Designing Better Influenza A Vaccines
with the Power of Data Science

Gianna August
Bellevue University
Bellevue, NE 68005, USA
gaugust@my365.bellevue.edu

Pushkar Chougule
Bellevue University
Bellevue, NE 68005, USA
pchougule@my365.bellevue.edu

Amy Femal
Bellevue University
Bellevue, NE 68005, USA
afemal@my365.bellevue.edu

Carlos Palomo
Bellevue University
Bellevue, NE 68005, USA
cpalomo@my365.bellevue.edu

Jonathan Renstrom
Bellevue University
Bellevue, NE 68005, USA
jrenstrom@my365.bellevue.edu

Abstract

The current COVID-19 outbreak generated a curiosity in our group to know about another commonly occurring but a major disease – Flu. From the information available about the impact of Flu, WHO estimates that around 290,000 to 650,000 deaths are caused by the Flu globally, every year. According to the CDC, in the United states alone, there are 12,000 to 61,000 deaths associated with Flu each year. Studying the data available for the effectiveness of existing flu vaccines, the rates vary between 20% - 60% over the last 10 years, with the majority of times around the 40% mark. The cell receptors of Influenza A undergo a high rate of mutation that can compromise the adaptive immune systems acquired immunity. This is one of the major factors for the effectiveness rate of vaccines being relatively low. Vaccines that target the proper viral epitopes can prevent infection. Unfortunately, Influenza vaccines are educated guesses based on prior viral strains. As a part of this paper, we aim at Studying the power of Data Science to predict the viral mutations and in turn help improve the accuracy of influenza vaccines.

Author Keywords

Bioinformatics; Neural Networks; Influenza; Decision Tree; Vaccine; Mutation Prediction; Machine Learning

ACM Classification Keywords

G.4 Mathematical Software: Algorithm design and analysis; I.2.1 Applications and Expert Systems (H.4, J): Medicine and science; I.5.1 Models: Neural nets, Statistical; I.5.2 Design Methodology: Classifier design and evaluation, Pattern analysis; J.3 Life and Medical sciences: Biology and genetics, Health, Medical information systems

Introduction

The deoxyribonucleic acid (DNA) strands are used as a template to create the ribonucleic acid (RNA) in a process known as transcription. However, unlike DNA, RNA is often found as a single-strand. One type of RNA is the messenger RNA (mRNA) which carries information from the ribosome, which is where the protein is synthesized. The sequence of mRNA is what specifies the sequence of amino acids which form the protein. DNA and RNA are also the main component of viruses. Some of the viruses are DNA-based, while others are RNA-based such as Newcastle, HIV, and flu. RNA viruses are different from DNA-based viruses in the sense that they have higher mutation rates, and hence, they have higher adaptive capacity. This mutation causes a continuous evolution that leads to host immunity, which in turn makes the virus become even more virulent [Salama, M 2016].

Influenza A is a serious illness in the United States that kills thousands every year. As a result, an annual influenza vaccination is recommended for young children, elderly people, other individuals at high risk for serious influenza-related complications, and those in close contact of these groups. Antigenic drift, which are small mutations in the genes of influenza viruses that can lead to changes in the surface proteins of the virus, necessitates frequent changes in the composition of influenza vaccines, and these changes must be specified 7–9 months in advance of the influenza season to allow for the production and distribution of vaccines [Belongia, 2009].

