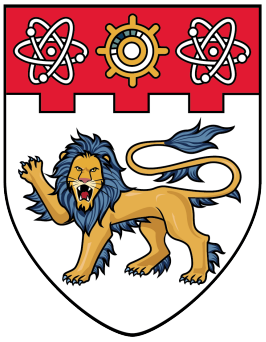


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

META-ANALYSIS ON THE LETHALITY OF INFLUENZA A VIRUSES USING MACHINE LEARNING APPROACHES

YIN RUI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

January 2020

META-ANALYSIS ON THE LETHALITY OF INFLUENZA A VIRUSES USING MACHINE LEARNING APPROACHES

YIN RUI

School of Computer Science and Engineering

A thesis submitted to Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

January 2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

2020-01-08 .....

Date

YIN Rui .....

YIN Rui

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

8 - 1 - 2020

Date

Associate Prof Kwoh Chee Keong

Authorship Attribution Statement

This thesis contains material from 3 papers published in the following peer-reviewed conference and journal where I am the first author.

Part of Chapter 2 is published as Yin R, Tran V H, Zhou X, et al. Predicting antigenic variants of H1N1 influenza virus based on epidemics and pandemics using a stacking model[J]. PloS one, 2018, 13(12): e0207777.

The contributions of the co-authors are as follows:

- I conceived, performed experiments, interpreted results, and completed the manuscript.
- Mr. Tran helped to perform experiments and analyze interpreted results.
- Ms. Zhou revised the manuscript and provided suggestions.
- A/Prof. Zheng involved discussion and overall supervision.
- Associate Prof. Kwoh provided direction and leadership to the research.

Chapter 3 is published as Yin R, Zhou X, Ivan F X, et al. Identification of Potential Critical Virulent Sites Based on Hemagglutinin of Influenza a Virus in Past Pandemic Strains[C] Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science. ACM, 2017: 30-36.

The contributions of the co-authors are as follows:

- I conceived, performed experiments, interpreted results, and completed the manuscript.
- Ms. Zhou helped to perform experiments and analyze interpreted results.

- Dr. Ivan revised the manuscript with discussions.
- A/Prof. Zheng provided the initial project direction and help to revise the manuscript.
- Prof. Vincent involved discussion and overall supervision.
- Associate Prof. Kwoh provided overall supervision, direction and leadership to the research.

Part of Chapter 2 and 4 is published as Yin R, Zhou X, Zheng J, et al. Computational identification of physicochemical signatures for host tropism of influenza A virus[J]. Journal of bioinformatics and computational biology, 2018: 1840023-1840023.

The contributions of the co-authors are as follows:

- I conceived, performed experiments, interpreted results, and completed the manuscript.
- Ms. Zhou helped to perform experiments and analyze interpreted results.
- A/Prof. Zheng revised the manuscript and provided suggestions.
- Associate Prof. Kwoh provided direction and suggestions.

2020-01-08 ..... YIN Rui .....

Date YIN Rui

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Kwoh Chee Keong for the patient and constant support of my Ph.D. study and related research, for his time, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this report. I could not have imagined having a better advisor and mentor for my Ph.D. study. Besides my advisor, I would also like to thank the rest of my instructor: Prof. Zheng Jie and Dr. Fransiskus Xaverius Ivan, for their insightful comments and encouragement, but also for the hard question which inspires me to widen my research from various perspectives.

Furthermore, my sincere thanks also go to Dr. Shamima Banu Binte, Mr. Lin Zhuoyi, Mr. Mohamed Ragab Mohamed Adam and Miss Zhang Yu, who have provided lots of valuable methods and suggestions. Without their precious support, it would not be possible to conduct this research. I also give my thanks to the team of Biomedical Informatics Lab. It is a quite homelike lab and responsible staff that provides me such a comfortable place and facility to conduct my research for the previous 4 years of Ph.D. study. I will not go smoothly without such a warm environment and friendly people. I want to especially thank our lab manager who is always readily available to give my technical support when there is a problem with my desktop computer with professional expertise.

Last but not least, I would like to thank my parents and Ms. Zhuang Pei, who unselfishly support me throughout the Ph.D. study and my life in general.

Table of Contents

Abstract

Influenza viruses are persistently threatening public health, causing annual epidemics, and sporadic pandemics. The majority of influenza viruses reside among the avian species due to host range restriction. However, some avian strains do acquire the capability to overcome host species barrier to cause human infections due to mutations and reassortments. These novel influenza strains may cause high mortality and morbidity. The main target of this thesis is to analyze the lethality of influenza A virus by profiling its virulence and antigenicity using machine learning approaches.

Firstly, potential critical virulent sites are investigated based on the hemagglutinin of the influenza A virus using past pandemic strains. Three rule-based algorithms are utilized to classify the pandemic and non-pandemic strains and extract the rules. These rules consist of mutations that occurred on the potential critical virulent sites. Fourteen out of the sixteen sites detected by our experiments are validated as receptor binding sites or antigenic sites.

Secondly, a host tropism framework to predict avian, human and swine strains. Seven physicochemical properties are used to generate features through Amino Acid Composition (AAC) and global descriptor CTD (Composition, Transition, Distribution). A novel computational method named HopPER is then developed based on the host tropism prediction system through random forest. In addition to the accurate prediction of reassortment for complete genomes, HopPER also demonstrates its effectiveness on incomplete genomes. The analysis of the evolutionary patterns of avian, human and swine strains by HopPER has further

revealed the reassortment history of the influenza viruses.

Thirdly, an integrative model is created to predict influenza virulence, incorporating prior mutation and reassortment information of influenza viruses. Using the mouse lethal dose 50, the virulence of the infections is classified as avirulent and virulent. The prior information on mutation and reassortment of input genomes are obtained by the previous computational models. By integrating this prior knowledge into all the predictive models using the posterior regularization technique, the proposed framework can improve the performance of virulence prediction. The experimental results validate the effectiveness of our proposed framework for virulence prediction. Moreover, the importance weights of the prior viral information will assist biologists to gain a better understanding of how the mutations influence the degree of virulence.

Lastly, it is shown that antigenicity is another crucial factor reflecting viral lethality. A novel algorithm is proposed to predict influenza antigenic variants of influenza A viruses via a 2D convolutional neural network. Specifically, the introduction of a new distributed representation makes it possible to deal with sequence and antigenic data of influenza strains. The squeeze-and-excitation mechanisms are integrated into the convolutional neural networks (CNNs), which enables networks to focus on informative residue features. Experimental results on three different influenza datasets have demonstrated superior performance over the existing state-of-the-art computational models.

In summary, this thesis elaborates novel methodologies for the analysis of the lethality of influenza A virus through predicting its virulence and antigenicity using machine learning approaches. This offers an improvement on existing influenza virologic surveillance and provides an early warning of an impending outbreak.

List of Tables

List of Figures

List of Abbreviations

AAC	Amino acid composition
AdaGrad	Adaptive gradient algorithm
Adam	Adaptive moment estimation
CNN	Convolutional neural network
BLAST	Basic local alignment search tool
CTD	Composition-Transition-Distribution
DNN	Deep neural network
GiRaF	Graph-incompatibility-based Reassortment Finder
GISAID	Global Initiative on Sharing All Influenza Data
GISRS	Global Influenza Surveillance and Response System
HopPER	Host-prediction-based Probability Estimation of Reassortment
HA	Hemagglutinin
HI	Hemagglutination inhibition
HPAI	Highly pathogenic avian influenza
IVA-CNN	Convolutional neural network model to infer influenza antigenic variants
KNN	K-nearest neighbor
LR	logistic regression
M1	Matrix 1 protein
M2	Matrix 2 protein
MAFFT	Multiple Alignment using Fast Fourier Transform
MCC	Matthew's correlation coefficient
MLD	Mouse lethal dose
MN	Micro-neutralization
NA	Neuraminidase

NCBI	National Center for Biotechnology Information
NLP	Natural language processing
NN	Neural network
NP	Nucleoprotein
NS1	Non-structural protein 1
NS2	Non-structural protein 2
OneR	One Rule
PA	Polymerase acidic protein
PART	Projective Adaptive Resonance Theory
PA-X	PA protein translated from alternate open reading frame
PB1	Polymerase basic protein 1
PB1-F2	Accessory protein F2 translated from PB1 segment
PB1-N40	Accessory protein N40 translated from PB1 segment
PB2	Polymerase basic protein 2
RF	Random forest
RIPPER	Repeated incremental pruning to produce error reduction
RMSProp	Root mean square propagation
RNP	Ribonucleoprotein
ROC	Receiver operating characteristic
RSR	Reassortant strain rate
SE	Squeeze-and-Excitation
SGD	Stochastic gradient descent
S-OIV	Swine-origin influenza viruses
SVM	Support vector machine
TPV	True positive value
VGG	Visual Geometry Group
WHO	World Health Organization

Chapter 1

Introduction

This chapter starts with the background of influenza. Following the objectives to be explored, in a receding sequence, the theme of the thesis is to establish computational models to analyze the lethality of influenza viruses through virulence and antigenicity prediction with machine learning approaches. Research contributions and the organization of this thesis are presented as well.

1.1 Background

Influenza is an infectious disease that can lead to headaches, body aches, fever, pneumonia and even death when humans get infected [1]. There are four types of influenza, that is A, B, C and D. Influenza A and B are the main types that infect humans and cause epidemics every year. Influenza C generally causes a mild respiratory illness and is not considered to cause epidemics [2], while Influenza D viruses primarily affect cattle and have not infected or caused illness in people yet [3]. Among these four influenza types, influenza A is the most widespread and influential that is my main research target. It will become an epidemic when influenza viruses rapidly infect many people within a short period in a given population, which occurs every year around the world [4]. Every year, there are about 250,00 to 500,000 deaths around the world according to the World Health Organization (WHO) [5]. However, an epidemic will turn to a pandemic if it spreads to other countries or continents and affects numerous people. In

history, there were mainly five influenza pandemics since the nineteenth century that include the H1N1 Spanish pandemic in 1918, the H2N2 Asian flu in 1957, the H3N2 Hong Kong flu in 1968, the H1N1 Russian flu in 1977 and the most recent swine-origin flu in 2009 [6] [7]. Millions of people are infected with enormous economic loss, intensively threatening public safety and health. Most recently, the outbreak of the new type of bird flu H7N9 in China has raised huge public concern due to the unusually high mortality rate [8], which has the potential of triggering another pandemic in the future.

Though restrained to the specific host species, influenza strains indicate high diversity. The viruses possess mutability and high frequency of genetic reassortment and have a great ability to be lethal and lead to disease. Virulence is the degree of damage caused to a host by parasite infection [9]. It denotes the capacity of an influenza virus to produce disease on its host. Eradicating the influenza viruses is an unrealistic goal due to the high evolutionary rates. Instead, formulating flu surveillance systems and vaccines are the main and achievable methods of preventing influenza-related diseases. Current systems of the flu vaccine by WHO rely on empirical determination of antigenicity by traditional biological experiments such as hemagglutination inhibition (HI) and micro-neutralization assays, which are time-consuming and labor-intensive.

Mutations on hemagglutinin (HA) and neuraminidase (NA) segments that cause antigenic drift will make the protein unrecognizable to pre-existing host immunity. This is a complex process and hence making a judgment based on sequence only will fail to estimate its perniciousness. Nevertheless, a single study may not provide sufficient confidence about influenza lethality, hence combining different aspects for a meta-analysis to estimate the lethality will provide novel insight.

1.2 Objectives

The main goal of this research is to analyze the lethality of influenza A viruses by inferring their virulence and antigenicity through computational models. To start with, the virulence of the viruses directly reflects the degree of lethality. However, single genomic marker or element would not be adequate to comprehensively determine the virulence level of influenza strains. Several factors are investigated that are associated with virulence, including viral mutation and reassortment. An integrative model is then constructed to detect the virulence label using prior mutation and reassortment information, which is the first goal of the thesis. Apart from the virulence, antigenicity is also a crucial metric that can evaluate the lethality of influenza strains. Hence, building computational models to predict influenza antigenicity is the second goal. Combining the predictive results of influenza virulence and antigenicity, we can further infer the lethality of the viruses and promote flu surveillance. The objectives of this thesis are presented below:

- Explore and identify the biological process and factors that are associated with the influenza virulence.
- Forecast the virulence of influenza A viruses through an integrative framework with prior viral knowledge
- Develop methodology for reliable prediction of influenza antigenicity

The main contributions of this thesis are:

- Build novel computational methods to identify the potential virulent mutation sites and detect the probability of influenza reassortment

- Develop an integrative framework to predict influenza virulence incorporating mutation and reassortment information with extensive experiments
- Propose a new 2D CNN-based model for influenza antigenicity prediction
- Provide strategies for the inference of viral lethality through meta-analysis of virulence and antigenicity

1.3 Thesis organization

The meta-analysis on the lethality of influenza viruses is characterized by using an integrative framework and a series of computational tools. The results from this work can help to infer the lethality of influenza strains and further facilitate the surveillance system and protect public safety from possible future pandemic and epidemic. Figure 1.1 shows a systematic overview of the research structure.

This thesis contains seven chapters that the research background and the main objectives are in Chapter 1. Chapter 2 comprises an overall up-to-date literature

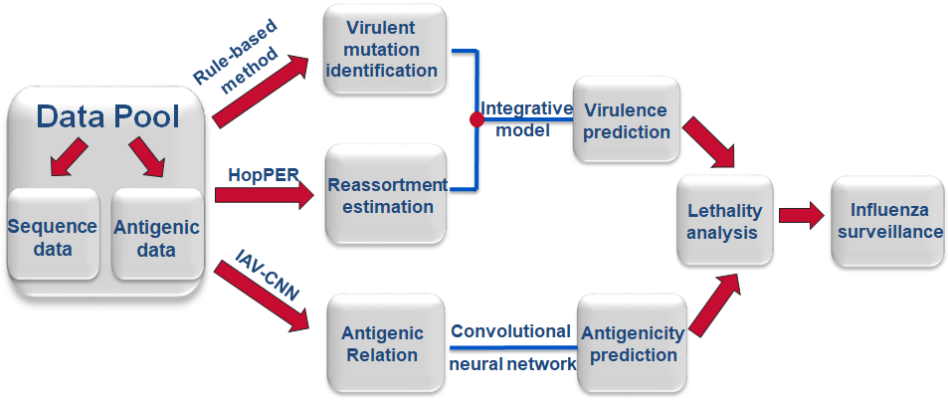


Figure 1.1: The overview of research workflow for the lethality analysis of influenza A viruses.

review on influenza A virus, viral factors on virulence i.e., mutations (antigenic drift) and reassortments (antigenic shift), current state-of-the-art computational models applied to the prediction of influenza virulence, as well as existing bioinformatic methods to detect its antigenicity.

Chapter 3 identifies virulence-associated sites on the hemagglutinin of influenza A virus using pandemic strains. Here a binary classification problem is defined that categorizes influenza strains into pandemic and non-pandemic classes based on their protein sequences. Three rule-based algorithms are applied, namely OneR (One Rule), JRip/RIPPER (Repeated Incremental Pruning to Produce Error Reduction) and PART (Projective Adaptive Resonance Theory), to extract rules, which comprise potential critical virulent sites for the pandemic strains.

Chapter 4 first constructs computational models to predict host tropism of influenza on eight different proteins of influenza viruses that use Amino Acid Composition (AAC) and global descriptor CTD (Composition, Transition, Distribution) through seven physicochemical attributes. Based on the host prediction model, a novel computational method named HopPER (Host-prediction-based Probability Estimation of Reassortment) is developed, that sturdily estimates reassortment probabilities with random forest. Additionally, not only can HopPER be applied to complete genomes but its effectiveness on incomplete genomes is also demonstrated. Further analysis of the evolutionary success of avian, human and swine viruses generated through reassortment across different years using HopPER has revealed the reassortment history of the influenza viruses.

Chapter 5 describes an integrative model to predict the virulence of influenza viruses using prior viral information, i.e., mutation (Chapter 2 and 3) and reassortment (Chapter 4) information, which enables us to incorporate heteroge-

neous discrete biological knowledge for the predictive models. Specifically, the proposed model leverages deep learning-based approaches as the basic predictive model with posterior regularization for the constraint posterior feature set, which automatically learns the bounds of constraint features. The experiments are implemented on the collected influenza dataset and the results show that the proposed model outperforms existing traditional machine learning classifiers and deep learning-based methods. Moreover, it can display the importance of different prior knowledge that contributed to the virulence prediction model. The experiments on individual influenza subtypes further demonstrate the utility and robustness of the proposed model.

