High Throughput Genetic Sequence Analysis

Ham Ching Lam
Department of Computer
Science and Engineering
University of Minnesota
Minneapolis MN 55455, USA
hamching@cs.umn.edu

Steve Cunningham
University of Minnesota
Minneapolis MN 55455, USA
cunni310@umn.edu

Srinand Sreevatsan
Veterinary Population
Medicine
University of Minnesota
Saint Paul MN 55108, USA
sreev001@umn.edu

Daniel Boley
Department of Computer
Science and Engineering
University of Minnesota
Minneapolis MN 55455, USA
boley@cs.umn.edu

1. INTRODUCTION

We present an application paradigm in which an unsupervised machine learning approach is applied to high dimensional influenza sequence datasets: (1) human A/H3N2, (2) avian H5, and (3) North American swine influenza H3N2 virus. Interesting visual patterns observed in the A/H3N2 influenza virus led us to hypothesize that vaccination could be one of the driving forces in the evolution of the human A/H3N2 influenza virus. We provide simulation study and statistical results to support our finding that the influenza virus evolves differently in a protected environment than it evolves in the wild. In the swine H3N2 case, our result suggests that the diversification of North American swine influenza virus can be attributed to the mutations at two positively selected sites on the hemagglutinin protein.

2. MATERIAL AND METHODS

We downloaded 239 human influenza A/H3N2, 288 avian influenza H5, and 883 North American swine influenza H3N2 hemagglutinin (HA) nucleotide and protein sequences from the NCBI influenza database [2]. Noncomplete sequences were discarded from the datasets. Human A/H3N2 and avian H5 HA sequences are from year 1968 to year 2010 and swine H3N2 sequences are from year 1999 to year 2013. Binary conversion method was used to convert genetic sequence into string of bits of 0 and 1. A similar conversion method has been used in protein sequence homology detection in [4]. After sequence conversion, Principal Component Analysis (PCA) [3] was applied to extract the dominant variation from the dataset. We also computed a weighted PCA for the swine influenza dataset. The weight ω which measures the variability of each HA position is obtained by computing its columnwise Shannon entropy $H(Y)$ [5] 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: $P_{ij} = \frac{1}{m}\sum_{l=1}^{m} I(Y_{lj} = y_i)$, where Y_{ij} is amino acid in strain l position j, and $I(\dot{})$ 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 $j = 1, ..., n$. A high variability position may indicate an antibody binding site that is under immune pressure. A low variability position may correlate to a 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 $\omega_j, j = 1, ..n$. We performed a statistical analysis based on multiclass scatter matrix computation and class labels randomization using the projected data points as coordinates of the viruses and the year of isolation of each virus as class label. The multiclass scatter matrix involves the computation of the between-class scatter matrix ($\mathbf{B}$) and the within-class scatter matrix ($\mathbf{W}$). The class separateness measure λ_o is the ratio of trace $\mathbf{B}$ over trace $\mathbf{W}$. 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 [1]. The procedure is listed as Alg. I.

```
Alg. I: Estimate Separateness Measures:
Let $\lambda_o = \frac{tr(B_o)}{tr(W_o)}$ be the observed separateness value.
Repeat $j = 1 : 1000$:
    Repeat $i = 1 : 1000$:
        Randomize class labels and obtain $B_i$ and $W_i$.
        Compute separateness measure: $\lambda_i = \frac{tr(B_i)}{tr(W_i)}$.
    Fit a gamma distribution to $\{\lambda_i\}_{i=1}^{1000}$
    obtaining gamma distribution parameters $\alpha_j$ and $\beta_j$.
    Compute distance of $\lambda_o$ from the mean of the gamma
    distribution, measured by # standard deviations.
```

3. RESULTS

We observed the human A/H3N2 viruses 1 clustered around vaccine seed strains chronologically since their introduction into humans in 1968, even though the dates were not en-

ACM International Conference on Computing Frontiers May 20-22, 2014
Cagliari, Italy

http://dx.doi.org/10.1145/2597917.2597957 ...$15.00.