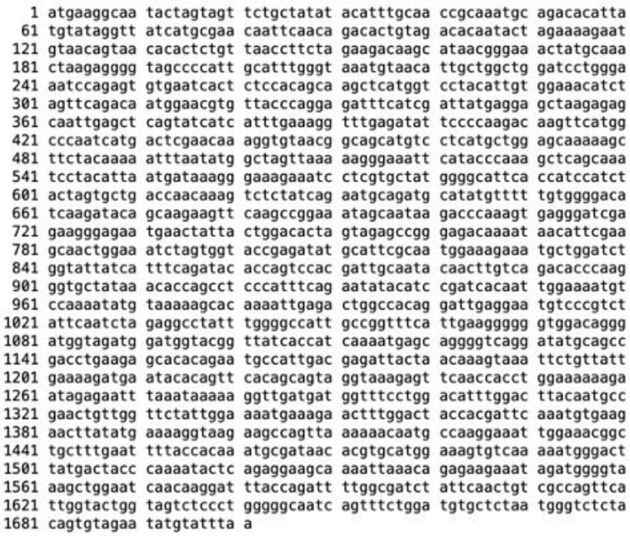
One of the important focuses in the field of human disease genetics is the prediction of genetic mutation. Information of the current virus generations and their past evolution could provide a general understanding of the dynamics of virus evolution and the prediction of future viruses and diseases. The evolutionary relationship between species is determined by phylogenetic analysis; additionally, it infers the ancestor sequence of these species. These phylogenetic relationships among RNA sequences can help in predicting which sequence might have an equivalent function [Salama, M 2016].

The analysis of the mutation data is very important, and one of the tools used for this purpose is machine learning. Machine learning techniques help predict the effects of non-synonymous single nucleotide polymorphisms on protein stability, function and drug resistance. Some of these techniques that are used in prediction are support vector machines, neural networks and decision trees. These techniques have been utilized to learn the rules describing mutations that affect protein behavior, and use them to infer new relevant mutations that will be resistant to certain drugs [E Cilia, 2014]. Another use is to predict the potential secondary structure formation based on primary structure sequences [Lotfi M 2015, Salama M 2016]. A different direction is to predict the discovery of single nucleotide variants in RNA sequence. Another tool in machine learning is Markov chains, which can describe the relative rates of different nucleotide changes in the RNA sequence. These models consider the RNA sequence to be a string of four discrete states, and hence, track the nucleotide replacements during the evolution of the sequence.

Methods

Data Collection

The viral genome contains eight segments of single-stranded RNA, which encode up to eleven proteins. The influenza virus comprises three types: A, B, and C. Among the three influenza types, the type A virus is the most virulent human pathogen and causes the most severe diseases [Attaluri 2010]. The Influenza receptor proteins, Neuraminidase (11 subtypes) and Hemagglutinin (18 subtypes), are largely responsible for the virulence of a particular influenza A strain. These N-H subtypes are recombinant. One source of data is the National Center for Biotechnology Information (NCBI). Through NCBI nucleotide sequence data for all eight segments, HA, NA, PA, NS, PB1, PB2, M, and NP can be downloaded.



```
   1 atgaaggcaa tactagtagt tctgctatat acatttgcaa ccgcaaatgc agacacatta
  61 tgtataggtt atcatgcgaa caattcaaca gacactgtag acacaatact agaaaagaat
 121 gtaacagtaa cacactctgt taaccttcta gaagacaagc ataacgggaa actatgcaaa
 181 ctaagagggg tagccccatt gcatttgggt aaatgtaaca ttgctggctg gatcctggga
 241 aatccagagt gtgaatcact ctccacagca agctcatggt cctacattgt ggaaacatct
 301 agttcagaca atggaacgtg ttacccagga gatttcatcg attatgagga gctaagagag
 361 caattgagct cagtatcatc atttgaaagg tttgagatat tccccaagac aagttcatgg
 421 cccaatcatg actcgaacaa aggtgtaacg gcagcatgtc ctcatgctgg agcaaaaagc
 481 ttctacaaaa atttaatatg gctagttaaa aagggaaatt catacccaaa gctcagcaaa
 541 tcctacatta atgataaagg gaaagaaatc ctcgtgctat ggggcattca ccatccatct
 601 actagtgctg accaacaaag tctctatcag aatgcagatg catatgtttt tgtggggaca
 661 tcaagataca gcaagaagtt caagccggaa atagcaataa gacccaaagt gagggatcga
 721 gaagggagaa tgaactatta ctggacacta gtagagccgg gagacaaaat aacattcgaa
 781 gcaactggaa atctagtggt accgagatat gcattcgcaa tggaaagaaa tgctggatct
 841 ggtattatca tttcagatac accagtccac gattgcaata caacttgtca gacacccaag
 901 ggtgctataa acaccagcct cccatttcag aatatacatc cgatcacaat tggaaaatgt
 961 ccaaaatatg taaaaagcac aaaattgaga ctggccacag gattgaggaa tgtcccgtct
1021 attcaatcta gaggcctatt tggggccatt gccggtttca ttgaaggggg gtggacaggg
1081 atggtagatg gatggtacgg ttatcaccat caaaatgagc aggggtcagg atatgcagcc
1141 gacctgaaga gcacacagaa tgccattgac gagattacta acaaagtaaa ttctgttatt
1201 gaaaagatga atacacagtt cacagcagta ggtaaagagt tcaaccacct ggaaaaaaga
1261 atagagaatt taaataaaaa ggttgatgat ggtttcctgg acatttggac ttacaatgcc
1321 gaactgttgg ttctattgga aaatgaaaga actttggact accacgattc aaatgtgaag
1381 aacttatatg aaaaggtaag aagccagtta aaaaacaatg ccaaggaaat tggaaacggc
1441 tgctttgaat tttaccacaa atgcgataac acgtgcatgg aaagtgtcaa aaatgggact
1501 tatgactacc caaaatactc agaggaagca aaattaaaca gagaagaaat agatggggta
1561 aagctggaat caacaaggat ttaccagatt ttggcgatct attcaactgt cgccagttca
1621 ttggtactgg tagtctccct gggggcaatc agtttctgga tgtgctctaa tgggtctcta
1681 cagtgtagaa tatgtattta a
```