Chapter 6 proposes a 2D convolutional neural network (CNN) model to infer influenza antigenic variants (IVA-CNN). Hemagglutinin inhibition (HI) assay is one of the main methods for the determination of influenza antigenicity and vaccine selection. However, it is costly and time-consuming that will not meet the increasingly available sequences. ProtVec, a new distributed representation of amino acids, is proposed in a variety of downstream proteomic machine learning tasks. After splittings and embeddings of influenza strains, a 2D squeeze-and-excitation CNN architecture is constructed that enables networks to focus on informative residue features by fusing channel-wise and spatial information with local receptive fields at each layer. The proposal of IVA-CNN in this chapter incorporates techniques to address the barrier of HI assay and presents a superior performance on the prediction of antigenic variants on three different influenza datasets.

Chapter 7 summarizes the previous work and proposes future directions in completing the construction of computational models on the inference of influenza lethality and flu surveillance.

Chapter 2

Literature review

[1]This chapter presents a literature review on the analysis of influenza lethality through virulence and antigenicity by computational models. It contains the introduction of influenza A virus, viral factors affecting the virulence (antigenic shift and drift), genetic markers associated with virulence, existing bioinformatics methods and machine learning on mutation and reassortment identification, as well as influenza antigenicity prediction.

2.1 Influenza A virus

The influenza A virus is a member of the Orthomyxoviridae family that consists of eight negative-sense, single-stranded RNA segments, encoding up to 16 classic proteins. The structure of the influenza A virus is shown in Figure 2.1 [10]. Among all the segments, HA and NA segments are the most important that characterize influenza A viruses [11]. The HA segment is responsible for binding the virus to cells with sialic acid on the membranes [12]. It is initially synthesized as a single polypeptide precursor (HA0), which needs to be cleaved into subunits

[1]Part of the work in this chapter has been published in [Yin R, Tran V H, Zhou X, et al. Predicting antigenic variants of H1N1 influenza virus based on epidemics and pandemics using a stacking model[J]. PloS one, 2018, 13(12): e0207777] and [Yin R, Zhou X, Zheng J, et al. Computational identification of physicochemical signatures for host tropism of influenza A virus[J]. Journal of bioinformatics and computational biology, 2018: 1840023-1840023]

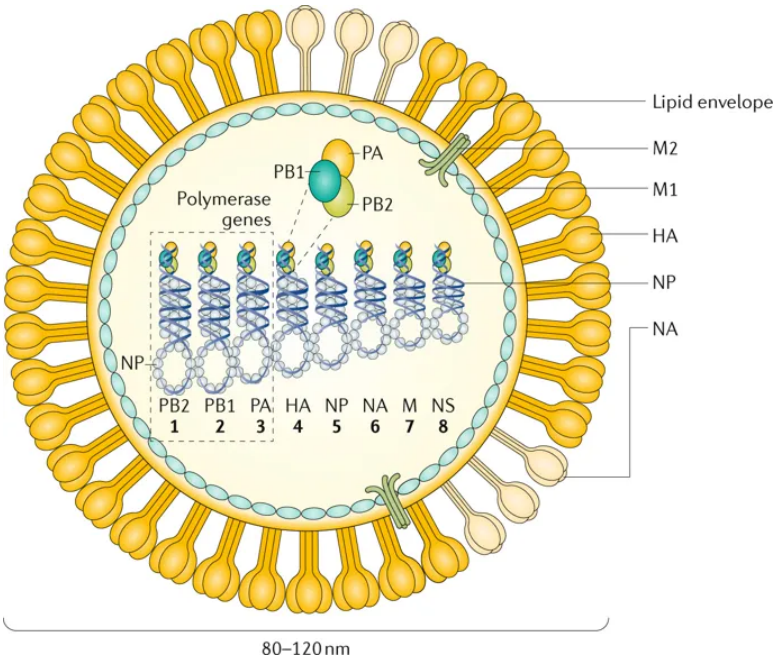


Figure 2.1: Structure of influenza A virus. Reprint with permission from reference Copyright 2018, Springer Nature [10].

HA1 and HA2 by cellular proteases to become biologically active [13]. NA segment functions as a tetramer that cleaves sialic acid from cells and virion glycoproteins to prevent clumping of released viruses [14]. The influenza A viruses are categorized into subtypes based on the antigenic and genetic properties of HA and NA segments. Up to now, 18 hemagglutinin subtypes (H1 to H18) and 11 neuraminidase subtypes (N1 to N11) have been identified [15].

Apart from the HA and NA segments, the influenza A virus particle comprises nucleoprotein (NP), non-structural proteins NS1 and NS2, matrix proteins M1 and M2, and three RNA polymerase subunits, namely, polymerase acidic protein (PA), polymerase basic protein 1 (PB1) and polymerase basic protein (PB2) [16]. The RNA polymerase complex plays crucial roles in both transcription and replication of the viral genomes [17]. The PA segment contains a second open reading frame that encodes the PA-X protein [15]. Similarly, the PB1 segment

can encode two more proteins, which are PB1-F2 and PB1 N40. The M1 protein mediates virion assembly, while M2 protein forms the channels for the viral entry. Moreover, NS1 protein functions as an antagonist that inhibits interferon related activities [18] and NS2 protein has been implicated in mediating the nuclear export of ribonucleoprotein (RNP) complexes and recruiting ATPase for efficient viral exit [19].

The influenza viruses are known to infect a broad range of hosts across species that include avian, human, swine and other mammals. There has been a long history of human infections by influenza viruses with high mortality and morbidity rates. Migratory waterfowls are the natural reservoirs of influenza viruses of all known subtypes. Usually, most of these strains are confined within their host species and do not cross the host-species boundary [20]. However, on rare occasions, adaptations had taken place over multiple loci in the virus genome, enabling the virus to the cross-species barrier and infect humans or other animals [21]. The cases of human infection were confirmed by the outbreaks of epidemics or pandemics. Table 2.1 shows several outbreaks of flu pandemic in history. The 1918 flu pandemic caused the systematic spread of the H1N1 viruses and killed as many as 50 million people worldwide, and then another pandemic arose that introduced avian viruses H2 HA and N2 NA genes into humans [22]. In 1968, the reassortant of human and avian strains that possessed an H3 HA gene of avian-origin viruses led to the Hong Kong flu pandemic. The attack rates were the highest (40%) among children of 10 to 14 years old [23]. Besides, the emergence of the highly pathogenic avian influenza (HPAI) H5N1, despite weak human-to-human transmissibility, caused severe acute respiratory diseases to those infected, having a 60% mortality rate [24]. Besides, several outbreaks of HPAI H7 viruses in poultry have resulted in human infections. In 2003, a large outbreak of an HPAI H7N7 virus in poultry in the Netherlands caused 89 cases

Table 2.1: The summary of several flu outbreaks in the history.

Year	Pandemic Name	Subtype	Death	Origin of virus genes
1918	Spanish Flu	H1N1	17 ∼50 million	Unclear origin, probably Kansas with m ammalian and avian genetic information
1957	Asian Flu	H2N2	2 million	Originated from an avian influenza A virus, including the H2 hemagglutinin and the N2 neuraminidase genes.
1968	Hong Kong Flu	H3N2	1 million	Two genes from an avian influenza A virus, including a new H3 hemagglutinin, but also contained the N2 neuraminidase from the 1957 H2N2 virus.
2009	Swine-origin Flu	H1N1	151,700 to 575,400	Resulted from reassortment, a process through which two or more influenza viruses can swap genetic information by infecting a single human or animal host.

of human infections [25]. Another novel influenza A virus H7N9, derived from multiple reassortments between avian strains, caused human infections with a fatality rate of approximately 30% and quickly spread to more than 18 provinces and municipalities in China [26]. Most recently, direct-contact and airborne transmission of influenza subtype H9N2 viruses were double or even triple re-assortants that have amino acid signatures in their hemagglutinin, indicating their potential to infect humans, which could lead to the next influenza pandemic among humans [27]. Researchers are employing word embedding approach to classify unstructured text data, e.g., Twitter, Facebook, and identify actionable information during early stages to stem the proliferation of an outbreak [28].

2.2 Mutation and reassortment of influenza A viruses

Mutations are the key mechanisms driving the evolution of influenza viruses. They have enabled changes in viral proteins with the emergence of novel strains, affecting the virulence when these new viruses adapt to different hosts. The occurrence of novel antigenic variants often results in vaccine failure due to the antigenic drift. [29]. The frequent mutations in RNA viruses enable them to form highly diverse strains, surviving and reproducing in the environment [30]. In particular, HA has the highest mutation rate ($\sim$5.7 $\times$ 10^{-3} substitutions per site per year) among all segments [31].

Tracing the evolution of influenza viruses could provide a comprehensive understanding of the dynamics of viral evolution. The mutations are the major driving force that contributes to the genetic and antigenic evolution of influenza viruses. Smith *et al.* developed the antigenic cartography to quantify and visualize the antigenic differences and site mutations of H3N2 using hemagglutination inhibition (HI) assay data [32]. Fleury *et al.* presented that some key mutations revealed by the binding affinity studies of H3N2 could escape antibody neutralization [33]. The study of Shih *et al.* suggested not only the positive selection has occurred in most of the time but also uncovered that multiple mutations at the antigenic sites cumulatively enhance the antigenic drift [34]. Du *et al.* outlined a network model to illustrate H3N2 virus evolutionary patterns and dynamics [35]. The analyses strengthened the findings that simultaneous multi-site mutations underpin the dynamics of human H3N2 evolution in known antigenic regions of the viral HA. Moreover, the mutations in the HA protein altered receptor-binding preference which allowed for the transmission of highly pathogenic avian H5N1 between mammals [36]. de Vries *et al.* found that three amino acid mutations on HA of H7N9 could change specificity to human-type receptors [37]. Overall,

most of the mutations in influenza viruses were deleterious or lethal that would impact the influenza evolution at the global scale [38] [39] [40].

Apart from the mutation, another important source of the diversity of RNA viruses is the error-prone nature of RNA synthesis. Reassortment is the key biological mechanism that drives the evolution of influenza viruses and significantly increases the diversity of circulating strains. The segmented structure of the viral genome allows for the exchange of the segments from different influenza strains to co-infect the host cell [41]. From a co-infection with two viral strains, each carrying eight segments, $2^8 = 256$ different progeny genotypes can be produced through reassortment [42]. Thus, the gene swapping through reassortment plays an important role in the evolution of influenza viruses and viral diversity.

Recent studies of viral evolutionary dynamics have displayed the surprisingly high frequency of mixed influenza virus infections and revealed the importance of reassortment in generating epidemic and pandemic [42]. The evidence suggested that the first pandemic in 1918 caused by influenza A H1N1 virus in Spanish is mixed recombination between the North American H1N2 swine influenza virus, European H1N1 swine influenza virus, North American avian influenza virus and H3N2 influenza virus [43]. Besides, three other influenza pandemics were also associated with emerging strains due to the reassortment process. For example, it was suggested that the reassortant strains from the combination of avian and human viruses are responsible for the two pandemics in 1957 and 1968, respectively [44] [45] [46]. Besides, it was believed that the 1957 Asian flu originated from the 1918 Spanish flu. The segments of HA and NA were derived from other avian viruses, while five of the rest of the segments were retained. Similarly, with six conserved gene segments (PA, PB2, NA, NP, NS and M), the 1968 Hong Kong flu was the descendant of previous strains from 1957 Asian flu. The reassortment of human and avian strains with an H3 HA gene

originated from the avian virus led to the 1968 H3N2 pandemic [47]. The most recent pandemics in 2009 was caused by a reassortant strain generated through the reassortment of human, avian and swine strains [7]. Furthermore, except for the pandemic strains, it was reported that the new emerging H7N9 strains were also uncovered as progeny through the reassortment between H7N9 and H9N2 subtypes [48] [49]. Therefore, the reassortment mechanism has made a significant impact on the occurrence of novel strains with high virulence and so as the outbreak of pandemics which infects a huge amount of people with illness and deaths. It is a crucial factor that characterizes the virulence of influenza strains.

2.3 Genetic markers associated with virulence of influenza A viruses

The measurements of virulence of the influenza strains remain a complicated and challenging issue. In regards to the genomic sequences, genetic signatures are usually utilized to investigate the influence on virulence through aspects such as interspecies transmission and receptor binding specificity [50]. The genetic markers are usually obtained from biological experiments or sequence analysis of amino acid residues with computational techniques [51]. Table 2.2 summarizes the molecular mutations of influenza A viruses that will influence the virulence.

The mutations in viral genes that are associated with virulence have been explored in many ways. These include the generation of mouse-adapted influenza avian viruses through lung-to-lung passage using plasmid-based reverse genetic techniques combined with mutagenesis methods. The analyses of the transmissible ability of swine-origin 2009 A/H1N1 by establishing ferrets and mice model [67]. All these techniques have provided various insights into viral mutations that

Table 2.2: Literature summary of the molecular mutation markers of influenza viruses that are associated with virulence or virulent factors.

Protein	Position	Mutation	Determinant	References
HA[1]	138	Ser (S) - Ala (A)	Sialic acid linkage	[52]
	163	Lys (K)- Glu (E)	Associated with pandemics strains	[53]
	185	Asp (D) - Ser (S)	Associated with pandemics strains	[54]
	187	Gly (G) - Asp (D)	Associated with pandemics strains	[54]
	190	Glu (E) - Asp (D)	Specificity mammalian transmissibility	[55]
	222	Asp (D) - Gly (G)	Increased pathogenicity	[53]
	225	Gly (G) - Asp (D)	Receptor-binding specificity	[56]
	226	Gln (Q) - Leu (L)	Receptor-binding specificity	[56]
	228	Gly (G) - Ser (S)	Receptor-binding specificity	[57]
NA[1]	223	Thr (T) - Ile (I)	Virulence	[58]
	275	His (H) - Tyr (Y)		[59]
PB2	591	Gln (Q) - Lys (K)	Virulence	[60]
	627	Glu (E) - Lys (K)	Polymerase activity	[61]
	701	Asp (D) - Asn (N)	Mammalian transmissibility	[62]
PA	35	Phe (F) - Leu (L)	Associated with pandemics	[53]
	97	Thr (T) - Ile (I)	Virulence	[63]
	224	Ser (S) - Pro (P)	Increased polymerase	[64]
	383	Asn (N) - Asp (D)	activity	[64]
NS1	92	Asp (D) - Glu (E)	Virulence	[65]
PB1-F2	66	Asn (N) - Ser (S)	Virulence	[66]

influence the virulence of viral infections. For example, previous studies have demonstrated the specific mutation on viral proteins may lead to increased virulence or enhanced transmission. Increased pathogenicity is revealed in macaques when the avian-type receptor binding ability was conferred by HA-222D and HA-222G [68]. The evidence suggested that the substitutions Q222L and G224S have contributed to the outbreak of the 1957 and 1968 pandemics, which changed the receptor binding of H2 and H3 avian influenza binding specificity to alpha (2,6) linked sialic acid [69]. Mutations in PB2 have also been considered as the most notable marker in host range restriction. The substitution E627K in avian viruses allowed for efficient replication in mammalian cells [61]. Also, the mutation D701N proved the increased transmission and replication of the virus [62]. Interestingly, another single mutation N66S in the accessory protein PB1-F2 could also contribute to the increase of virulence [66].

Moreover, the genetic markers from other segments can also influence virulence and pathogenicity. The mutations at position 223 and 275 in the NA protein [58] [59], 97 in PA [63] and 92 in NS1 [65] were related to enhanced virulence in mammalian hosts. Moreover, it was shown that mutation in multiple sites of a specific influenza protein and multiple genes would make a synergistic effect on the virulence of influenza viruses. For example, the synergistic effect of dual mutations N383D and S224P in PA led to the increased polymerase activity and has been used as the hallmark for natural adaption of H1N1 and H5N1 viruses to mammals [64]. Another example was the synergistic action of two mutations D222G and K163E in HA protein with the mutation F35L in PA of 2009 H1N1 pandemic strain that causes lethality in the infected mouse [53] [62].

