Comparison of CIV, SIV and AIV using Decision Tree and SVM

Hyorin Park[1], Yoojin Park[1], Yerin Moon [1] and Taeseon Yoon[2]

[1]*Natural Science, Hankuk Academy of Foreign Studies, HAFS, Yongin-si, Republic of Korea*
[2] *Department of Computer Science and Engineering, Korea University, KU, Seoul, Republic of Korea*

Abstract. The H3N2, the canine influenza virus has numerous types of animal hosts that can live and reproduce on. They mostly settle on pigs and birds. However, some concerned voices are rising that there is high possibility that humans could be an additional victim for the canine flu. Consequently, our project group expect that the information about the H3N2's DNA are valuable, since the information could attribute to development of vaccine and medicine. In the experiments of analysing the properties of CIV, Canine Influenza Virus with the comparison of SIV, Swine Influenza Virus and AIV, Avian Influenza Virus with the decision tree and SVM, Support Vector Machine. The result came out that CIV, SIV and AIV are alike but also different in some aspects.

1 Introduction

Recently, various types of influenzas have broken out worldwide. The influenzas can be thought of as similar as each other, but can also be thought as the origins of separate diseases. Determining the ranking of them with their impact to human, to seek the effective treatment, maybe the three-AIV, SIV, and CIV- will win the most of all the influenzas. In our experiment, the sample of viruses was composed of Nev, Nef, Tat and Vif each originating from human, swine and avian influenza virus. The assortments of the isolate implies that gene classification exists. [1]

Specific investigation on gene assortments and the 8 particular gene segments and DNA sequences among the H3N2/AIV/SIV/CIV is the core of the report in terms of finding the treatment of H3N2 CIV. The remedy of H3N2 AIV and SIV could be substituted to potential cure of H3N2 CIV. 'SVM' and 'decision tree' algorithm would be used in order to analogue the resulting DNA sequences and similarities of the 3 viruses. The common gene of the 3 viruses includes the fact that swine, avian, human cells are the possible host or incubator of the virus and the restriction of virus from being cloned could result the destruction of the virus's ability.

2 Background knowledge

2.1 Virus

2.1.1 Swine Influenza virus

The H3N2 Swine Influenza Virus(SIV) originates from the H1N1 Influenza A virus. It produces fever, lethargy, sneezing, coughing, difficulty breathing and decreased appetite in pigs. [2] In humans, on the other hand, the symptoms of the 2009 "swine flu" H1N1 virus are similar to those of influenza-like illness in general. Symptoms include fever, cough, sore throat, body aches, headache, chills and fatigue. [3] The 2009 outbreak has also shown an increased percentage of patients reporting diarrhoea and vomiting.

2.1.2 Canine Influenza virus

Canine Influenza Virus(CIV) is well known as one of the most highly pathogenic subtype of the influenza A virus. Symptoms vary from mild to harsh ones, including a cough which lasts for approximately 30 days, possibly a nasal discharge, high fever and - in extreme cases – pneumonia. [4] The vaccine for this virus has been developed in 2009, but the effectiveness or safety has not yet been verified. [5]

2.1.3 Avian Influenza virus

Avian Influenza Virus (AIV) is an influenza caused by viruses adapted to birds, and it is traced back to its original form, the H5N1 AIV, which is noted for its fatalness. Symptoms include fever, cough, sore throat, body aches, headache, chills and fatigue. [6]

2.2 Algorithm

2.2.1 Decision Tree

A decision tree is a tree-like graph or model composed of node and branches. The decision support tool are openly utilized statistical analysis for classification of category type inputs. [7] Several internal nodes contain questions associated with data items, branches down the original source into subsets. Following branches represents the outcome of test and leaf nodes are directed to a class node. [8]

2.2.2 Support vector machine

SVM(Support Vector Machine) is supervised learning model for data analyzing algorithms and it is used to classification and regression analysis. When a group of data is given, SVM training algorithm builds a model that decide which category to involve the data, based on existing data. SVM model is represented as a widest gap, border of separated categories, in space. [9] New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. SVM is used by not only linear classifications but also non-linear classification. For SVM to perform a non-linear classification efficiently, kernel trick is needed to implicitly map their inputs into high-dimensional feature spaces.