Fig 1: Sample data of Influenza A virus (H1N1) segment HA from 01/2010

Multiple Sequence Alignment
Multiple sequence alignment is used in order to facilitate the classification analysis to perform better. A sequence alignment is a way of arranging two sequences one by one where the residues under the same column are supposed to have a common evolutionary origin. Through the evolutionary relationship, a set of sequences share a lineage and are descended from a common ancestor [Attaluri 2010]. Multiple sequence alignment is when three or more sequences are involved.

One tool to perform the multiple sequence alignment is MUSCLE, which stands for multiple sequence comparison by log-expectation, and is one of the most popular multiple alignment software for protein and nucleotide sequence [Attaluri 2010]. MUSCLE uses two distance parameters: k-mer and Kimura for a pair of sequences. The K-mer distance is used for an unaligned pair of sequences and Kimura distance is used for an aligned pair of sequences. A k-mer is a contiguous subsequence of length k. Sequences having more k-mers in common tend to be similar to each other. For an aligned pair of sequences, the pairwise identity is computed and converted to distance estimate applying Kimura correction for multiple substitutions at a single site. MUSCLE uses UPGMA for clustering distance matrices. A new profile function is used to apply pairwise alignment to profiles [Attaluri 2010].

Machine Learning
Various machine learning processes can be performed. For this project, the primary goals will be developing a neural network, to aid in prediction, and a decision tree, to aid in classification of influenza subtypes.

Expected Results

Decision Tree

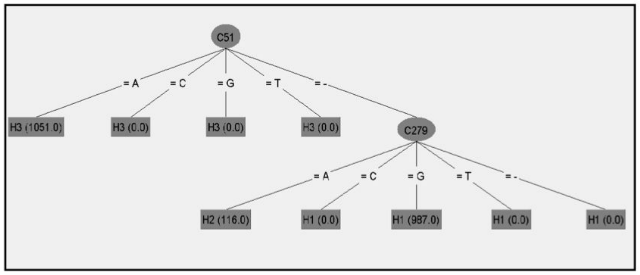


Fig 2: This image of a decision was taken from “Classifying influenza subtypes and hosts using machine learning techniques” Attaluri, 2010

The viral Neuraminidase-Hemagglutinin genotype can be determined based on point mutations within specific sequences of these genes. The aligned sequences from the dataset will be used to train decision tree classifiers. In each of the iteration steps, one or more critical positions, in which different subtypes can be most likely identified, can be determined. These positions can then be collectively utilized to build more precise models for further subtype prediction as well as better understanding which nucleotide positions are conserved, and which positions are under high selective pressure. Neural networks will be used in two ways for this project. The first is to perform classification analysis. This will allow us to have a set of results to compare against the decision tree results [Attaluri 2010].