2.4 Computational methods to identify mutation and reassortment of influenza A viruses

With the increasing progress of computing ability and the development of machine learning, computational techniques enable researchers to discover knowledge more efficiently. Data mining approaches have been demonstrated to be more powerful than manpower in some fields including biology. Large-scale available sequences make it possible to retrieve clues from past strains and to estimate the mutations of influenza viruses. The prediction of genetic mutation and its effect on influenza viruses and has attracted considerable attention in recent years. Rapid progress has been achieved in predicting the mutations and analyzing mutation effects including single nucleotide variant prediction [70] [71] [72], the prediction of potential protein secondary structure in the generation of post-mutation [73] [74] and the detection of the resistance of the virus to drugs after mutations [75] [76]. Salama *et al.* applied neural networks to predict the possible point mutations that occurred on the alignments of primary RNA sequence-structure [77]. Neher *et al.* proposed a model to predict the properties of viruses that have not been characterized antigenically using phylogenetic trees [78]. By mapping the antigenic changes along the path connecting viruses in phylogenetic trees, it allowed the prediction of antigenicity through HA sequence data and further estimated the makeup of future H3N2 seasonal influenza virus population. Marta Łuksza and Michael Lässig described a fitness model for haemagglutinin to predict the frequency of its descendant strains in the following year [79]. Zhu *et al.* found that some mutations in polymerase genes enhanced the virulence of the 2009 pandemic H1N1 influenza virus in mice using statistical analysis [80]. Peng *et al.* identified genome-wide nucleotide sites associated with mammalian virulence in influenza A viruses by computational analysis [81].

Table 2.3: Summary of studies identify influenza reassortment with computational methods.

Datasets	Methods	Results	References
156 H3N2 human genomes of influenza A viruses collected between 1999 and 2004	The combinational analysis of individual and whole gene segment with phylogeny	Multiple co-circulating clades with different population frequencies	[82]
18 H1N1, H1N2 and H3N2 viruses that circulated in North America from 1997 to 2005	Genetic and phylogenetic analyses with cycle sequencing and amplification	Wholly human and reassortant virus genotypes	[83] [84] [85]
16 Resembled swine-origin influenza viruses (S-OIV) from Thailand	Enumerating maximal bicliques with a defined incompatibility graph	Facilitate identification of reassortment patterns and shed lights on the cause of S-OIV	[86]
6 Triple reassortant H3N2 viruses isolated from pigs and turkeys throughout Canada in 2005	Phylogenetic analysis by the method of maximum parsimony with bootstrap resampling	Suggest a fast and complicated interspecies transmission of reassortants	[87]
39 complete and incomplete genome sets	Statistical analysis by diversity and entropy measures of each segment and its correlations	Not only HA and NA but also the PB1 segment reassorted more frequently in swine viruses	[88]
36 well-supported candidate reassortants with strong confidence	Using neighborhood of each segment and nucleotide distance matrix	Provide evidence to draw a more well-rounded picture of the origin of some previous strains	[89]
93 comprehensive reassortment study including human, avian and S-OIV influenza populations	Graph mining technique by searching large collections of Markov chain Monte Carlo-sampled trees	Account for uncertainties in the inferred phylogenies of reassortant strains	[90] [91]
Synthetic datasets produced by the flu evolution simulator	Reconstructed phylogenetic trees of the individual segments and the full genome	Achieve 0% false positive rate and 10% or less false negative rates on the test dataset	[92]
Simulated dataset and a small collection of real viral data isolated in Hong Kong in 1999	Maximum likelihood and Bayesian approach	Effectively identify reassortment events in small viral datasets	[93]
Human H3N2 viruses isolated in New York State (1995-2006) and New Zealand (2000 - 2005)	A quantitative way to measure genetic shifts	Suggest that the patterns of reassortment in the viral population are not random	[94]

Despite the improvement and availability of sequencing analysis and computational ability, the antigenic evolution and how it connects with reassortment is still under exploration [95]. One of the main reasons is that the unavailability of automated tools that can fast and accurately detect the reassortment. Table 2.3 summarizes some computational approaches that can identify reassortments by reconstructing and analyzing species and segments tree. From the table, it can be observed that bioinformatic methods are widely adopted in detecting reassortment of influenza A viruses. Edward performed a phylogenetic analysis of 156 complete genomes of human H3N2 influenza A viruses collected between 1999 and 2004 and the results revealed that multiple reassortment events occurred among these clades [82]. Kingsford *et al.* conducted a comprehensive computational search of all available sequences of the surface proteins of H1N1 swine influenza isolates and found that a similar strain to swine-origin influenza viruses (S-OIV) appeared in Thailand [86]. de Silva *et al.* presented a new approach to identify reassortants from large data sets of influenza whole genome nucleotide sequences [89]. Svinti *et al.* described the implementation of two approaches, namely maximum likelihood and the second Bayesian approach for robustly identifying reassortment events. The algorithms rested on the idea of the significance of the difference between phylogenetic trees and subtree pruning and regrafting operations, which mimic the effect of reassortment on tree topologies [92]. Moreover, Niranjan described a novel computational method, called GiRaF (Graph-incompatibility-based Reassortment Finder), that robustly identifies reassortments in a fully automated fashion while accounting for uncertainties in the inferred phylogenies [90]. Yurovsky and Moret developed a fully automated flu virus reassortment finder, which is inspired by the visual approach to reassortment identification and thus uses the reconstructed phylogenetic trees of the individual segments and the full genome [93]. These studies reflect the

more advanced methods to investigate the reassortment of influenza viruses.

2.5 Antigenicity prediction of influenza A viruses

In addition to the virulence, antigenicity is another indicator that can be used to evaluate the degree of lethality of influenza strains. Table 2.4 illustrates typical experimental and computational approaches for the detection of influenza antigenicity. Among all the experimental methods, hemagglutinin inhibition (HI) assay is the primary way to determine the antigenicity of influenza viruses and quantitative antibody titers for vaccine selection [96]. However, HI assay is a labor-intensive and time-consuming method, which prompts the development of computational techniques for the prediction of antigenic similarity between antisera and antigens to identify the antigenic variants. Sequence-based methods and imputation-based methods [97] are the most common computational methods for antigenic prediction. Smith et al. constructed an antigenic map to determine the antigenic evolution of influenza A H3N2 virus from 1968 to 2003 [32]. Lorusso et al. used antigenic cartography to analyze the antigenic properties of 2008 H1 viruses and demonstrated that the viruses in the different phylogenetic clusters are also antigenically divergent [98]. The antigenic patterns and evolution of human influenza A (H1N1) viruses were investigated by Liu et al., who inferred the antigenic clusters from a large-scale sequence data covering the whole epidemic history of H1N1 [99]. Bedford et al. and Du et al. constructed the maps of the global circulation patterns of seasonal flu strains and antigenic evolution, respectively [100] [101]. These previous works depicted the evolutionary paths of influenza and provided the foundation for computational models of antigenicity prediction.

As for the sequence-based models, Ren et al. applied random forest regres-

Table 2.4: The summary of methods for influenza antigenicity detection.

	Category		Reference
Experimental methods	Hemagglutination inhibition (HI) assay		[102]
	Micro-neutralization (MN) assay		[103]
	Enzyme-linked immunosorbent assay (ELISA)		[104]
Computational methods	Imputation-based methods	Mapping the antigenic and genetic evolution	[32]
		3D antigenic cartography construction	[97]
	Sequence-based methods	Random forest and support vector regression	[105]
		Universal model based on conserved antigenic structures	[106] [107] [108]
		Graph-guided multi-task sparse learning model	[109]
		Mutation-based antigenicity prediction	[110] [111]
		Bootstrapped ridge regression model	[112]

sion and support vector regression to identify antigenicity-associated sites in the hemagglutinin protein of A/H1N1 seasonal influenza virus [113]. Besides, one of the most crucial factors for the success of influenza vaccination is the timely determination of emerging influenza virus antigenic variants. Sun et al. provided a novel, experimentally validated, computational method for determining influenza virus antigenicity based on HA sequences [114]. Liao et al. proposed a method by incorporating scoring and regression methods to predict antigenic variants of influenza A/H3N2 viruses [115]. Yao et al. proposed a joint random forest method for predicting influenza H3N2 antigenicity from hemagglutinin sequence data [105]. Lee and Chen used the number of amino acid changes located on the five epitope regions for the antigenic variants prediction [116]. Lees et al. provided an update for the frequently referenced five antigenic sites and increase additional assignments to establish five canonical regions [117]. Qiu et al. incorporated the structural context of HA protein to calculate the antigenicity for influenza virus

A/H3N2 with an accuracy of 0.875 [107]. Furthermore, by building a universal model for all HA subtypes of influenza A viruses based on conserved antigenic structures, Peng et al. achieved an accuracy of 0.77 for predicting antigenic variants of avian influenza H9N2 viruses [106]. Furthermore, Richard Neher et al. showed the antigenic differences measured by serological data are well described by antigenic changes along the path connecting viruses in phylogenetic trees [78]. It allows predicting antigenicity from HA sequences by mapping on the trees. Luksza and Laessig developed a fitness model for haemagglutinin that predicts the evolution of the viral population, which maps the adaptive history of influenza A and suggests guidance for vaccine selection [79]. Han et al. developed a Graph-Guided Multi-Task Sparse Learning model that uses multi-sourced serologic data to learn antigenicity-associated mutations and infer antigenic variants [109]. This method enables the rapid characterization of antigenic profiles and the identification of antigenic variants in real-time and on a large scale. All these methods above not only demonstrate the feasibility in predicting the influenza antigenicity but also guarantee the timeliness of identification. Future predictive models would have to consider the antigenicity of influenza strains when evaluating the virulence level of the viruses.

2.6 Influenza surveillance

It has been decades for exploring in influenza field. Although significant progress has been made in mutation and co-mutation identification [118] [119], protein structure prediction [120], reassortment detection [121], and vaccine recommendation [122], etc., it is still not possible to completely prevent the outbreak of influenza epidemics or pandemics. Current methods in flu surveillance remained limited and immature, which mainly relies on the application

of influenza vaccines. To track the genetic composition of circulating influenza strains for the appearance of novel antigenic variants, the WHO has run the Global Influenza Surveillance and Response System (GISRS) [123]. Traditionally, the circulating strains are prepared and sent to WHO by some influenza centers from different countries. These samples that encompass genetic, antigenic and epidemiological data are collected from patients with influenza-like illness. The WHO collaborative centers produce the viruses of selected strains in hens' eggs. Then the 'reassortant laboratories' create reassortants, which are primarily used for vaccine production [124]. The antigenic changes of these reassortants are examined for further evaluation. Selected viruses representative of circulating strains will be characterized based on hemagglutination inhibition and virus micro-neutralization (MN) assays for vaccine virus decision-making purposes. The final vaccine strains for recommendation will be decided in the WHO Influenza Vaccine Consultation Meeting. However, the vaccine strains must be updated periodically to match with circulating viruses for antigenicity determined by HI and/or MN assays. The process of vaccine manufacturing is complicated and time-consuming and that takes 6 months from the time the vaccine strains are recommended [125]. By the time new vaccines are distributed, the dominant viruses in circulation may be different from the vaccine strains selected earlier. Meanwhile, the measurements by HI assays, affected by the substitutions in the host cell receptor binding [111], may also impel influenza evolution and the mechanisms of how to drive viral evolution have been not fully understood [126]. In this regard, timely inference of influenza lethality could vastly facilitate flu surveillance and prevention, which allows for accurate analysis of influenza strains to formulate strategies for disease prevention and control.

2.7 Chapter summary

This chapter describes the literature review of the research objectives in this thesis. There are primarily four types of influenza viruses and influenza A is the most important. Influenza strains are mainly circulating in animals while some of them will manage to cross the host species barrier and infect humans, which could cause the outbreak of epidemics and even pandemics. The large diversity of influenza virus is due to their error-prone RNA polymerase leading to high mutation rates and frequent reassortment of viral RNA segments during co-infection with two or more distinct strains. Mutation and reassortment are also the key mechanisms driving the evolution of influenza viruses and associated with the virulence of influenza strains. The identification of genetic markers is usually used to investigate how they influence the virulence of influenza viruses. With the improvement of computing ability and increasing available biological sequences, many studies have employed the computational model to analyze influenza-related problems that contain genetic marker identification, mutation prediction, reassortment event detect and antigenicity prediction. The flu vaccine is the major preventive countermeasure by eliciting antibody responses against circulating viruses. However, current surveillance methods do not have the capability to estimate the lethality of influenza viruses on infected humans. Further research on identifying virulence-associated factors as well as more advanced computational methods are needed to develop models to better analyze the degree of viral lethality.

Chapter 3

Identification of potential critical virulent sites based on hemagglutinin of influenza A virus in past pandemic strains

[2]The influenza pandemics have caused millions of deaths and enormous economic loss. Current circulating influenza viruses in human, avian, swine and other animals are potential to evolve into novel strains that may cause another pandemic in the future. Hence, recognizing the determinants of pandemic strains helps to raise the alarm of future pandemics. With increasingly large biological data, computational modeling is a good technique for analyzing data, providing novel insight into significant patterns and rules. Here a binary classification problem is defined that the influenza strains are categorized into pandemic and non-pandemic classes based on amino acid sequences. Three rule-based algorithms are applied, namely OneR, JRip and PART, to extract rules, consisting of potential critical virulent sites. The results present good performance in terms of accuracy, specificity, sensitivity and F-measure (more than 0.9 on average for each). Fourteen out of the sixteen potential critical virulent sites detected in the experiments are overlapped with receptor binding sites or antigenic sites.

[2]The work in this chapter has been published in [Yin R, Zhou X, Ivan F X, et al. Identification of Potential Critical Virulent Sites Based on Hemagglutinin of Influenza a Virus in Past Pandemic Strains[C] Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science. ACM, 2017: 30-36]

Besides, some variations occurred in these sites are known to affect the virulence of influenza viruses or to cause more severe symptom in the infected patients.

3.1 Introduction

The influenza A virus is a member of the Orthomyxoviridae family, containing a genome makeup of eight single-stranded, negative-sense RNA segments. It may cause fever, headaches, sore throat, body aches, nose congestion, bronchitis, pneumonia and even death when humans get infected [1]. Influenza infection becomes an epidemic when it rapidly spreads to many people within a short period in a given population, which occurs every year around the world [4]. While an epidemic is usually restricted to the locations, if it spreads to other countries or continents and affects numerous people, then it will turn to a pandemic. As mentioned in Chapter 1, there are mainly five influenza pandemics since the nineteenth century, which caused millions of deaths as well as enormous economic loss, intensively threatening public health and safety. In 2009, WHO made a restatement description of the pandemic to make clear the meaning of the word pandemic and the way to recognize it quoted as ”An influenza pandemic may occur when a new influenza virus appears against which the human population has no immunity [127].” Recently, a new influenza type H7N9 appeared in China, cumulative reported human infections with 359 deaths out of 918 cases by 16 Jan 2017. People are concerned about its potential of triggering another pandemic of this virus and the risk of impacting public health. Therefore, the identification of critical sites from previous pandemic strains that caused human diseases with high virulent influenza viruses is a very important step in promoting the surveillance system and protecting public safety from a possible future pandemic.

The determinants of pandemic influenza strains remain an open and challeng-

ing problem. Animal models, such as mice, ferrets and non-human primates, have been significant helpers in analyzing the host-range, antigenicity and virulence (from mild to severe) of influenza virus. Guinea pigs are also often used for transmission studies. Maines et al. analyzed the transmissible ability of swine-origin 2009 A/H1N1 by establishing ferrets and mice model, shedding light on the atypical symptoms, including gastrointestinal distress and vomiting [67]. Some works demonstrated increased virulence or enhanced transmission with the acquisition of specific mutations. For example, Tokiko et al. showed increased pathogenicity in macaques when conferred the avian-type receptor binding ability by HA-222D and HA-222G [68]. Sander at al. selected mutant A/H5N1 viruses, indicating Q222L and G224S changed the receptor binding of H2 and H3 avian influenza binding specificity to alpha (2,6) linked sialic acid, which contributed to the outbreak of 1957 and 1968 pandemics [69]. Chinh et al. found that mutation R289K-induced conformation in H7N9 suggests potential adaption of the virus itself for future drug-resistance [128]. Although experiments have detected several determinants of pandemic influenza strains, these animal models are expensive, difficult to work with and it is hard to analyze all factors thoroughly.