2.3 Protein

2.3.1 Nef protein

Nef(Negative Regulatory Factor) protein is a 27-35kDa protein. The protein is encoded by primate lentiviruses which include AIV-1, AIV-2 and SIV. Nef promotes T-cell activation and the establishment of two basic attributes of AIV infections. Nef regulates the cell surface expression of CD4 and Lck. [10] Enrichment of active Lck induces the production of Interleukin 2 which activates the growth, differentiation and proliferation of T-cells. Differentiated T-cells creates a new population of cells in which AIV-1 could infect. In short, Nef protein stands for manipulation of the host's cellular machinery AIV-1 replication in terms of "Negative Factor."

2.3.2 Env protein

Env(Envelop Protein) protein is a protein that forms the viral envelope. The expression of the env gene induces retroviruses to attach to specific cell target and to infiltrate the cell membrane. The env gene codes for the gp160 protein, which is cleaved into gp120 and gp41 by Furin. The glycoprotein gp120 bins to CD4 receptor on the target cell with a receptor including the helper T-cell. Replication cycle of AIV is related with the env gene in which gp120 has been the subject of AIV vaccine research while CD4 receptor binding is an important step in AIV infection. In addition, glycoprotein gp41, which bound to gp120, enables AIV to enter the cell in another step of AIV infection.

2.3.3 Tat protein

In molecular biology, Tat(Trans-Activator of Transcription) is a protein that is encoded for by the tat gene in AIV-1. Tat is a regulatory protein that drastically enhances the efficiency of viral transcription from the LTR promoter and replication. Tat has an unusual property for a transcription factor. It can be released and enter cells freely, yet still retain its activity, enabling it to up-regulate a number of genes.

2.3.4. Vif protein

Vif(Viral infectivity factor) is a protein found in AIV and other retroviruses and is essential for viral replication. It have human enzyme APOBEC(cytidine deaminase enzyme that mutates viral nucleic acids) cause hypermutation of the viral genome, rendering it dead-on-arrival at the next host cell by ubiquitination and cellular degradation . [11] As result, it disrupts antiviral activity. Thus APOBEC plays key role to defend retroviral which AIV-1 has overcome by the acquisition of vif. [12]

3 Method

3.1 Decision Tree

We used decision tree to analyse the overlapped rule among AIV, SIV, and CIV. Dividing the DNA sequence in 10 parts, we tested the particular protein sequence of env, nef, tat, and vif in 5,7,9 window using 10-fold cross validation. Sequences with confidence above 75% were only considered as 'valid'.

3.2 SVM

We used RBF, Polynomial function, normal function, and sigmoid function in this experiment. The X-axis of graph is the kind of function we used, and Y-axis is accuracy. We progressed the experiment in each windows and calculated the average. To test SVM's validity, we compare Accuracy rate. At a higher accuracy, more similarity. [13]

4 Result

4.1 Decision Tree

Decision Tree was used to analyse common rules between DNA sequences of AIV, SIV and CIV. The DNA sequences are divided into 10 subsets and the experiment was held in 10 fold cross validation for each 5 window, 7 window and 9 window rules of specific proteins: ENV, NEF, TAT and VIF. Under all the extracted rules, only those with frequency of over 0.75 were considered as valuable. [14] In the experiment, AIV refers to class 1, SIV refers to class 2 and class 3 represents CIV. In interpreting the result, a particular amino acid has a tendency of being repeatedly observed in a certain position. In the protein ENV, AIV seemed to have the most powerful impact among the three viruses in

classification. In protein NEF and TAT, CIV was dominant under rules. On the other hand, SIV and CIV were the main subjects of classification, so the influence of AIV was small in protein VIF.[15]

Table 1. Rule extraction under 5 window of ENV.