The second use of neural networks is to predict the probability of nucleotide mutations within the primary RNA sequence. This analysis can determine the mutation probability for the Hemagglutinin-neuraminidase genotype. These receptor binding genes undergo high selective pressure and are largely responsible for the differing virulence of influenza strains. The average rates of nucleotide mutation for specific sequences can be initially determined through analysis of a large data set using WSPmaker to detect the synonymous and nonsynonymous substitutions for every 3 nucleotides that make up a codon within the gene. Approximations of variable rates of mutation can be inferred for chunks of nucleotide sequences within the Hemagglutinin-Neuraminidase genes and used to refine the neural network model by assigning weights based on the probability of mutation for each chunk. Each nucleotide is assigned one of four values. The values are inputted into the network and the weights are modified based on the general difference in rates of mutation for purines vs. pyrimidines. The neural network is non-linearized using the sigmoid function. Multiple generations of the genes are fed through the neural network to predict the probability of mutation within specific sequences. The training process continues until test sequences fed through the network generate a 70% or higher accuracy of prediction for nucleotide mutations. [Attaluri 2010, Salama M 2016].

Secondary structure can also potentially be predicted using thermodynamic algorithms. Existing programs such as RNAMute can determine the effect of mutations on the secondary folding structure of an RNA sequence. Another potential possibility is the use of tensorflow with deep learning tools to examine the relationship between viral genes [Attaluri 2010, Salama M 2016].

The output generated by these predictive models can be used to generate more effective seasonal vaccines with a higher degree of accuracy than exists with conventional models. The average effectiveness of the influenza vaccine is currently 40%, although this varies from year to year. Even a couple percentage points of improvement in vaccine efficacy has the potential to save thousands of lives.

Acknowledgements

We would like to thank our Professor and classmates for providing helpful comments and suggestions for our project. We would also like to thank the many authors who have provided us with the knowledge to make this paper possible. Finally, we would like to thank the brave health care workers who are currently risking their lives fighting the COVID-19 pandemic.