On the contrary, computational techniques will enable researchers to discover knowledge more efficiently. Data mining approaches have been demonstrated to be more powerful than manpower in some fields including biology. Large-scale sequences available makes it possible to retrieve clues from past strains and to estimate the mutations of influenza viruses. Wu et al. gave a review from a computational mutation viewpoint on the mutation trend of hemagglutinin of influenza A virus [129]. Conenello et al. proved that a single mutation in the PB1-F2 of H5N1(HK/97) and 1918 influenza A viruses contributed to increased virulence [66]. Miotto et al. identified a catalog of sites as human-to-human transmission markers [130], and Chen et al. explored the mutation co-occurrence

in HA1 sequences and detected co-mutation sites under strong selective pressure, predicting the potential drifts with specific mutations of the viruses [110].

However, those works do not explore the virulence of influenza directly. To investigate the critical sites of virulent strains, a binary classification problem is defined. Sequences from recent pandemics are labeled as high virulent, while the seasonal flu or epidemics strains are labeled as mild or less virulent. By applying three typical rule-learning based classification algorithms, namely OneR, JRip and PART, the results indicate high accuracy classifying the strains. Besides, rules generated by those algorithms can provide insights on the crucial virulent sites and assist their detection. To evaluate the outcomes, they are compared with experimental results of animal models and it is found that 14 out of 16 detected sites are either located in epitope regions or relevant to the binding site to the host. Besides, the rest are also indicated to be highly potential to play an important role in the virulence of an influenza strain.

3.2 Materials and methods

3.2.1 Overview

The hemagglutinin (HA) sequences of past pandemic and non-pandemic strains were first collected with three subtypes of influenza, namely, H1N1, H2N2 and H3N2. Then the data cleaning and preprocessing were carried out to acquire qualified strains. Next, three benchmark rule-based algorithms were applied to classify pandemic strains from non-pandemic strains using H1N1 dataset with ten-fold cross-validation. The rules of classifying virulent pandemic strains and their corresponding critical sites were obtained. Finally, these sites were mapped into H2N2 and H3N2 pandemic strains based on amino acid distribution. The

workflow of this work is summarized in Figure 3.1.

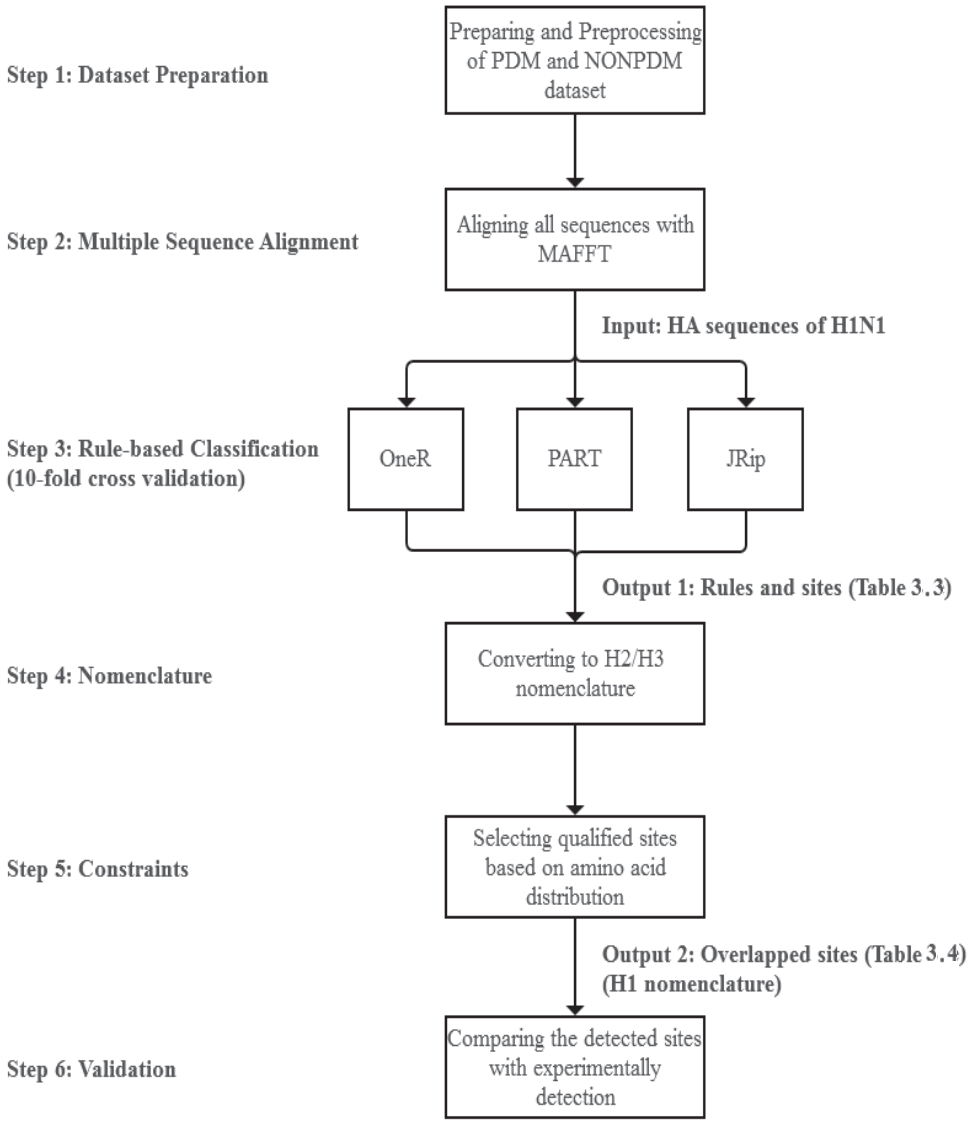


Figure 3.1: The flowchart for binary classification of pandemic and non-pandemic strains

3.2.2 Data collection

Influenza sequences were obtained from Influenza Virus Resource (IVR) on 31 Dec 2016 [131]. These HA sequences were retrieved with full length and human host from 1918 to 2016. The experimental strains can be categorized into two types, pandemic and non-pandemic classes. The pandemic class is denoted as PDM that includes sequences from five pandemics in recent centuries, namely H1N1 for 1918 Spanish flu, 1977 Russian flu and 2009 swine originated influenza virus, H2N2 for the 1957 Asian flu and H3N2 for 1968 Hong Kong pandemic. Viruses could circulate for a while even after the pandemic. The sequences from two consecutive years of each pandemic were also labeled as PDM sequence. Correspondingly, the dataset collected for the non-pandemic class denoted as NONPDM was made up of all the strains of subtype H1N1, H2N2 and H3N2 from 1918 to 2016 excluding the pandemic strains, during which there were only some small-scale outbreaks that happened seasonally. The datasets ended up with 3275 samples (collapsing identical sequences) for PDM class, with 3216 H1N1 strains, 35 H2N2 strains and 24 H3N2 strains respectively. Meanwhile, 8989 samples are obtained for the NONPDM class, with 3329 H1N1 strains, 45 H2N2 strains and 5615 H3N2 strains. The details of the datasets collected for each class are in Table 3.1.

3.2.3 Data cleaning and preprocessing

Due to the different lengths of each subtype of HA, the strains were classified from multiple clades of H1, H2 and H3. The PDM and NONPDM strains of H1N1 were first assembled. Multiple sequence alignment was implemented by MAFFT (Multiple Alignment using Fast Fourier Transform) with FASTA output

Table 3.1: Data collection of subtype H1N1, H2N2 and H3N2 in pandemic and non-pandemic years of recent centuries.

Period	Name	Circulating	Subtype	Number	Class
1918-1919	Spanish flu	1950s	H1N1	1	PDM
1957-1958	Asian	1960s	H2N2	35	PDM
1968-1969	Hong Kong	circulating	H3N2	25	PDM
1977-1978	Russian	circulating	H1N1	14	PDM
2009-2010	China	circulating	H1N1	3201	PDM
1920-1976	Seasonal	–	H1N1	54	NONPDM
1979-2008	Seasonal	–	H1N1	1072	NONPDM
2011-2016	Seasonal	–	H1N1	2203	NONPDM
1918-1956	Seasonal	–	H2N2	0	NONPDM
1959-2016	Seasonal	–	H2N2	45	NONPDM
1918-1967	Seasonal	–	H3N2	0	NONPDM
1970-2016	Seasonal	–	H3N2	5615	NONPDM

format in the same order as input samples. FFT-NS-2 mode was selected as the running strategy. The results showed many deletions and insertions of aligned sequences because of different HA lengths in H1N1, which turned out to have a great impact on the distribution and alignment of amino acid sites. The length of 566 amino acids was the most common length for a complete HA segment of H1N1 of the influenza A virus that was used as a criterion to eliminate the strains that have more than 566 residues. The remaining samples of H1N1 included 3201 PDM strains and 3327 NONPDM strains. Multiple sequence alignment was repeated by MAFFT with the same parameter settings. This process was also applied to subtype H2N2 and H3N2 samples.

A comparison of residues between subtypes of the influenza virus has been

increasingly utilized for comparative studies across subtypes. An analysis of N-terminal cleavage sites for thirteen subtypes of influenza A hemagglutinin sequence has been described by Nobusawa and colleagues [132]. F. Burke and J.Smith extended this work for eighteen subtypes [133]. The analysis of known structures of HA of influenza virus allows us to define structurally and functionally equivalent amino acids across all subtypes using a numbering system based on mature HA sequences. The N-terminal signal peptide cleavage site of HA was predicted using signals for all HA subtypes [134]. On account of this numbering system, it was able to unify equivalent sites across subtypes H1, H2 and H3 of influenza A viruses by deleting signal peptides. The three representatives with N-terminal sequences of mature HA proteins starting with "DTICIGYHANNS", "DQICIGYHANNS", and "QDLPGNDNSTATLCLGHHAVPN" are "A/California/04/2009/H1N1pdm", "A/Singapore/1/1957/H2N2" and "A/AICHI/2/68/H3N2", respectively [133].

For each sequence in the preprocessed dataset, every site was regarded as one feature by concatenating the residue and its position (for example T190). If there was an insertion or deletion, '-' would be used for replacement. As a result, there are 549, 547 and 550 features for every sequence of H1N1, H2N2 and H3N2 respectively, followed by the label at the end of each sequence that class PDM is denoted as '1' and NONPDM is '0'.

3.2.4 Binary classification

Machine learning has been widely applied in bioinformatics for many years and the diverse range of rapidly expanding data produced by modern molecular biology has fueled a need for accurate classification and prediction algorithms [135]. Compared with traditional machine learning approaches, rule-based meth-

ods suggest a more efficient way to manipulate knowledge and interpret information [136]. They are usually leveraged for identifying a set of context-dependent rules that collectively store and apply knowledge in a piece-wise manner to make predictions, which is selected for the binary classification task and detect critical virulent sites in PDM strains. However, there is no unique classifier or rule that can be used to outperform others for all problems. Hence, three different rule-based classification algorithms are applied, namely, OneR, JRip and PART, to integrate the results of all the classifiers to complement each other. A comparative analysis of their performance of different classifiers will also be described and all methods are implemented by Weka [137].

OneR, short for "One Rule", is a simple and accurate learning classification algorithm that generates one rule for each predictor. It only selects one rule with the smallest total error. The drawback of this method is that it can only identify a single feature that matters most in the dataset without informing more rules that are potentially related. The second algorithm JRip implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which is based on association rules with reduced error pruning, a very common and effective technique in decision tree algorithms. Rules are optimized by generating and pruning two variants of each rule from randomized data, from which the one with the minimal decision length is selected. Finally, rules that would increase the decision length were deleted from the ruleset and the optimal rule set is obtained. PART (Projective Adaptive Resonance Theory) is the third algorithm that uses partial decision trees to generate the decision combining two dominant rule mining strategies C4.5 and RIPPER, which is straightforward for global optimization by either discarding or adjusting individual rules.

Considering different lengths of H1N1, H2N2 and H3N2 features as well as sample distributions for the three subtypes, H1N1 samples manifested the best

quality with sufficient numbers of PDM and NONPDM strains in a balanced way, compared with the other two subtypes, for which there is a lack of enough PDM strains or imbalanced class distribution. Therefore, H1N1 samples are used as the input dataset. Ten-fold cross-validation was applied to identify the rules and sites that are related to pandemic strains in the H1N1 subtype. The detected sites were mapped across H2 and H3 subtypes with equivalent sites by a numbering system based on mature HA sequences. Overlapped sites were found using H2N2 and H3N2 strains based on amino acid distribution.

3.3 Results

The results show that these three classification algorithms have achieved competitive performance. To evaluate the performance of three classification algorithms, we use precision, sensitivity, specificity, F-measure and ROC area as metrics. We define that sequences in the PDM class are taken as positive instances, while the NONPDM sequences are negative. Thus, precision denotes the ratio of correctly identified PDM over all positive samples. To measure the proportions of positives and negatives that are correctly identified, sensitivity and specificity are also leveraged to better understand the performance of PDM and NONPDM strains classification. F-measure is the harmonic mean of precision and sensitivity. Table 3.2 shows the details of performance for the classifiers.

The rules were discovered by OneR, JRip and PART with H1N1 datasets. OneR selected only one feature, site 185 that gets the smallest total error with 0.883 precision. JRip discovered six rules for the classification of PDM strains with twenty different sites identified with 0.926 precision, which performed the best among these three methods. In the method of PART, which is based on partial decision trees, generating more rules than the previous two methods. Here

only the top two rules classified as PDM by PART are presented that occupied most of the PDM strains with the precision of 0.921, ignoring NONPDM rules that were less important with the identification of critical virulent sites. Details of rules and detected sites of these three methods are shown in Table 3.3.

On obtaining the sites from generated rules by using the HA sequences of H1N1 PDM and NONPDM datasets, these identified sites are transferred into equivalent ones that located in H2N2 and H3N2 by a numbering system based on mature HA sequence. Due to the limited strains in H2N2 (45 NONPDM strains and 35 PDM strains) and imbalanced strains in H3N2 (5614 NONPDM strains and 24 PDM strains), directly applying these rule-based algorithms in H2N2 and H3N2 datasets did not present many rules associated with critical sites, neither making consistent results with the identified sites in H1N1 experiments. Considering that, the sites identified in H1N1 are mapped into equivalent ones in H2N2 and H3N2 strains using a cross-subtype numbering scheme proposed by Burke and Smith [134]. As a result, the sites are selected that are related to the virulence of influenza strains using H2N2 and H3N2 datasets based on the following two constraint conditions. The first constraint is that the type of amino acid for each site in H2/H3 PDM strains is almost the same, but different from most amino acids in NONPDM strains in the same site. The second constraint is

Table 3.2: Performance using OneR, JRip and PART methods

Metric	OneR	Jrip	PART
Precision	0.883	0.926	0.921
Sensitivity	0.910	0.928	0.923
Specificity	0.858	0.925	0.919
F-measure	0.881	0.926	0.921
ROC area	0.880	0.952	0.971

Table 3.3: Rules and sites generated by OneR, JRip and PART that are associated with PDM and NONPDM strains.

	Rules $(x/y)^a$	Sites
OneR	G185→PDM, N185→PDM, S185→PDM, V185→PDM A185→NONPDM, D185→NONPDM, I185→NOMPDM, K185→NONPDM, M185→NONPDM, P185→NONPDM, R185→NONPDM, T185→NONPDM (5751/775)	185
JRip	(S185+E374+A186+S183)→PDM (1846.0/17.0) (S185+I216+V272+K163+S451+A134)→PDM (730.0/23.0), (T190+K274)→PDM (441.0/169.0) (E499+S143+D97+E374+S162)→PDM (91.0/17.0) (E499+S143+D97+V520+S162+P271)→PDM (192.0/66.0) (S451+T190+K274+D187+V153)→PDM (20/0) ~ [b] → NONPDM (3209.0/180.0)	97, 134, 143, 153, 162, 163, 183, 185, 186, 187, 190, 216, 271, 272, 274, 374, 451, 499, 520
PART[c]	(S185+K239+I216+T133+A521+N473+I372+K494+T281+ T25+K146+A423+L526+T13+A134+S143+V272+S451+, S263+N228+K169+S71+E491+K154+K163+K308+ G202+V234+H273+Q223) → PDM (2195.0/11.0) (E374+K458+V479+N31+D97+K409+N540+I266+R259+ K302+I510+S190+A197+S84) → PDM (339.0/12.0)	13, 25, 31, 71, 84, 97, 133, 134, 143, 146, 154, 163, 169, 185, 190, 197, 202, 216,223, 228, 234, 239, 259, 263, 266, 272, 273,281, 302, 308, 372, 374, 409, 423, 451, 458,473, 479, 491, 494, 510, 521, 526, 540

[a] x stands for the number of instances covered by the rule, while y is the number of misclassified instances.

[b] ~ means all the other features.