Class	Rule	Frequency
class1	pos 1=G pos 4=M	0.750
class2	pos 1=G pos 4=V	0.800
class3	pos 1=L pos 2=R	0.800

In Table 1, the result represents that AIV has considerably mixed with SIV and CIV. In addition, class 2 and class 3 overlapped common sequences with other classes, so the DNA sequences between three viruses are quite evenly mixed.

Table 2. Rule extraction under 7 window of ENV.

Class	Rule	Frequency
class1	pos 2=A pos 4=F	0.750
class2	pos 4=D pos 5=T	0.750
class3	pos 2=S pos 4=S	0.857/0.833

In Table 2, AIV showed high tendency to follow the rules of class 1 while SIV almost equally combined with other classes. On the other hand, CIV refer to class 2 in relatively high percentage.

Table 3. Rule extraction under 9 window of ENV.

Class	Rule	Frequency
class1	pos 7=Q pos 9=H	0.800
class2	pos 7=Q pos 9=R	0.800
class3	pos 7=Q pos 9=N	0.800

In Table 3, AIV, SIV and CIV have the common amino acid Q in position 7 and the three viruses are mixed equally.

Table 4. Rule extraction under 5 window of NEF.

class	Rule	Frequency
class1	pos 3=V pos 5=Q pos 4=H _	0.750 0.750
class2	pos 3=V pos5=L pos 3=M _	0.750 0.800
class3	pos 3=V pos5=V	0.750

In Table 4, we observed the considerably combined DNA sequences among AIV, SIV and CIV.

Table 5. Rule extraction under 7 window of NEF.

class	Rule	Frequency
class1	pos 1=T _	0.750
class2	pos 1=A pos7=G	0.750
class3	_	_

In Table 5, under certain combining of the three viruses, AIV had high tendency to class 3 and SIV was likely to refer to class 2 in highest percentage.

Table 6. Rule extraction under 9 window of NEF.

class	Rule	Frequency
class1	pos 2=C _	0.750

In Table 6, largest number of appearances were observed in pos 2=C of AIV, pos 2=F of SIV and pos 9=N in CIV. The percentile showed that the three viruses are evenly mixed in protein NEF.

Table 7. Rule extraction under 5 window of TAT.

class	Rule	Frequency
class1	pos1=C pos5=P pos1=H	0.83
class2	pos1=D pos5=P	0.75
class3	pos1=E pos5=P	0.82

In Table 7, AIV, SIV, and CIV share considerably similar amino acid sequence. The first table represents representative rules of each class and their frequency. Second table notice the percentage which AIV,SIV,CIV refers class1,2,3 with tat_window.

Table 8. Rule extraction under 7 window of TAT.

class	rule	frequency
class1	pos6=I pos7=L	0.80
class2	pos6=L pos7=I	0.75
class3	pos2=W pos6=V pos7=K	0.75

In Table 8, AIV is similar with SIV. SIV has similar amino acid with AIV. CIV is relatively different from others.

Table 9. Rule extraction under 9 window of TAT.

class	rule	frequency
class1	pos2=L pos6=W pos4=D pos9=R pos5=E pos8=R	0.80
class2	-	0.75
class3	pos2=I pos4=T	0.75

In Table 9, AIV is similar with SIV and CIV. SIV is similar with AIV. CIV is Smilar with AIV.

Table 10. Rule extraction under 5 window of VIF.

class	rule	frequency
class1	pos2=L pos4=R	0.80
class2	pos2=T	0.75
class3	pos4=W	0.75

In Table 10, AIV, SIV, and CIV share considerably similar amino acid sequence.

Table 11. Rule extraction under 7 window of VIF.

