| 6 | A/Beijing/353/1989 |
| 7 | A/Shangdong/9/1993 |
| 8 | A/Johannesburg/33/1994 |
| 9 | A/Sydney/5/1997 |
| 10 | A/Moscow/10/1999 |
| 11 | A/Fujian/411/2002 |
| 12 | A/California/7/2004 |
| 13 | A/Wisconsin/67/2005 |
| 14 | A/Brisbane/10/2007 |
| 15 | A/Perth/16/2009 |

Figure caption

- Figure 1: 1st and 2nd Principal Components plot reveals influenza H3N2 evolution trajectory based on the HA protein. Each black dot is the vaccine strain listed in table 1. The trajectory starts at lower left corner and ends in lower right corner.

- Figure 2: Top: PCA projection of H3 and H5 protein sequences without applying hydrophobicity information. Bottom: A plot of the percent variability explained by each principal component.

- Figure 3: Top: PCA projection of H3 and H5 protein sequences with hydrophobicity information incorporated. Bottom: A plot of the percent variability explained by each principal component.

- Figure 4: A complete view of all subtypes from 1918 to 2009 based on their HA protein sequence. The three active evolving subtypes (H1, H3, and H5) are spread out to each tripod leg indicating their dominance in establishing their own lineage.

- Figure 5: Plot of reassortant virus (RV) genome projected onto principal components computed using flu virus (FV1) genome of human origin. Each dot represents a gene sequence from the genome. RV protines are represented by circules. FV1 proteins are represented by triangles.

- Figure 6: SW/CO/77 genome projected onto principal components computed using human origin flu viruses genomes. Light grey colored open circles represent genes from human host genome; black crosses represent proteins from Sw/CO/77 genome.

- Figure 7: SW/CO/77 genome projected onto principal components computed using swine virus genome as control. Black Crosses represent SW/CO/77 proteins and open circles are proteins from the control genome.

- Figure 8: Bald eagle strain (black cross) against avian H1N1 (triangle) and H2N1 (plus) substypes. The strong overlapping of shapes suggests the high genetic similarity between the bald eagle strain and the two subtypes.

- Figure 9: NA gene plot of bald eagle (black cross), avian H1 (circle) and avian H2 (triangle) subtypes. The NA gene of bald eagle strain falls in the path of the H1 and H2 subtype.

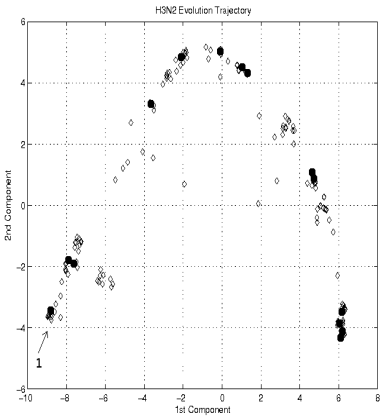


Figure 1

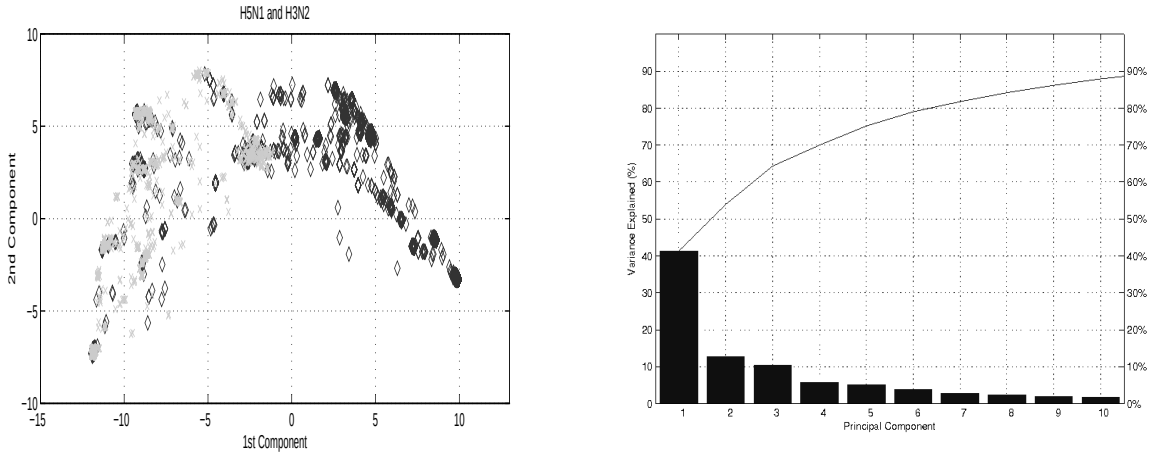
H5N1 and H3N2
2nd Component
1st Component
Variance Explained (%)
Principal Component

Figure 2

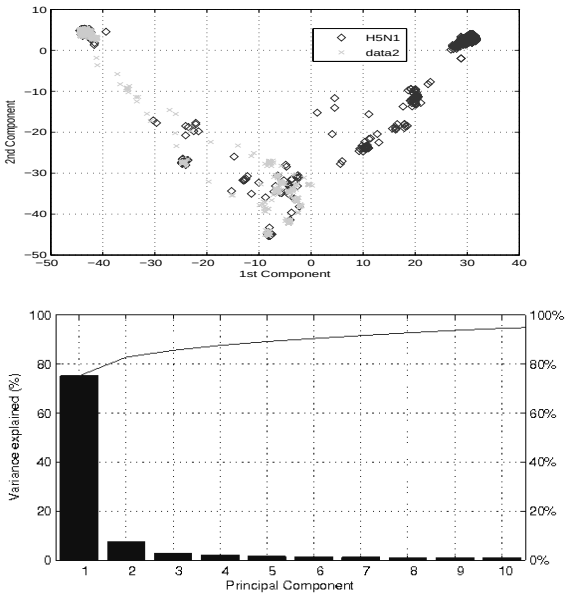
H5N1
data2
2nd Component
1st Component
Variance explained (%)
Principal Component