[c] Only two main rules are presented that identifying PDM class, for which the sites are important related to virulence.

the amino acid existing in the PDM strains of each site but not the majority of the second majority amino acid. After cross-type site transformation and selection, some critical sites generated by the rule-based classification are acquired that not only existing in the H1N1 subtype but also make sense in H2N2 or H3N2 subtypes. The selection sites of different subtypes are presented in Table 3.4.

Table 3.4: Summary of features in the generated rules across H1N1, H2N2 and H3N2 of PDM strains (H1 numbering system)

	Features(a)	(b/a)*
OneR	**185**	1/1
JRip	**134, 143, 153, 185, 186, 187, 273, 274,** 451	8/9
PART	**134, 143, 185, 190, 202, 216, 223, 239, 259, 273,** 451, 540	10/12

* a stands for the number of features included in the extracted rules and b is the number of features located in epitope regions. The features located on the epitope regions are in bold font. Features with reported effects but not on the epitope are underlined.

3.4 Discussion

Since the first emergence of pandemic after the nineteenth-century, influenza A viruses have been circulating and evolving up to now, during which several other pandemics occurred with the emergence of novel influenza strains, causing millions of human deaths and enormous economic loss, seriously affecting the public health. This study aims at identifying some critical sites that are related to the virulence of PDM strains. The analysis performed in this study using the HA gene of PDM and NONPDM sequences revealed the presence of sixteen distinct sites that may affect virulence, leading to the outbreak of the pandemic. To evaluate the sites identified by the algorithms, all the sites are mapped on the HA protein presented in Table 3.4 that 14 out of 16 are found in epitope regions [138]. The results visualized in Figure 3.2 also show that sites 134, 153, 186 and 187 identified by JRip and the sites 134 and 216 identified by PART are the receptor-binding sites among the 13 serotype HAs. The receptor specificity of HA has been considered as one of the determinants of the tissue tropism and host range in the influenza virus [139]. Specific mutations on sites 134, 187, 190 and 223 could lead to increased virus binding to alpha (2,6) linkage [133]. Also,

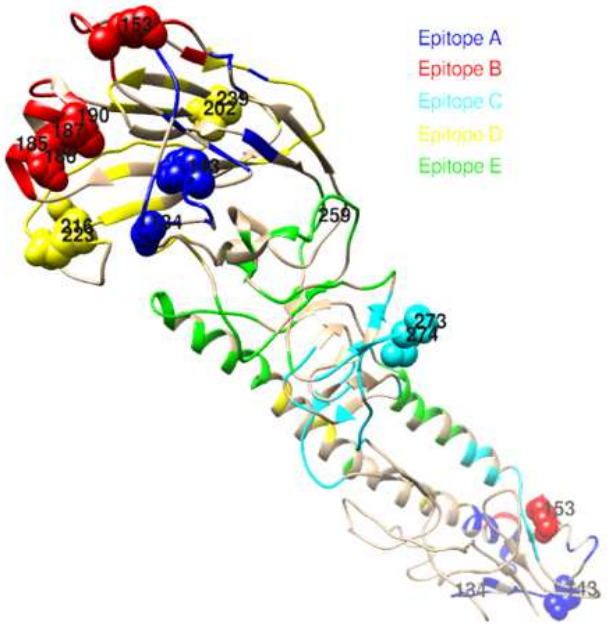


Figure 3.2: Mapping the detected sites onto hemagglutinin of the H1N1 virus (PDB ID: 4EDB). The blue, red, cyan, yellow and green regions represent the five epitope regions A-E respectively. The spheres with corresponding numbers are the detected sites locating at the epitope regions.

some more sites are in four major antigenic regions defined by Brownlee and Fodor [140] including sites 185, 186 and 187 from region Ca1 and site 202 from Sb. The accumulation of HA mutations by antigenic drift arising from population immune pressure is a significant cause of the emergence of new seasonal human influenza viruses [141].

Apart from the sites that are receptor binding sites or antigenic sites, there are several residue positions and mutations that could be indicated by other experimental results. For example, the amino acid change S143G, which occupied 29/55 (52.7%) of viruses analyzed during the 2009-2010 season, is located next to the antigenic region Ca [142]. Meanwhile, a previous report indicates that two

viruses with the S143G substitution were observed to have reduced antigenicity against the A/California/07/2009 vaccine virus in the HI test [143]. Some amino acid changes in H3N2, H1N1 and H1N2 viruses between seasons in site 239 suggest that this site could have been through the cross-type transmission. Mutation E274K occurring between 2004 and 2007 were maintained in the oseltamivir-resistant viruses that caused epidemics in the 2007-2009 seasons [144], which is also located next to the antigenic site 273. Substitution S451N, found by Antonio Piralla et al., may increase the severity of a symptom of infected patients [142]. Annotations and references on those sites are summarized in Table 3.5.

To sum up, three rule-based classification algorithms provide competitive performance in classifying pandemic strains from non-pandemic. By retrieving experimentally determined functional features and epitope regions, it is not difficult to find that the detected sites are highly overlapped with the experimentally determined functional sites, mostly affecting the antigenicity or receptor binding mechanisms with host cells. When it comes to the virulence, the adaption of an influenza A virus to recognize host cells and immune responses are always considered critical measurements. Here, this work focused on virulence to humans, and the binding mechanism to human cells will affect the pathogenicity to individuals and transmission efficiency among populations, both of which are key requirements of a pandemic. Avian influenza viruses prefer an alpha (2,3) sialic linkage, which is preferentially expressed on cells deeper in the human lungs besides the avian intestinal epithelium. Normally an avian influenza virus has little chance to infect humans, whose upper respiratory cells are mostly alpha (2,6) sialic linkages, but it will cause more acute symptoms once humans get infected. However, when the avian influenza virus has conferred the ability to bind the alpha (2,6) linked sialic acid, it would be more pathogenic and contagious. Another factor, epitope, the antigenic determinant of viruses, is the main target

Table 3.5: Summary of annotations of features bearing specific functions (H1 numbering system).

Features	Annotations	References
134	Receptor binding site, mutation S134A increased virus binding to alpha2-6 glycans	[133]
143	S143G observed reduced antigenicity against the A/California/07/2009 vaccine virus in the HI test	[143]
153	Receptor binding site	[133]
185	Antigenic site	[140]
186	Antigenic site, receptor binding site	[133] [140]
187	Antigenic site, receptor binding site, mutation E187G increased virus binding to alpha2-6 glycans	[133] [140]
190	Receptor binding site, mutation K190R increased virus binding to alpha2-6 glycans	[133]
202	Antigenic site	[140]
216	Receptor binding site	[133] [142]
223	Receptor binding site, mutation Q223L increased virus binding to alpha2-6 glycans	[133]
239	Amino acid changes in H3N2, H1N1, H1N2 viruses between seasons	[141]
273	Antigenic site	[140]
274	Mutation E274K were maintained in the oseltamivir-resistant viruses that caused epidemics during 2007-2009	
451	Substitution S451N may increase the severity of infected patients	[142]

recognized by the immune system. Thus, it makes a significant sense that sites on epitope regions are classifiers between the pandemic and the non-pandemic strains.

According to the experimental results, a few other sites detected, 451 and 540, are found outside the epitope regions or link to the binding of the human host. No significant mutations are identified at site 540 affecting pathogenicity, which may indicate it is not as important as the top ones. However, it should also be noted that they are still potential key sites in causing pandemics. More biological experiments are needed for validation on these sites with the guidance of computational techniques.

3.5 Chapter summary

In this chapter, a binary prediction task is defined and three rule-based machine learning algorithms have been applied to classify pandemics strains as well as identify the potential virulent sites using past pandemic strains, which prove to be a powerful and efficient way of biological knowledge discovery. The experimental results show a good performance with over 0.9 precision rate on average, detecting 16 potential virulent sites in total. Further validation has confirmed that 14 out of 16 sites are located in epitope regions or relevant to the binding of host cells, which contributes a better understanding for the rapid detection of genetic variants with the potential of causing pandemics or epidemics. The findings in this work will be leveraged for the construction of the virulence prediction model in the subsequent sections.

Chapter 4

HopPER: an adaptive model for probability estimation of influenza reassortment through host prediction

[3]Influenza reassortment, a mechanism where influenza viruses exchange their RNA segments by co-infecting a single cell, has been implicated in several major pandemics since the 19th century. Owing to the significant impact on public health and social stability, great attention has been received on the identification of influenza reassortment. A new computational method named HopPER (Host-prediction-based Probability Estimation of Reassortment) is proposed, that sturdily estimates reassortment probabilities through host tropism prediction using 147 new features generated from seven physicochemical properties of amino acids. The experiments are conducted on a range of real and synthetic datasets using HopPER and several other state-of-the-art methods for comparison. The results indicated 280 out of 318 candidate reassortants have been successfully identified. Additionally, not only can HopPER be applied to complete genomes but its effectiveness on incomplete genomes is also demonstrated. The analysis of the evolutionary success of avian, human and swine viruses generated through reassortment across different years using HopPER further revealed the reassort-

[3]Part of the work in this chapter has been published in [Yin R, Zhou X, Zheng J, et al. Computational identification of physicochemical signatures for host tropism of influenza A virus[J]. Journal of bioinformatics and computational biology, 2018: 1840023-1840023]

ment history. This chapter presents a new method for influenza reassortment estimation. The method could facilitate rapid reassortment detection and shed light on the evolutionary patterns of influenza viruses.

4.1 Introduction

Influenza A viruses, as highly infectious respiratory pathogens, can transmit across host species and evade host immune responses. As illustrated in Chapter1, a complete influenza genome consists of eight independent gene segments, where the subtype of influenza is characterized by the HA and NA segments [145]. Transcription and replication take place by the viral RNA-dependent polymerase complex PA, PB1 and PB2 [15]. The rest of the segments encode the NP, M1, M2 and two non-structural proteins. This segmented structure of the virus enables the exchange of different segments between influenza viruses when co-infecting a cell [146]. The mechanism of genetic recombination, named reassortment, may lead to the occurrence of novel progeny viruses [147].

It has been well recognized that reassortment is an evolutionary mechanism of segmented viruses that play important roles in the interspecies transmission and generating novel strains. The reassortment could accelerate the rate of acquiring new genetic markers that would faster overcome host barriers than the slow process of incremental accumulation of mutations [148]. In Chapter 2, we know that the outbreak of three major influenza pandemics was associated with the reassortment that produced new strains infecting humans since the 19th century [149]. In more detail, the evidence indicated that the HA and NA segments of the 1957 Asian pandemic were substituted by genes related to avian strains. The reassortment of human and avian strains led to the 1968 H3N2 pandemic that the H3 HA gene was derived from avian-origin viruses [47]. Besides, reassortment be-

tween two different swine influenza viruses, which themselves comprised genes from previous strains of different hosts, caused another pandemic in 2009 [7]. These pandemics have not only killed numerous people but also led to enormous economic losses. Therefore, early identification of influenza reassortment and potential reassortant strains are crucial for the surveillance and prevention of pandemics in the future.

With the rapid growth of flu data in recent years, increasing complete influenza genomes are publicly available [150]. Many efforts have been made to detect influenza reassortment events using the influenza genomic data. The common approach of identifying influenza reassortment is to construct fixed phylogenetic trees relating each segment of the strains [82] [151] [152]. Two methods were proposed for identifying reassortment events based on the difference between phylogenetic trees [92]. These trees are compared to detect disagreements of different strains, but it is a laborious and time-consuming process. Moreover, it provides no guarantee that all reassortments have been found. To account for the uncertainty in the inferred phylogenies, a novel computational method named GiRaF was developed to identify reassortment [90]. In GiRaF, large collections of Markov chain Monte Carlo sampled trees were searched for groups of incompatible splits. This successfully detected some known reassortments in avian, human and swine influenza strains. Yurovsky and Moret presented a fully automated flu reassortment finder called FluRF that employed a bottom-up search on the reconstructed phylogenetic trees of full and segment-based genomes [93]. However, the computational cost of phylogeny laid a formidable barrier for reassortment detection using phylogenetic analysis with a large scale of the dataset. Silva et al. aimed to solve this problem by formulating a phylogeny independent method that only utilized nucleotide distance matrices as input for reassortment detection [89]. Furthermore, Rabadan et al. provided a quantitative method to measure

the genetic shift from nucleotide sequence data that did not rely on phylogenetic analysis for reassortment detection [94]. Villa and Lässig determined rate and average selective effect of reassortment process in human influenza H3N2 using a new method to map reassortment events from joint genealogies of multiple genome segments [153].

Despite the growing data of genomic sequences and powerful computational capability for constructing various phylogenies to detect reassortment events, these approaches are generally applicable in a small scale of the dataset with well-defined phylogenetic trees. In this chapter, a novel approach named HopPER (Host-prediction-based Probability Estimation of Reassortment) is developed that employs machine learning techniques to calculate the reassortment probability by predicting the host tropism in a given collection of genomic sequences. HopPER first generates the feature vectors by seven physicochemical properties from avian, human and swine strains with global descriptors CTD (Composition, Transition and Distribution). It then applies a kernel perspective on host probability estimation by the random forest [154] for a single sequence and then combines all segments of the genome to produce an overall estimation of reassortment probability. The HopPER is tested on both real datasets and synthetic datasets for the evaluation of the capacity of estimating the reassortment possibility. Compared with several other computational methods, it is shown that HopPER has successfully identified reassortments with better precision. Furthermore, HopPER is efficient in detecting reassortment for even incomplete genomes (with at least two available genomic segments) and in analyzing large datasets. Thus, the development of HopPER can assist in flu surveillance and prevent future pandemics.

4.2 Materials and methods

4.2.1 Problem formulation

The concepts of reassortment are broadly applicable to other multipartite genomes, most of which have been studied. Here, only influenza reassortment is investigated in this chapter. As far as we know, the reassortant strains are responsible for the majority of flu pandemics in history and will continuously threaten public health. While any exchange of genetic material between different the segments of influenza viruses can be considered as reassortment. This work mainly focuses on identifying interspecies reassortments that have occurred across hosts. It is similar to definitions of host tropism predictors in the literature, except that here the problem is formulated probabilistically to enable a quantified estimate of host origin. Hence, host tropism is modeled by quantifying the reassortant probabilities. The model can also detect intra-host reassortments, for instance between different viral strains that originate from one single host category such as avian. In the model, the actual host in which the mixing occurred is disregarded and the focus is mainly on detecting past reassortants and the potential evolutionary relationships. For all practical purposes, only avian, human and swine strains are used that occupy the overwhelming majority of the existing sequence data. The following subsections respectively elaborate on the dataset and the structure of the proposed model. Figure 4.1 presents the flowchart.

4.2.2 Data collection and preprocessing

The amino acid sequences of all segments with avian, human and swine hosts are downloaded from NCBI on 31 Dec 2017 [131]. The datasets contain

influenza sequences with annotations such as accession number, host, collection year, region. Only full-length sequences are acquired and duplicate strains are removed from the collection. The results are presented in Table 4.1. The proteins of PB1-F2 and PA-X are excluded as they are completely contained in PB1 and

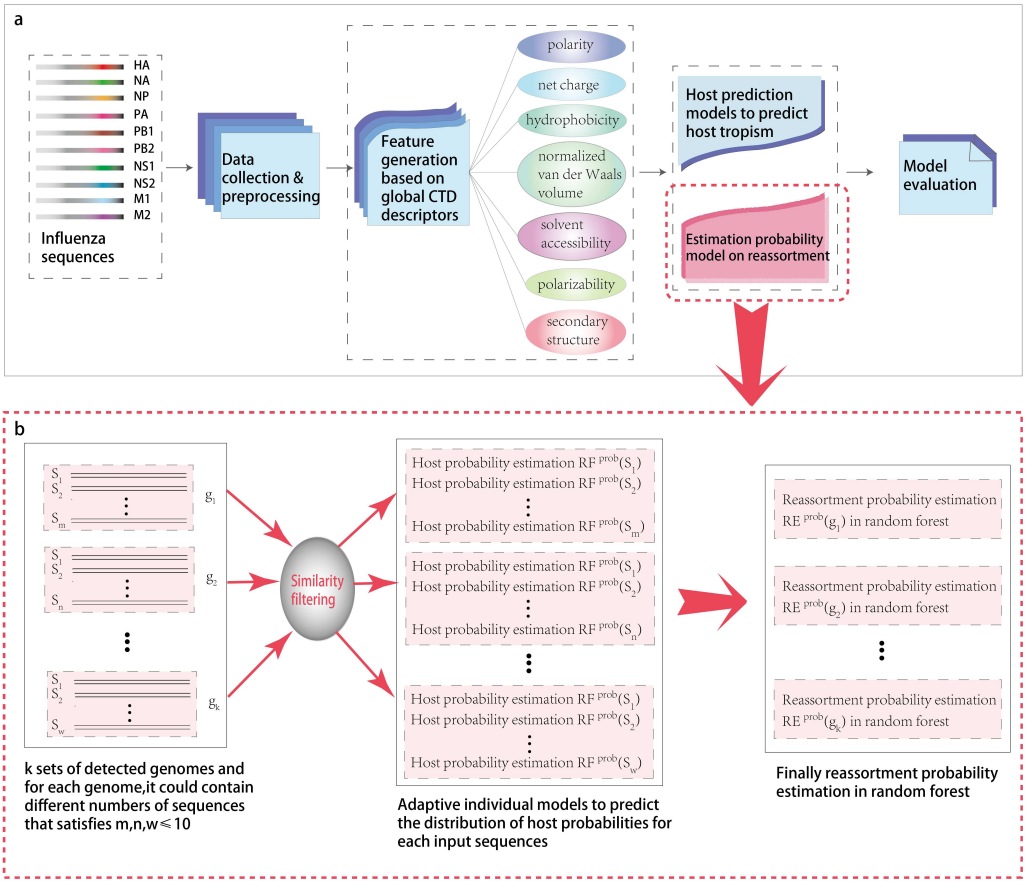


Figure 4.1: Schematic overview of analysis workflow in HopPER. a) The general diagram of the host prediction model based on seven physicochemical properties and reassortment probability estimation in the random forest. b) Specific algorithmic steps for estimation probability model on influenza genome reassortment detection.