Class	rule	frequency
class1	pos2=L pos6=W pos4=D pos9=R pos5=E pos8=R	0.80
class2	-	0.75
class3	pos2=I pos4=T	0.75

In Table 11, AIV is different from others. However, SIV, and CIV are also different but not clearly.

Table 12. Rule extraction under 9 window of VIF.

class	rule	frequency
class1	pos1=F	0.75
class2	pos1=E pos9=G	0.75
class3	pos9=N	0.80

In Table 12, AIV is similar with SIV. SIV and CIV are similar each other as well as similar with AIV.

4.2 SVM

The experiments were held for each protein of 5window, 7window and 9window. In each result, we calculated the mean value of the data and drew the bar graph to enhance the visual recognition with the relative value. [16]

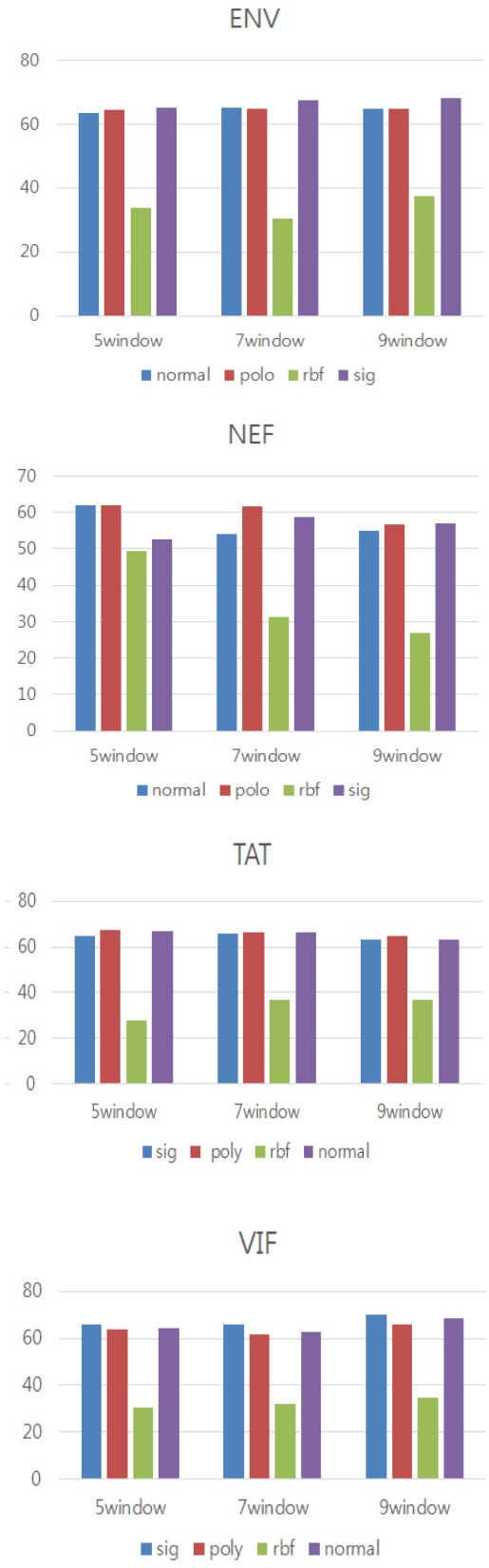

Figure 1. Mean value under 5 window, 7 window, 9 window of SVM

In Fig. 1, the average value of the rbf is evidently lower than that of normal, polynoml and signoid. The results means that the relation between the three viruses. SIV, CIV and AIV are in nonlinear relationship, so the three are seemed to be quite mixed.

5 Conclusion

The paper applied the approach of comparing the properties of CIV, SIV and AIV. In order to find out the genetic information of AIV, the 4 kinds of protein such as nef, tat, vif and env protein were the subjects of the experiments. Decision tree and SVM were used in the experiments. According to the rule which were extracted form the decision tree, the three viruses are evenly mixed in considerable amount of DNA sequences. From SVM, we realized that the relation between SIV, AIV and CIV are below the unlinear relationship. This similarities and differences discovered through our experiment may support other studies of the treatment of not only CIV but also SIV and AIV. We hope this study to be extended to other fields also in logical way and content aspects.