References

1. Anon. 2019. Predicting correct serotypes using machine-learning models based on codon usage patterns of influenza A viruses. (January 2019). Retrieved April 23, 2020 from https://www.biorxiv.org/content/10.1101/528083v1.abstract
2. Anon. 2020. CDC Seasonal Flu Vaccine Effectiveness Studies. (February 2020). Retrieved April 23, 2020 from https://www.cdc.gov/flu/vaccines-work/effectiveness-studies.htm
3. Anon. Classification of Host Origin in Influenza A virus by ... Retrieved April 23, 2020 from http://www.naun.org/main/NAUN/bio/2017/a182010-049.pdf
4. Pavan Attaluri et al. 2009. Applying machine learning techniques to classify H1N1 viral strains occurring in 2009 flu pandemic. (2009).
5. Pavan Attaluri et al. Applying neural networks to classify influenza virus antigenic types and hosts. Retrieved April 23, 2020 from https://ieeexplore.ieee.org/abstract/document/5510726
6. Pavan K. Attaluri, Zhengxin Chen, and Guoqing Lu. 2010. Applying neural networks to classify influenza virus antigenic types and hosts. 2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (2010). DOI: http://dx.doi.org/10.1109/cibcb.2010.5510726
7. Pavan Attaluri. 2010. Classifying influenza subtypes and hosts using machine learning techniques. (May 2010). Retrieved April 22, 2020 from http://ezproxy.bellevue.edu/login?url=https://search-proquest-com.ezproxy.bellevue.edu/docview/219921882?accountid=28125
8. Pavan K. Attaluri, Zhengxin Chen, Aruna M. Weerakoon, and Guoqing Lu. 2009. Integrating Decision Tree and Hidden Markov Model (HMM) for Subtype Prediction of Human Influenza A Virus. (June 2009). Retrieved April 23, 2020 from https://link.springer.com/chapter/10.1007/978-3-642-02298-2_8
9. Edward A. Belongia et al. 2009. Effectiveness of Inactivated Influenza Vaccines Varied Substantially with Antigenic Match from the 2004–2005 Season to the 2006–2007 Season. The Journal of Infectious Diseases 199, 2 (2009), 159–167. DOI:http://dx.doi.org/10.1086/595861
10. Eng CLP et al. Predicting host tropism of influenza A virus proteins using random forest. BMC Medical Genomics. 2014;7:1-11. https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=135645438&site=eds-live.
11. Claire Gillespie. This Is How Many People Die From the Flu Each Year. Retrieved April 23, 2020 from https://www.health.com/condition/cold-flu-sinus/how-many-people-die-of-the-flu-every-year
12. R. Parida, M.s. Shaila, S. Mukherjee, N.r. Chandra, and R. Nayak. 2007. Computational analysis of proteome of H5N1 avian influenza virus to define T cell epitopes with vaccine potential. Vaccine 25, 43 (2007), 7530–7539. DOI:http://dx.doi.org/10.1016/j.vaccine.2007.08.044
13. Mostafa A. Salama et al. 2016. The prediction of virus mutation using neural networks and rough set techniques. EURASIP Journal on Bioinformatics and Systems Biology 2016, 1 (2016). DOI:http://dx.doi.org/10.1186/s13637-016-0042-0
14. Reatha Sandie and Stéphane Aris-Brosou. 2013. Predicting the Emergence of H3N2 Influenza Viruses Reveals Contrasted Modes of Evolution of HA and NA Antigens. Journal of Molecular Evolution 78, 1 (2013), 1–12. DOI:http://dx.doi.org/10.1007/s00239-013-9608-6
15. Nermeen Shaltout et al. 2014. Information Gain as a Feature Selection Method for the Efficient Classification of Influenza Based on Viral Hosts. (2014). Retrieved April 22, 2020 from

16. Lin Tang. 2013.Machine Learning Methods for Annual Influenza Vaccine Update. Masters Thesis. Waterloo University, Ontario,Canada.
17. J.J. Treanor et al. 2012. Effectiveness of Seasonal Influenza Vaccines in the United States During a Season With Circulation of All Three Vaccine Strains. Clinical Infectious Diseases 55, 7 (2012), 951–959. DOI:http://dx.doi.org/10.1093/cid/cis574
18. Jia Wang et al. 2013. Using Amino Acid Factor Scores to Predict Avian-to-Human Transmission of Avian Influenza Viruses: A Machine Learning Study. Protein & Peptide Letters 20, 10 (January 2013), 1115–1121. DOI: http://dx.doi.org/10.2174/0929866511320100005
19. WHO. Coronavirus disease (covid-19) outbreak situation. https://www.who.int/emergencies/diseases/novel-coronavirus-2019
20. Xiaowei Xu et al. Deep learning system to screen coronavirus disease 2019 pneumonia. DOI: https://arxiv.org/ftp/arxiv/papers/2002/2002.09334.pdf
21. Li Yan et al. Prediction of criticality in patients with severe COVID-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. DOI: https://www.medrxiv.org/content/10.1101/2020.02.27.20028027v
22. Shaomin Yan and Guang Wu. 2010. Prediction of Mutation Positions in H5N1 Neuraminidases From Influenza A Virus by Means of Neural Network. Annals of Biomedical Engineering 38, 3 (2010), 984–992. DOI:http://dx.doi.org/10.1007/s10439-010-9907-7
23. Rui Yin, Yu Zhang, Xinrui Zhou, and Chee Keong Kwoh. 2020. Time series computational prediction of vaccines for influenza A H3N2 with recurrent neural networks. Journal of Bioinformatics and Computational Biology (2020). DOI: http://dx.doi.org/10.1142/s0219720020400028