coded in the data. The same chronological order cannot be observed in the avian H5 viruses 2. The isolation year of the virus was used only as a cluster label when computing the cluster scatter for seasonal A/H3N2 virus and wild type avian H5 influenza virus. The hypothesis is that seasonal A/H3N2 viruses tended to cluster toward the yearly vaccine seed strain for each season, thus leading to the unique chronological evolution clusters observed in the PCA plot. Following this observation, we computed the class or clusters separateness λ_o of both A/H3N2 and the wild type H5 viruses within the PCA two dimensional space using the multi-class scatter matrix computation formulation. The computed λ_o for human A/H3N2 virus is 30.52 and for avian H5 virus is 2.8. These measures were compared to those from 1000 randomizations of the labels. The result from the 1000 runs showed that the observed separateness measure λ_o of human A/H3N2 influenza virus (vaccinated sample) was consistently at about 100 standard deviations away from the mean of gamma distribution and the observed separateness measure λ_o of avian H5 influenza virus (unvaccinated sample) is consistently at about 10 standard deviations away from the mean. The area under the tail of the distributions beyond the observed separateness values was below rounding error which made the computation of p-value not possible. Five distinct clusters of the North American swine influenza H3N2 virus are well separated in the PCA 3D space (Figure 3). The Shannon Entropy weighting scheme yielded a weighted PCA that revealed a cluster signature of each individual cluster consisting of the combination of residues 142 and 144 of the HA gene. We applied the weight matrix (using weight values from positions 142 and 144 exclusively) to the 2013 North American swine H3N2 virus sequences (cyan) and projected it onto the PCA clusters which was pre-computed using prior years data up to 2012. The result (Figure 3) demonstrated that 2013 North American swine H3N2 virus segregated into the correct clusters based on their amino acid residues at positions 142 and 144.

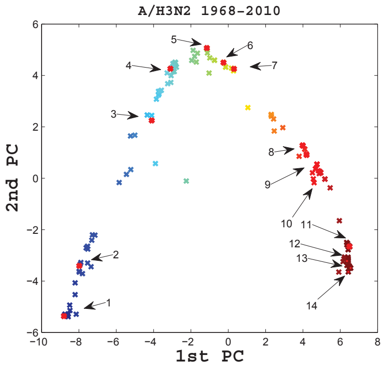

Figure 1: Human influenza A/H3N2 virus clusters. Clusters with vaccine seed strain (pointer) as cluster center can be observed along the evolution path.

4. ACKNOWLEDGMENTS

This research was supported in part by NSF grants IIS-0916750, IIS 1319749. Influenza research in Srinand Sreevatsan lab is funded by the National Institute of Allergy and Infectious Diseases of NIH under Contract No. HHS N266200700007C. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or NSF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

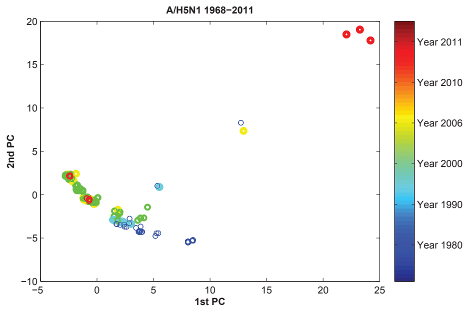

Figure 2: Avian influenza H5 virus.

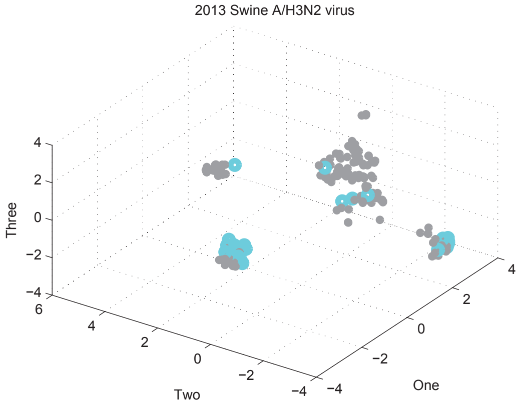

Figure 3: North American swine A/H3N2 virus clusters up to 2012. Cyan ●: overlay of 2013 data.

5. REFERENCES

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

[2] 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.

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

[4] 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.

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