References

1. Colin R. Parrish, Yoshihiro Kawaoka (2005)."The Origins of New Pandemic Viruses: The Acquisition of New Host Ranges by Canine Parvovirus and Influenza A Viruses", Annual Review of Microbiology.
2. Characterization of a Novel Influenza A Virus Hemagglutinin Subtype (H16) Obtained from Black-Headed Gulls.
3. Clinical Signs and Symptoms Predicting Influenza Infection; Arnold S. Monto, MD; Stefan Gravenstein, MD; Michael Elliott, MD; Michael Colopy, PhD; Jo Schweinle, MD.
4. Chulseung Lee, Kwonil Jung, Jinsik Oh, Taehoon Oh, Sangyoon Han, Jeongmin Hwang, Minjoo Yeom, Donghyun Son, Jongman Kim, Bongkyun Park, Hyoungjoon Moon, Daesub Song ,Bokyu Kang(2009)."Protective efficacy and immunogenicity of an inactivated avian-origin H3N2 canine influenza vaccine in dogs challenged with the virulent virus", Vet Microbiol.
5. Shuo Su, Ye Chen, Fu-Rong Zhao, ji-Dang Chen, Jie-Xiang Xie, Zhong-Ming Chen, Zhen Huang, Yi-Ming Hu, Min-Ze Zhang, Li-Kai Tan, Gui-Hong Zhang, Shou-Jun Li(2013). "Avian-origin H3N2 canine influenza virus circulating in farmed dogs in Guangdong, China", Infection, Genetics and Evolution' Volume 19, Pages 251–256, October 13
6. Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus
7. Wentworth DE1, McGregor MW, Macklin MD, Neumann V, Hinshaw VS. "Transmission of Swine Influenza Virus to Humans after Exposure to Experimentally Infected Pigs" ; Departments ofPathobiological Sciences and Veterinary Sciences,UniversityofWisconsin-Madison, and Agracetus, Inc., Middleton, Wisconsin
8. Webby RJ1, Swenson SL, Krauss SL, Gerrish PJ, Goyal SM, Webster RG. "Evolution of Swine H3N2 Influenza Viruses in the United States"
9. White JM, Hoffman LR, Arevalo JH, et al. (1997). "Attachment and entry of influenza virus into host cells. Pivotal roles of hemagglutinin". In Chiu W, Burnett RM, Garcea RL. Structural Biology of Viruses. Oxford University Press. pp. 80–104.
10. Heo, C. and Yoon, T. (2014) Deeper Understanding about Attributes of HIV Employing Support Vector Machine. International Journal of Bioscience, Biochemistry and Bioinformatics, 4, 336-339.
11. Lim, S.J., Heo, C., Hwang, Y. and Yoon, T. (2015) Analyzing Patterns of Various Avian In-fluenza Virus by Decision Tree. International Journal of Computer Theory and Engineering,7.
12. Jang, S. P., Park, K. H., Kim, Y. L., Cho, H. N., & Yoon, T. S. (2015). Comparison of H5N1, H5N8, and H3N2 Using Decision Tree and Apriori Algorithm. Journal of Biosciences and Medicines, 3(06), 49.
13. Go, E., Lee, S. and Yoon, T. (2014) Analysis of Ebolavirus with Decision Tree and Apriori Algorithm. International Journal of Machine Learning and Computing, 4.
14. Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics 21.3 (1991): 660-674.
15. Friedl, Mark A., and Carla E. Brodley. "Decision tree classification of land cover from remotely sensed data." Remote sensing of environment 61.3 (1997): 399-409
16. Furey, Terrence S., et al. "Support vector machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics 16.10 (2000): 906-914.