Table 4.1: The number of influenza sequences for selected segments on avian, human, swine hosts and combined dataset.

Protein	Host type			
	Avian	Human	Swine	Combined
HA	12248	13607	6257	32112
NA	9452	10107	5734	25293
NP	4841	2659	2292	9792
PA	8428	5498	3059	16985
PB1	7699	4869	2892	15460
PB2	8106	5490	2901	16497
NS1	6115	4133	2662	12910
M2	2237	1404	1534	5175

PA respectively. It would be impossible for PB1-F2 and PA-X to have different host designation to PB1 and PA. Similarly, segment M consists of M1 and M2, while segment NS comprises NS1 and NS2. Only NS1 and M2 proteins are selected as representatives for host tropism prediction. This is because many more samples on NS1 and M2 are available to construct the model. Finally, the data of eight different proteins is obtained and the avian, human and swine sequences are labeled as '0', '1' and '2', respectively.

Besides, whole-genome datasets are also collected from NCBI on the same date and settings. To analyze the global patterns of reassortment events from the year 1918 to 2017, It is ended up with 13598, 20614 and 4380 complete and incomplete genomes of avian, human and swine hosts respectively after data preprocessing. Further analysis is performed to illustrate the potential reassortants using genomic sequences. Also, synthetic genomes are collected from the Global Initiative on Sharing All Influenza Data (GISAID) [155]. These strains

are synthesized from the laboratory and labeled as true reassortants that contain 87 complete genomes and 25 incomplete genomes to evaluate the performance of HopPER. The incomplete genomes have at least two different segments so that we could calculate the probability of host tropism for each segment and exert statistical probability estimation to identify the reassortment. Apart from synthetic genomes, The HopPER is also validated through real samples that have been tested by some state-of-the-art methods.

4.2.3 Feature transformation

The feature transformation of protein sequences is conducted based on AAindex, a database of amino acid physicochemical properties, substitution matrices and statistical protein contact potentials [156]. The method developed by Dubchak is performed to transform protein sequences into feature vectors [157]. The transformation is implemented by using three global descriptors: composition (C), transition (T) and distribution (D) to calculate the numerical values for each amino acid properties. The amino acid physicochemical properties contain polarity, net charge, hydrophobicity, normalized van der Waals volume, solvent accessibility, polarizability and secondary structure [158]. These amino acids are divided into three different groups based on the physicochemical properties of amino acid indices [159] that can be seen in Table 4.2. Meanwhile, the equations for three global descriptors are formulated as follows:

$$Composition = (\frac{C_{G1}}{N}, \frac{C_{G2}}{N}, \frac{C_{G3}}{N}) \tag{4.1}$$

$$Transition = (\frac{T_{G1G2}}{N-1}, \frac{T_{G1G3}}{N-1}, \frac{T_{G2G3}}{N-1}) \tag{4.2}$$

Table 4.2: The division of amino acid groups based on physicochemical properties and amino acid indices.

Attributes	Group 1	Group 2	Group 3
Hydrophobicity	Polar Q, E, R, K, D, N	Neutral G,P, H, A, S, T, Y	Hydrophobic C, V, F, L, I, M, W
Polarizability	0-1.08 S, D, G, A, T	0.128-0.186 C, Q, I, P, N, V, E, L	0.219- 0.409 Y, M, K, R, H, F, W
Normalized Van der Waals	0-2.78 S, C, G, A, T, P, D	2.95-4.0 E, Q, N, V, I, L	4.0-8.1 K, F, M, H, R, Y, W
Polarity	4.9-6.2 W, C, L, I, F, M, V, Y	8.0-9.2 T, G, P, A, S	10.4-13.0 K, N, H, Q, R, E, D
Solvent Accessibility	Buried A, I, F, C, G, L, V, W	Exposed R, K, Q, E, N, D	Intermediate M, S, P, T, H, Y
Secondary Structure	Helix E, A, L, M, Q, K, R, H	Strand V, I, Y, C, W, F, T	Coil G, N, P, S, D
Charge	Positive K, R	Neutral A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V	Negative D, E

$$Distribution = (\frac{D_{i0}}{N}, \frac{D_{i25}}{N}, \frac{D_{i50}}{N}, \frac{D_{i75}}{N}, \frac{D_{i100}}{N}) \tag{4.3}$$

Composition describes the percentage frequency of each amino acid property group across the entire protein sequence. N is the number of amino acids and C_{Gi} is the frequency of amino acid property of group *i* in the sequence. Transition characterizes the percentage frequency with which amino acid from a group is followed by another group denoted as T_{GiGj}. It means the property in group *i* is followed by group *j* or the other way around such that *i, j* = 1,2,3 and $G_i \neq G_j$. The third descriptor illustrates the distribution of each attribute in the sequence and D_i represents the percentage in these positions of the amino acid properties in group *i*. The distribution is based on the first, 25%, 50%, 75% and 100% of the amino acids for each attribute [157]. Therefore, 21, 21 and 105 new features are generated based on seven amino acid physicochemical and structural properties for global CTD descriptors respectively. In total, 147 amino acid feature vectors have been used to build the model for host tropism prediction.

4.2.4 Host tropism prediction

The experiments are first carried out on the host tropism prediction for selected proteins, which is the basis for the probability estimation of influenza reassortment. The effectiveness of host tropism prediction on influenza HA proteins and zoonotic strains prediction has been demonstrated by Eng et al. [160]. Previous work supplemented this work on the host prediction of human-adapted subtypes using random forest that achieved better results over other classifiers [161]. By constructing a multitude of decision trees, it applies the general technique of bootstrap aggregating to tree learners and then splits leaf nodes in the trees by random subset of feature space [154]. This comes at the expense of a

small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model [162]. To ensure the robustness of the constructed models, all the datasets are split into an independent training dataset and testing dataset with a ratio of 0.8:0.2. Ten-fold cross-validation is adopted to evaluate the training process with random forest classifier. To assess the ability of HopPER, the independent testing data is used to predict the host tropism. The evaluation metrics include accuracy, precision, recall, G-means [163] and Matthew's correlation coefficient (MCC) [164].

4.2.5 Construction of training data

In Figure 4.1b, for the given input genomes for reassortment detection, the genomes are split into segments. The host tropism prediction for each segment is performed by individual independent models with random forest. To reduce the overfitting of the model for host prediction, an algorithm named Ratcliff-Obershelp [165] is introduced and this method measures the similarity between input sequences and training sequences using gestalt pattern matching. The similarity between a pair of sequences ranges from 0 to 1. The threshold is set as 0.99 to filter the sequences from the training data that are similar to input sequences. The remaining sequences are utilized to train the host prediction model and construct HopPER. Removal of similar sequences establishes the independence of train and test datasets. It ensures the cross-validated results are a "true reflection" of model performance. and make HopPER adaptive to the distinct input genomes for reassortment detection.

4.2.6 Reassortment probability estimation

In the reassortment probability estimation, x_{ia} denotes the influenza sequence and y_j is the possible host. The variable x_{ia} represents the influenza protein type a in genome i. Here, a belongs to one of the selected proteins while the ordered elements in set j = 0,1,2 correspond to avian, human and swine hosts, respectively. To better calculate the reassortment probability, it is assumed that the distribution of pairs of influenza sequences and its host labels are independent and identical, that is x_{ia} and y_j are related according to an unknown conditional class probability function $\mathbb{P}(y_j|x_{ia})$. Typical classification is to discriminate whether $\mathbb{P}(y_j|x_{ia})$ $\geq$ 0.5 to predict the class of a new input sequence as described in the subsection of host tropism prediction above. However, the goal is to directly estimate the probability of host tropism for each protein in a genome.

To the best of our knowledge, there is no literature regarding reassortment probability estimation in random forest models. This is probably that virologists would usually check for reassortment by a homology search or by phylogenetic analysis of influenza segments. Meanwhile, a previous study has indicated that random forests are difficult to calibrate by standard calibration methods [166]. However, random forest achieves the best performance of estimation among machine learning classifiers after calibration [167]. Some other researchers have investigated the effect of utilizing corrected probability estimates in random forests by Laplace and m-estimates at the nodes have demonstrated its usefulness [168]. Though there still exists limited empirical evidence for the effect of random forest probabilities estimation [169], the framework of kernel regression in the random forest probability estimation produces better results [170].

Consisting of a collection of T un-pruned decision trees, where one tree is built from each bootstrap sample, random forests allow consistent estimation

of individual probabilities [171]. A tree is constructed by introducing recursive binary splits to the data based on the covariates and only a subset of covariates of predefined size $mtry$ is randomly selected at each node. The randomness in each tree is represented by a random variable $\theta \in \mathbf{\Theta}$, which is an indicator to index the trees in the forest. The class probability estimates for a terminal node are obtained by the relative frequency of the class in that terminal node. For example, the probability estimate of the tree for a new item is the class probability of the corresponding terminal node. The decision tree will partition the input space by the terminal nodes that would be denoted in the tree generated through $\theta \in \mathbf{\Theta}$, where a point x_0 belongs to $R_\theta(x_0)$. And the number of the samples in this node will be represented by $N_\theta(x_0)$. Under these assumptions, the probability estimation for a single tree at a point x_0 could be defined as function $f(\theta, x_0)$ formulated below.

$$f(\theta, x_0) = \sum_{i=1}^{n} \frac{\prod(x_i \in R_\theta(x_0))y_j}{N_\theta(x_0)} \tag{4.4}$$

A random forest is composed of a set of independent random draws θ_1,...,θ_t, and the associated trees $f(\theta_1, \cdot)$,...,$f(\theta_t, \cdot)$. In the case of host tropism prediction of influenza sequences, the probabilities are estimated by making the host label for each tree round($f(\theta_t, x_0)$) and counting the fraction of trees that vote for its class. The results are aggregated by averaging the probability estimates denoted by $RF^{prob}(\cdot)$ for the new input data over all trees (Figure 4.2). Here the function is defined as $RF^{prob}(x_0)$ that approximates the conditional class probability $\mathbb{P}(y_j|x_0)$, calculating the probability of each possible binding host for input sequence x_0.

$$RF^{prob}(x_0) = \frac{1}{T} \sum_{t=1}^{T} round(f(\theta_t, x_0)) \tag{4.5}$$

The random forest sustains significant basis for host tropism prediction of in-

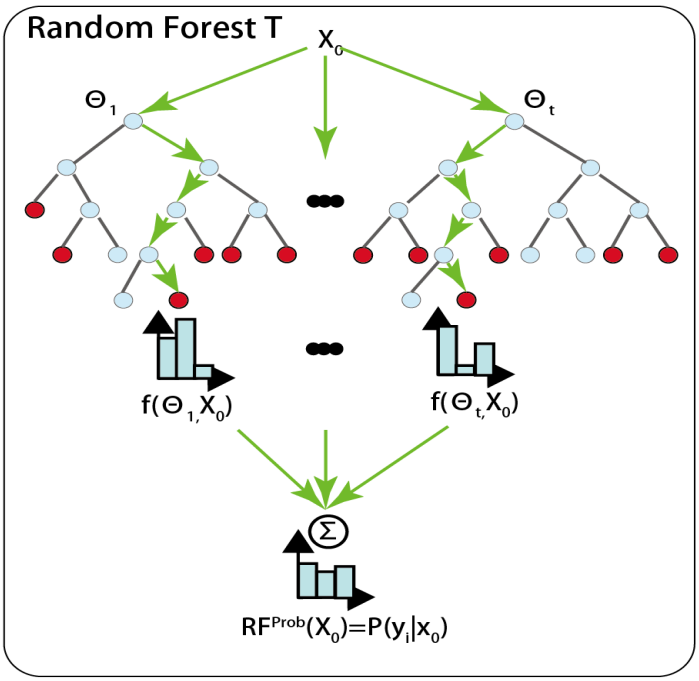


Figure 4.2: The structure of random forest T for probability estimation. θ_t is an independent random draw and $f(\theta_t, x_0)$ stands for the probability estimate by associated tree t at point x_0. $\mathbb{P}(y_j|x_0)$ characterizes the aggregation of conditional probability of all trees for label y_i.

fluenza sequences but cannot directly identify reassortment or reassortant strains. To perform reassortment probability estimation, it is necessary to know the original host types for all influenza sequences of the genome. In practice, the sequence in genome i is denoted as x_{ia} where $a \in S\{$HA, NA, NP, PA, PB1, PB2, NS1, M2$\}$. The probability estimation for certain host of protein is represented as x_{ia_s} and could be calculated by $RF^{prob}(x_{ia_s})$, that is, $\mathbb{P}(y_j|x_{ia_s})$, where y_i indicates different host labels and a_s is certain protein. For a candidate genome g_i containing n different proteins, a_n is used to denote the possible segments in g_i, where $a_s \subseteq a_n \subseteq S$ and $x_{ia_s} \in g_i$. The probability estimation of g_i being non-reassorted

could be represented as $NonRE^{prob}(g_i)$ and is formulated below.

$$\begin{aligned} NonRE^{prob}(g_i) &= P(y_j|g_i) \\ &= \prod^{a_n} P(y_j|x_{ia_s}) \end{aligned} \tag{4.6}$$

Taking all the available sequences into the calculation, the estimate of influenza reassortment probability is given as $RE^{prob}(g_i)$ shown in eq.7. Algorithm 1 clarifies the detailed steps of estimating reassortment probability. It not only allows the estimation of reassortment probabilities in complete genomes but also displays effectiveness in incomplete genomes. Random forest probability estimation provides a principled way to view the reassortments in terms of conditional probability functions. Hence, such a problem formulation motivates the discussion of estimation function as a fundamental and quantitative way to predict influenza reassortment. For instance, a prediction that an avian host origin is more likely than a human or swine host can narrow the sequence or homology search space for a virologist, given a sequence of interest.

$$RE^{prob}(g_i) = 1 - \sum_{j=0}^{2} NonRE^{prob}(g_i) \tag{4.7}$$

A genome is regarded as a reassortant strain if the estimated probability is greater than 0.5, otherwise, it is a non-reassortant strain. The true positive value (TPV) in equation (8) is leveraged to measure the ability of HopPER in reassortant detection. Apart from the detection of reassortment, HopPER can also predict the non-reassortant strains with the same principle. However, as far as we know, the study of non-reassortant strains attracts less attention and it is usually difficult to confirm or deny a strain without reassortment. Direct validation of true negative samples by HopPER poses great challenges. As an alternative, it is intended to sketch the contours of the distribution of reassortant strains

Algorithm 1 Probability estimation for influenza reassortment

Input: Training sequences $t_k \in G$, detected genome g_i that contains sequences x_{ia}