Figure 3

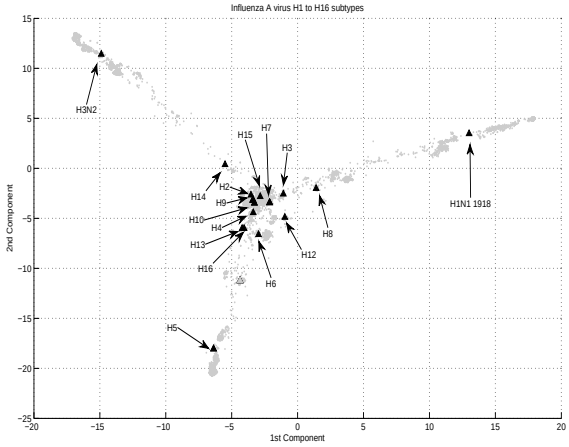
Influenza A virus H1 to H16 subtypes
2nd Component
1st Component
H3N2
H15
H7
H3
H14
H2
H9
H10
H4
H13
H16
H6
H12
H8
H1N1 1918
H5

Figure 4

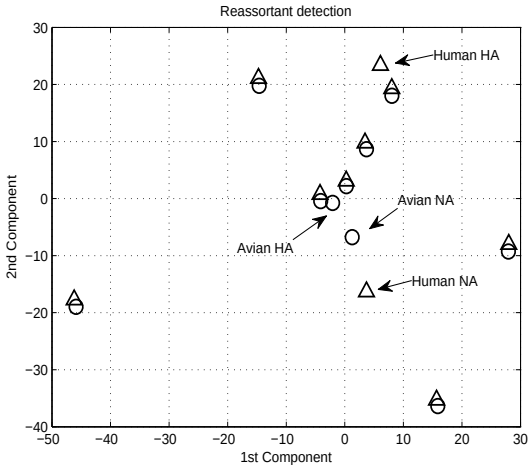
Reassortant detection
Human HA
Avian NA
Avian HA
Human NA
2nd Component
1st Component
30
20
10
0
−10
−20
−30
−40
−50
−40
−30
−20
−10
0
10
20
30

Figure 5

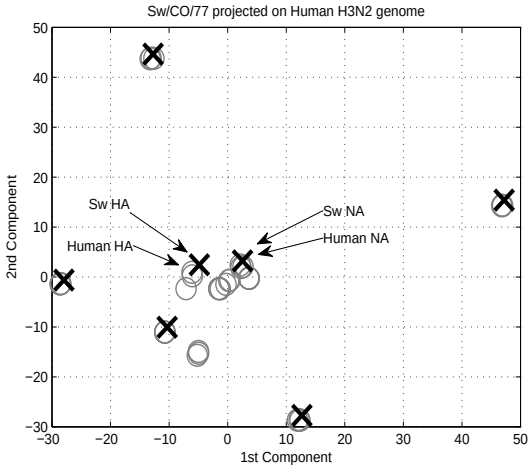
Sw/CO/77 projected on Human H3N2 genome
2nd Component
1st Component
Sw HA
Human HA
Sw NA
Human NA
50
40
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30
40
50

Figure 6

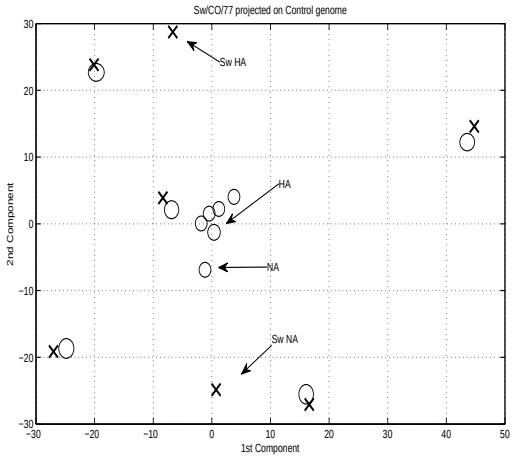
Sw/CO/77 projected on Control genome
Sw HA
HA
NA
Sw NA
2nd Component
1st Component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30
40
50

Figure 7

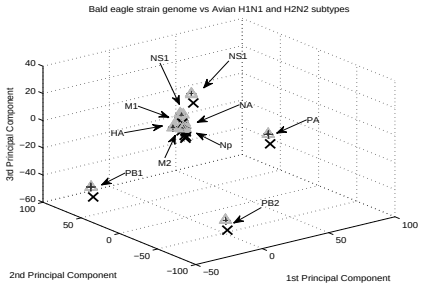
Bald eagle strain genome vs Avian H1N1 and H2N2 subtypes
NS1
NS1
M1
NA
HA
PA
Np
M2
PB1
PB2
3rd Principal Component
2nd Principal Component
1st Principal Component

Figure 8

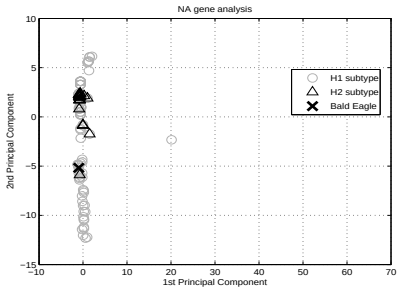
NA gene analysis
H1 subtype
H2 subtype
Bald Eagle
2nd Principal Component
1st Principal Component
10
5
0
−5
−10
−15
−10
0
10
20
30
40
50
60
70

Figure 9