Output: Reassortment probability estimation $RE^{prob}(g_i)$

$n \leftarrow$ Number of sequences in genome g_i

for s = 1 to n **do**

$a_s \leftarrow$ Certain protein type of input sequences x_{ia}

if *Similarity*(x_{ia_s}, t_k) ¡= threshold **then**

$G^{'} \leftarrow$ Remove t_k from G

feature$(G^{'}_{a_s}) \leftarrow$ Feature generation on updated datasets $G^{'}$ on protein a_s

end if

for $m = 1$ to t **do**

$\theta_m \leftarrow$ Select independent random draw

$f(\theta_m, x_{ia_s}) \leftarrow$ Probability estimation for single tree based on *feature*$(G^{'}_{a_s})$

end for

$RF^{prob}(x_{ia_s}) \leftarrow$ Aggregate probabilities of t trees on protein a_s

$\mathbb{P}(y|x_{ia_s}) \leftarrow$ Obtain probabilities for different host labels y on a_s through $RF^{prob}(x_{ia_s})$

end for

$RE^{prob}(gi) \leftarrow$ Calculate the final reassortment probability of genome g_i by taking $\mathbb{P}(y|a_{s=1,\ldots,n})$ under reassortment rules

return $RE^{prob}(gi)$

across different years and analyze the rate variation of the evolutionary success of viruses generated through reassortment by HopPER. The reassortant strain rate (RSR) is defined as the ratio of reassortments that have occurred and the strains reproduced from the past reassortants to total genomic strains, which is a measurement of the subsequent evolutionary success of viruses generated

through reassortment. It could be calculated by identifying the reassortant and non-reassortant strains by HopPER. As a result, we could outline RSR variation by year and analyze the potential evolutionary patterns of the avian, human and swine strains.

$$TPV = \frac{number\,of\,correct\,predictions}{number\,of\,genomes} \tag{4.8}$$

4.3 Results and discussion

4.3.1 Performance of individual protein on host tropism prediction

After data preprocessing and feature generation for all available sequences from NCBI, individual prediction models for each protein were built by random forest. Table 4.3 presents the performance of predictive models for individual proteins on independent training and testing data. It is shown that the constructed models achieved outstanding performance in both training data and independent test data. In more detail, the HA model obtained the highest accuracy of 0.966 (G-means = 0.953, MCC = 0.943), whereas the lowest was M2 model with 0.876 accuracy (G-means = 0.854, MCC = 0.805) in the training set. Regarding independent test results, HopPER showed comparative performance with accuracy ranging from 0.865 to 0.965 for different proteins, which further demonstrated the robustness of the proposed models on host tropism prediction. Furthermore, it was also reported the predictive performance based on each class of avian, swine and human to help increase the confidence of HopPER (Table 4.4).

The results suggest that all the prediction models have presented high pre-

Table 4.3: Performance of host tropism predictive models for individual proteins on independent training and testing data.

Model	Training data					Testing data				
	Accuracy	Precision	Recall	G-means	MCC	Accuracy	Precision	Recall	G-means	MCC
HA	0.966	0.967	0.956	0.953	0.943	0.965	0.969	0.956	0.955	0.947
NA	0.961	0.962	0.953	0.953	0.939	0.957	0.958	0.95	0.949	0.933
NP	0.947	0.944	0.933	0.931	0.912	0.954	0.951	0.944	0.943	0.927
PA	0.929	0.916	0.893	0.89	0.881	0.922	0.906	0.892	0.888	0.875
PB1	0.931	0.927	0.907	0.902	0.887	0.937	0.933	0.914	0.912	0.898
PB2	0.943	0.937	0.912	0.913	0.906	0.945	0.938	0.923	0.921	0.911
NS1	0.934	0.928	0.917	0.916	0.896	0.931	0.93	0.919	0.917	0.896
M2	0.876	0.866	0.856	0.854	0.805	0.865	0.86	0.853	0.848	0.795

dictive performance, capable of classifying avian, human and swine strains. The viruses transmit between different host species, which allows for the mixture of gene segments and produces reassortant strains. This might enhance the pathogenicity of the virus, assisting novel strains to adapt to new host species [172]. However, it is still a challenge to directly predict the interspecies transmission and identify the capability of an avian strain to cross the species barrier and infect humans. But the results have proved the effectiveness of all models in predicting host tropism, which paves the way for further reassortment probability estimation through host prediction.

Table 4.4: The predictive performance of host tropism using HopPER based on each class of avian, human, and swine, labeled as '0', '1' and '2', respectively.

Model	Class	Training data				Testing data			
		Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
HA	0	0.98	0.99	0.98	9849	0.97	0.98	0.98	2399
	1	0.96	0.97	0.96	10897	0.95	0.98	0.96	2710
	2	0.97	0.9	0.94	4943	0.98	0.9	0.94	1314
NA	0	0.97	0.99	0.98	7613	0.97	0.98	0.98	1839
	1	0.95	0.96	0.96	8101	0.94	0.96	0.95	2006
	2	0.97	0.91	0.94	4520	0.96	0.9	0.93	1214
NP	0	0.96	0.99	0.97	3867	0.96	0.99	0.98	974
	1	0.93	0.9	0.92	2120	0.95	0.92	0.93	539
	2	0.93	0.91	0.92	1846	0.94	0.93	0.93	446
PA	0	0.95	0.98	0.97	6783	0.95	0.99	0.97	1645
	1	0.9	0.93	0.91	4351	0.92	0.9	0.91	1147
	2	0.89	0.78	0.83	2454	0.85	0.79	0.82	605
PB1	0	0.95	0.98	0.97	6175	0.95	0.99	0.97	1524
	1	0.9	0.92	0.91	3881	0.92	0.92	0.92	988
	2	0.93	0.81	0.87	2312	0.93	0.84	0.88	580
PB2	0	0.96	0.99	0.97	6499	0.96	0.98	0.97	1607
	1	0.92	0.94	0.93	4348	0.93	0.95	0.94	1142
	2	0.93	0.82	0.87	2350	0.92	0.84	0.88	551
NS1	0	0.96	0.98	0.97	4922	0.96	0.98	0.97	1193
	1	0.91	0.92	0.91	3300	0.91	0.92	0.91	833
	2	0.92	0.85	0.88	2106	0.93	0.85	0.89	556
M2	0	0.93	0.96	0.94	1812	0.89	0.96	0.93	425
	1	0.82	0.78	0.8	1092	0.87	0.75	0.8	312
	2	0.83	0.84	0.83	1236	0.82	0.85	0.83	298

4.3.2 Evaluation on real datasets

To measure the effectiveness of HopPER, several other independent influenza datasets are tested on HopPER. Genomic sequences for 18 typical reassorted H1N1, H1N2 and H3N2 genomes isolated from pigs in North America were studied in Karasin *et al.* [84], [85], [83]. The datasets of 16 resembled novel 2009 swine-origin isolates, 6 triple reassortant H3N2 strains throughout Canada and 39 reassortment events in swine influenza strains were constructed from a large-scale whole-genome sequences [86], [87], [88]. More comprehensively, 36 well supported candidate reassortants [89], 93 single-taxa and multi-taxa reassortment candidates [91], [90] were also selected for validation. A brief description of these genome datasets follows below.

1) Karasin *et al.* investigated the genetic characterization of H1N1, H1N2 and H3N2 viruses that circulated in North America from 1997 to 2005. Due to the occurrence of influenza pandemics in 1957 and 1968, the reassortant human/avian viruses have circulated in the world and been collected from pigs in Europe [173]. After that, the wholly avian H1N1 viruses adapted to the swine population and spread to North America with classical swine influenza virus [174], [175], [176]. The results obtained by genetic and phylogenetic trees indicated the evidence for wholly human and reassortant virus genotypes.

2) Kingsford *et al.* dataset contained 16 genome sequences that were similar to swine-origin influenza viruses (S-OIV) that appeared in Thailand. These swine-origin isolates only caused one cause of human infection A/Thailand/ 271/2005 (H1N1). The comparison between these earlier resembled S-OIV reassortant strains and the real S-OIV strains could facilitate the identifica-

tion of reassortment patterns.

3) Olsen *et al.* dataset contained 6 triple reassortant H3N2 viruses isolated from pigs and turkeys throughout Canada in 2005. These were human classical swine/avian reassortants similar to the viruses in 1998 except a distinct human-lineage NA segment, which suggested a fast and complicated interspecies transmission of reassortants.

4) Khiabanian *et al.* dataset was used to explore the process and patterns of viral reassortment with 39 complete and incomplete genome sets. The analysis of reassortment phenomena in swine influenza viruses was performed by several statistical techniques. The finding indicated that not only the surface glycoprotein coding proteins (HA and NA) but also the PB1 segment reassorted more frequently compared with other segments in swine viruses.

5) de Silva *et al.* dataset presented 36 well-supported candidate reassortants with strong confidence. The results indicated that the combination of novel HA and/or NA genes from different circulating viruses led to reassortment events. Reassortment patterns of the identified strains drive us to outline a more well-rounded picture of the evolution of some previously existing reassorted strains.

6) Nagarajan *et al.* presented a more comprehensive reassortment study including human, avian and S-OIV influenza populations. The framework GiRaF was applied to detect the sets of taxa based on a fast biclique enumeration algorithm.

HopPER is compared with the above 6 described computational methods by their ability to detect reassortment on real test datasets. Table 4.5 presents the

Table 4.5: The results of reassortant strains identified using HopPER that was validated by alternative methods for reassortment analysis.

Datasets	Number of genomes	Original Methods	Identified number by HopPER	TPV
Karasin *et al.*	18	Genetic and phylogenetic analyses with cycle sequencing and amplification by reverse transcription-PCR.	16	0.889
Kingsford *et al.*	16	Enumerating maximal bicliques with a defined incompatibility graph to detect high-probability inconsistencies between the distributions of trees.	14	0.875
Olsen *et al.*	6	Phylogenetic analysis by the method of maximum parsimony with bootstrap resampling for the genetic characterization of reassortant H3N2 viruses.	6	1.000
Khiabanian *et al.*	39	Applying statistical methods such as diversity and entropy measures of each segment and its correlations to investigate reassortment patterns.	33	0.846
de Silva *et al.*	36	Comprehensive analysis based on neighborhood of each segment and using only nucleotide distance matrix as input to formulate the phylogeny.	29	0.806
Niranjan *et al.*	93	Graph-incompatibility based reassortment finder that searches large collections of Markov chain Monte Carlo-sampled trees for groups of incompatible splits using a graph mining technique.	80	0.860

results of the number of reassortants identified by HopPER and other methods. The threshold is set as 0.5 to classify the reassorted and non-reassorted strains in HopPER. The results have demonstrated that HopPER easily picked up reassortants where the strains varied in hosts across different periods. Overall, 178 out of 208 strains were successfully detected as reassortants. Looking at outcomes in each dataset, it is apparent that all the similar swine-origin H3N2 influenza strains were recognized as reassortants. Perhaps the number of test genomes on this dataset was not significant and the TPV was only 0.806 on de Silva *et al.* dataset, slightly worse than other datasets. Some of the reassortant strains identified by Silva *et al.* were reported for the first time. This could decrease the confidence of the candidates as true reassortant strains. Nevertheless, the evaluation of the real datasets displayed strong evidence for the characterization of reassortment by HopPER, e.g., the validation on Nagarajan *et al.* dataset achieved TPV of 0.860, which contained a larger quantity of genomes with a diversity of strains.

One of the most critical strains A/CALIFORNIA/04/2009 was estimated to be reassortant with the probability of 0.885. Of particular interest was the potential host adaptation for individual segments of the genome. Selected avian, human and swine genomic strains are shown in Table 4.6, indicating the reassortment patterns based on host tropism and reassortment probabilities. The results incorporated the most likely host adaptation for each protein. Most of the reassortants displayed a diversity of host adaptation of influenza sequences in the genome. Table 4.6 indicates that more than one host species exists in all genomes except the strain A/domestic teal/Hunan/79/2005, which is estimated as a reassortant with the probability of 0.701, with the host tropism for each segment being the same. Another finding was that the reassortment probability of strain A/domestic teal/Hunan/79/2005 was not high compared with others. It may infer that interspecies transmission of influenza viruses had a direct impact

Table 4.6: Reassortment patterns on host distribution of selected avian (0), human (1) and swine (2) strains and the gap '-' denoted the missing sequence in the genome.

	Strain	Subtype	HA	M2	NA	NP	NS1	PA	PB1	PB2	RE^{prob}
Avian	A/domestic teal/Hunan/79/2005	H5N1	0	0	0	0	0	0	0	0	0.701
	A/pekin duck/California/P30/2006	H4N2	0	0	0	0	0	2	0	0	0.856
	A/mallard/Pennsylvania/454069-12/2006	H5N4	0	0	1	0	0	0	0	0	0.804
	A/chicken/Hubei/C1/2007	H9N2	0	2	0	0	2	2	0	0	0.976
Human	A/California/05/2009	H1N1	1	1	1	1	-	1	1	1	0.888
	A/Texas/05/2009	H1N1	1	2	2	1	1	1	1	1	0.993
	A/California/04/2009	H1N1	1	2	1	1	1	1	1	1	0.885
	A/New Jersey/1976	H1N1	2	0	1	1	1	1	1	1	0.984
Swine	A/Thailand/271/2005	H1N1	1	-	1	0	2	2	2	2	0.995
	A/swine/Ontario/00130/97	H3N2	2	1	1	2	2	1	1	2	1
	A/swine/Ontario/53518/03	H1N1	2	2	2	2	2	2	1	2	0.959
	A/swine/Hong Kong/273/1994	H1N1	1	2	1	1	2	2	1	2	0.999

on probability estimation. Correspondingly, more credible reassortment events would be obtained if the sequences in the genome stemmed from different species can be demonstrated.

Reassortant strains are implicated in several major pandemics in history with reassortments occurring across different hosts. An example is a swine-origin reassortant, which comprises genes derived from avian, human and classical swine [7]. More attention is needed for the reassortant strains when the complement of individual protein sequences are from three or more different host species detected by HopPER. Besides, the emergence of novel HA segment in a reassorted genome is crucial for the outbreak of potential pandemics that has to

be considered.

Moreover, It can further identify latent breakdowns in the ancestry of known reassortants and give insights for interspecies transmission and evolution of influenza viruses. For example, in A/swine/Ontario/53518/03, It is found that the segment PB1 was derived from human influenza virus lineages whereas the rest of the segments were in classical swine lineage [83]. The H3N2 viruses recovered from Canada in January 1997 like A/swine/Ontario/00130/97 from Ontario isolates, which were regarded as wholly human influenza viruses [84]. It was consistent with the experimental results that four segments M2, NA, PA and PB1 originated from human influenza viruses, suggesting strong interspecies transmission of the different clades. Similarly, the highly pathogenic avian influenza (HPAI) H5N1 lineage has demonstrated various combinations of its genes to form several generations of multiple reassortants [177]. The precursor of H5N1 strain A/Goose/Guangdong/1/96 and the re-emerging strain A/peregrine falcon/Hong Kong/2142/2008 were reassortants with probabilities 0.748 and 0.546 respectively. The complex reassortment mechanism and the manifold possibility of combination could adversely affect the host tropism prediction and overestimate the probability of reassortment, but HopPER has manifested the robustness of its capability to identify reassortment and also provided perspectives for evolutionary patterns.

4.3.3 Evaluation on synthetic datasets

To further verify the model's ability to identify induced reassortants and assess performance in a controlled setting, experiments on lab-synthesized reassortant strains were carried out. These synthetic strains were regarded as the true label on the detection of reassortants. The synthetic dataset was divided into

complete and incomplete genomes that contained 85 and 25 samples respectively. According to the rules, the data of incomplete genomes contained two different sequences at least. The results of reassortment detection on both complete and incomplete strains by HopPER were summarized in Table 4.7. HopPER correctly identified 19 out of 25 reassortants for incomplete genomes and 83 out of 85 reassortants for complete genomes on synthetic strains. Though the incomplete information of genomes likely influenced the prediction of reassortment, the TPV achieved by HopPER on the laboratory dataset (0.927) was more persuasive compared with the real dataset (0.855). On observation, the false positives reported by HopPER were dominated by incomplete samples. It was found that all these false positives only contain HA and NA proteins while most of the rest of incomplete genomes have more than two different segments (Table 4.8). In general, it can be inferred that the number of available segments in a genome is a critical factor impacting the reliable estimation of reassortment probability. Despite this, the false positive rate was still less than 0.1 on synthetic datasets.

Table 4.7: The number of predicted reassortant strains identified by HopPER for complete and incomplete genomes in both real and synthetic datasets.

Genomes	Integrity of genome	Predicted reassortants /total number
Real	Complete	154/173
	Incomplete	24/35
Synthetic	Complete	83/85
	Incomplete	19/25

Table 4.8: The reassortant strain names and their reassortment patterns of incomplete synthetic strains that '0' is avian host, '1' is human host, '2' is swine host and '-' stands for - sequences.

Index	Strain name
0	A/REASSORTANT/X157(NEW YORK/55/2004 X PUERTO RICO/8/1934)
1	A/REASSORTANT/NYMC X-185
2	A/REASSORTANT/NYMC X-197
3	A/REASSORTANT/IVR-148(BRISBANE/59/2007 X TEXAS/1/1977)
4	A/REASSORTANT/CBER_RG1(DUCK/LAOS/3295/2006 X PUERTO
5	A/REASSORTANT/IDCDC-RG18(TEXAS/05/2009 X NEW YORK/18/2009 X PUERTO RICO/8/1934)
6	A/REASSORTANT/IGYRP16(CALIFORNIA/07/2004 X PUERTO RICO/8/1934)
7	A/REASSORTANT/IDCDC-RG56B(HONG KONG/125/2017 X PUERTO RICO/8/1934)
8	A/REASSORTANT/IDCDC-RG20(TEXAS/05/2009 X PUERTO RICO/8/1934)
9	A/REASSORTANT/NYMC X-183
10	A/REASSORTANT/IDCDC-RG56N(HONG KONG/125/2017 X PUERTO RICO/8/1934)
11	A/REASSORTANT/IDCDC-RG13(EGYPT/3300-NAMRU3/2008 X PUERTO RICO/8/1934)
12	A/REASSORTANT/RESVIR9(NANCHANG/933/1995 X PUERTO RICO/8/1934)
13	A/REASSORTANT/X-99
14	A/REASSORTANT/X161(WISCONSIN/67/2005 X PUERTO RICO/8/1934)
15	A/REASSORTANT/RESVIR17(PANAMA/2007/1999 X PUERTO RICO/8/1934)
16	A/REASSORTANT/IVR147(BRISBANE/10/2007 X PUERTO RICO/8/1934)
17	A/REASSORTANT/JLUMV_RG1(CHICKEN/EGYPT/VSVRI/2009 X PUERTO RICO/8/1934)
18	A/REASSORTANT/IDCDC-RG42A(SICHUAN/26221/2014 X PUERTO RICO/8/1934)
19	A/REASSORTANT/NIBRG268(ANHUI/1/2013 X PUERTO RICO/8/1934)
20	A/REASSORTANT/X-73
21	A/REASSORTANT/IDCDC-RG56B
22	A/REASSORTANT/NYMC X-191
23	A/PHILIPPINES/2/82/BS [A/PR/8/24 X A/PHIL/2/82 REASSORTANT]
24	A/REASSORTANT/X147(WYOMING/3/2003 X PUERTO RICO/8/1934)

Index	HA	M2	NA	NP	NS1	PA	PB1	PB2
0	-	0	-	0	0	2	0	0
1	-	0	-	1	1	-	-	-
2	-	0	-	-	1	-	-	-
3	1	0	1	1	-	1	1	1
4	0	-	1	-	-	-	-	-
5	1	-	1	-	-	-	-	-
6	1	1	1	-	0	1	0	0
7	0	-	0	-	-	-	-	-
8	1	-	1	-	-	-	-	-
9	1	0	1	1	1	1	-	1
10	0	-	0	-	-	-	-	-
11	0	-	1	-	-	-	-	-
12	-	1	-	2	0	0	0	0
13	1	0	1	1	1	1	-	1
14	-	2	-	0	1	2	0	0
15	-	2	-	0	1	0	1	0
16	-	1	-	2	0	2	1	2
17	0	-	0	-	-	-	-	-
18	1	-	0	-	-	-	-	-
19	0	-	0	-	-	-	-	-
20	1	0	1	1	1	1	-	1
21	0	-	0	-	-	-	-	-
22	1	0	1	1	1	-	-	-
23	-	0	-	-	1	-	-	-
24	-	2	-	2	0	0	0	2

Detecting the reassortment is usually hard or impossible by either HopPER or other methods if the input genome is incomplete. It poses great challenges for any other computational tools to identify reassortment events with lots of missing information in the genome. Estimating the probabilities without constraining the integrity of genomes enables us to explore the reassortant strains in synthetic genomes. Though the reassortment analysis on incomplete genomes brings uncertainty of probability estimation and increases the difficulty of identifying reassortment, the results are not greatly affected using HopPER. 24 out of 35 and 19 out of 25 incomplete strains have been successfully identified in real and synthetic datasets respectively. The TPVs of reassortment detection on incomplete strains have achieved noteworthy performance in comparison to complete ones. However, a look into the unsuccessful cases of incomplete strains finds most of the failures in genomes with only 2 segments. The predicted reassortant strains by the number of available sequences in the genome are listed in Table 4.9. It demonstrates the effectiveness of HopPER in predicting the reassortment of incomplete strains.

4.3.4 Analysis on reassortment history

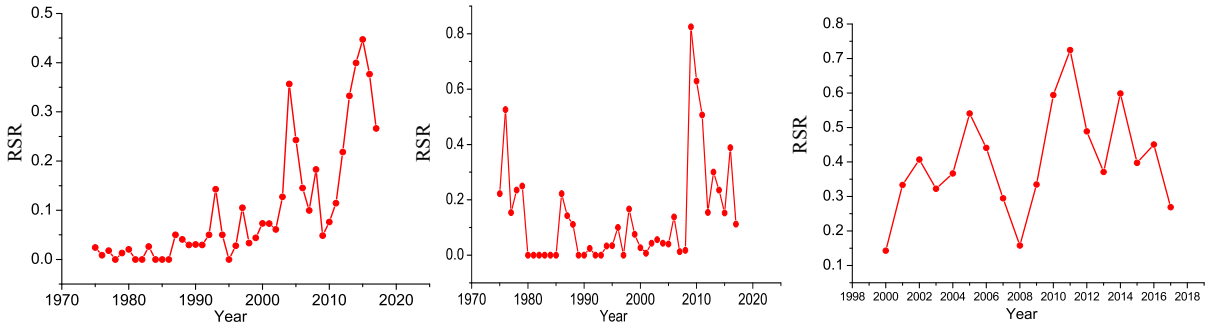
Influenza viruses have caused substantial morbidity in humans since the emergence of the Spanish pandemic [178]. Despite the long-term existence of the viruses, the influence of the reassortment in the expected transmission properties of influenza viruses is still an area of active research. A study on 71 representative complete genomes sampled between 1918 to 2006 showed reassortment occurred frequently throughout the evolutionary history of the virus [179]. Though some reassortment events would not cause severe infections or lead to outbreaks, reassortment still plays an important role in the process of evolution and epi-

Table 4.9: The number of predicted reassortant strains identified by HopPER in the case of different number of available sequences contained in the genome.

Available sequence of a genome	Number of test strains (real and synthetic datasets)	Number of reassortant identified by HopPER
2	14	6
3	4	4
4	6	4
5	8	7
6	15	12
7	13	10
8	258	237

demiology for influenza viruses, particularly when considering transmission from avian or swine host populations into human populations. For example, pigs have been known as a mixing vessel with multiple reassortment events occurring. While most of the cases were mild to humans, three out of four pandemics are related to the reassorted swine strains. The reassortment between influenza viruses from different host species can generate novel pandemic-potential strains. These antigenic and genetic novel strains are usually not well matched to the contemporaneous vaccines, and so existing vaccines offer little protection [82]. Detecting reassortment frequency among influenza viruses is also a crucial aspect to capture evolutionary history [180].

The HopPER is applied to investigate the reassortment history on avian, human and swine species respectively. RSR is utilized to illustrate the variety of reassortant strains. Figure 4.3 presents the RSR of influenza strains on three distinct species across different years. The experiments are conducted on the years with more than 20 genomes. The results reflect the complex reassortment histories and



(a) Avian (b) Human (c) Swine

Figure 4.3: The reassortant rate of three distinct host species of influenza strains across different years detected by HopPER. (a) The reassortant rate of avian species from 1988 to 2017. (b) The reassortant rate of human species from 1975 to 2017. (c) The reassortant rate of swine species from 2000 to 2017.

suggest the reliability compared with the actual evolutionary patterns. In Figure 4.3a, the RSR sustains a relatively low level until 2004, when the HPAI H5N1 virus has re-emerged [181]. These viruses have managed to transmit through both human and avian hosts, leading to novel reassortant strains [182]. The human species describes the different situations in which RSR reaches the local peaks around the pandemic years. The RSR starts to decrease after the outbreak of the 1976 pandemic when a new H1N1 strain predominated. After that, another pandemic caused by a triple reassortant swine-origin human strain occurred in 2009, when there is a rapid increase of RSR in Figure 4.3b. The RSR of swine species varies differently from avian and human and it gains high value, except in 2008. It can be inferred that the swine species, as the mixing vessel, more frequently participate in the reassortment process with both avian and human strains. According to Figure 4.3, the RSR remains at a relatively high level after 2009. This is because the progeny strains of these 2009 strains are still circulating

the world. Haemagglutination inhibition (HI) tests with post-infection ferret antisera indicate that the majority of A(H1N1)pdm09 viruses are antigenically homogeneous and closely related to the vaccine virus A/CALIFORNIA/7/2009 [183]. It is noteworthy as a possible indication for the resurgence of another potential pandemic or epidemic after the current reassortant strains have been in circulation.

4.4 Chapter summary

HopPER has been proposed for the probability estimation of influenza reassortment based on host prediction. Though the development of HopPER mainly focuses on influenza datasets, it could also be helpful for the research of other viral datasets that contain different host species. Demonstrated by different real and synthetic datasets, HopPER is validated by the comparison with alternative methods for reassortment detection. Meanwhile, it can also be leveraged to detect any known complete or incomplete strains for reassortment identification and reassortant strains with robustness, which makes it possible to monitor the transmission of influenza viruses and assist flu surveillance. This chapter witnessed the success of reassortment detection through HopPER, which paves the basis for the next chapter. More specifically, HopPER not only achieves a high true positive rate of identifying reassortant strains. but also calculates the estimated probabilities of viral genomes being reassortant. The results would be utilized as prior knowledge incorporated into the input data to establish the model for influenza virulence prediction.

Chapter 5

Virulence prediction of influenza A viruses with prior mutation and reassortment knowledge using all 8 segments

Influenza viruses greatly pose threats to public health and cause economic loss. Previous work has been investigated to reveal the viral factors that influence the virulence of influenza viruses. However, none of the existing work explicitly predicting influenza virulence. In this chapter, a novel and general framework is proposed for virulence prediction tasks by incorporating prior mutation (Chapter 2 and 3) and reassortment (Chapter 4) information into predictive models using all 8 segments. Firstly, 488 samples of influenza A virus infections are collected that are classified as virulent and avirulent labels using mouse lethal dose (MLD) 50. The posterior regularization technique is applied to convert prior mutation and reassortment knowledge into constraint features. Three deep learning-based approaches are leveraged for modeling, incorporating transformed constraint features. Experimental results on the collected dataset validate the effectiveness of the proposed framework for virulence prediction compared with other baselines. Moreover, it can display the importance weights of the prior viral information, which suggests the biological significance of the model and provides assistance for the accurate detection of influenza virulence.

5.1 Introduction

Influenza virus can easily infect the respiratory and cause a highly contagious disease, composed of 8 single-stranded RNA segments [145]. The viruses possess mutability and high frequency of genetic reassortment [148] [184]. They will have a great ability to be virulent and are in high mortality and morbidity infecting humans, which usually leads to epidemics. As illustrated in Chapter 1, apart from seasonal influenza epidemics, when millions of people are infected worldwide and up to 500,000 people are killed every year [5], influenza viruses are responsible for severe outbreaks in history. The pandemic strains are still circulating among humans and continuously cause recurrent epidemics. Apart from the strains that caused epidemics and pandemics, it is found that some other subtypes i.e., H5N1, H6N1, H7N9 and H9N2, etc. have also infected on humans [185] [186]. Among them, H5N1 and H7N9 suggest a high fatality rate on infected humans that has the potential to lead to the outbreak for the next epidemics even pandemics [187]. Overall, the detection of virulence level is needed to estimate the lethality of viruses and facilitate flu surveillance for better precautions.

Influenza virulence indicates the degree of pathogenicity and the capacity of the viruses to cause disease, which has a direct correlation to lethality on infected humans. Mutations and reassortment in the influenza viruses will cause antigenic drift and shift that makes the protein unrecognizable to pre-existing host immunity, which will increase the viral virulence. The signatures that influence the virulence of influenza A viruses have been investigated in many methods. For example, the mutations in the region 130-loop, 190-helix and 220-loop during the adaptation of influenza A virus have increased its virulence [55]. The mutations E627K and D701N in protein PB2 have also been considered as biological

markers for the virulence of influenza viruses [62]. Surprisingly, it was shown that a single mutation N66S on PB1-F2 protein has contributed to the increased virulence [66]. The dual mutation S224P and N383D in PA protein caused the increase of polymerase activity [64]. The evidence suggested that the substitutions Q222L and G224S have contributed to the outbreak of the 1957 and 1968 pandemics, which changed the receptor binding of H2 and H3 avian influenza binding specificity to alpha (2,6) linked sialic acid [36]. The mutation R289K-induced conformation in H7N9 showed the potential adaptions of the virus itself for future drug-resistance [128]. In addition to the mutations on HA, PB2 and PB1 proteins, the substitutions at site 223 and 275 of NA protein [58] [59], site 97 of PA protein [63] and site 92 of NS1 protein [65] have also made an effect with enhanced virulence in mammalian hosts [61].

Deep neural networks (DNNs) have been widely applied to a variety of fields for learning patterns from massive data, including image classification [188], machine translation [189], speech recognition [190], protein secondary structure prediction [120], etc. Despite the significant progress on the performance of computational models, DNN methods still have limitations. The high predictive results have heavily relied on large amounts of training data. The purely deep learning neural networks have created a black box that leads to uninterpreted and sometimes counter-intuitive results [191] [192]. On the other hand, the cognitive process of humans indicates that people not only learn from concrete samples but also different types of experience and knowledge [193]. However, it is easy to ignore the importance of general information and domain knowledge to construct computational models and regulate the learning process.

In this chapter, a novel framework is presented that enables neural networks to learn simultaneously from label genome sequences as well as prior knowledge (Chapter 3 and 4) to predict the virulence of influenza A viruses. This

is through an iterative rule knowledge distillation process [194] that converts the encoded discrete prior knowledge into the network parameters. The prior knowledge includes mutation information and the reassortment probability of the input genomes. At each iteration, posterior regularization [195] is leveraged to convert the prior genomic information by converting the posterior distribution into the constraint feature sets. Compared with the baseline methods, it achieves better or comparable performance on the collected influenza dataset. As far as we know, it is the first time that attempts to utilize prior knowledge, i.e., mutation and reassortment information for virulence prediction of influenza A viruses. The encouraging predictive results indicate that this model can be used to estimate the lethality of viruses and facilitate flu surveillance of the novel emerging influenza strains.

5.2 Materials and methods

5.2.1 Definition of virulence

The definition of virulence is generally regarded as the ability of a pathogen or microbe to infect or damage a host [9]. More specifically in animal systems, it refers to the degree of damage caused by a microbe to its host [196]. Only the virulence of influenza A virus is explored in this work. As far as we know, virulence has not been clearly defined with a rigorous mathematical definition. Here the mouse lethal dose (MLD) 50 is leveraged to measure the virulence of infections. The level of virulence is categorized into avirulent and virulent types that are labeled as “0” and “1”, respectively. If the MLD50 is greater than 10E6.0, it is regarded as avirulent. Otherwise, it is virulent. Besides, if the virulence label cannot be obtained from the upper or lower bound of MLD50, the RULE 1 to 6

from [197] is utilized to classify the remaining samples.

5.2.2 Data collection

The datasets contain viral sequences and their virulence information. The virulence information was based on Ivan et al. dataset [197], which was collected through previous publications and experiments. The MLD 50 information was recorded in each infection with specific influenza A virus strain and mouse strain. The corresponding sequence data found in the literature information was downloaded from NCBI [198] and GISAID [155]. The sequence data consists of all segments of the strains. For those of incomplete influenza genomes, basic local alignment search tool (BLAST) [199] [200] was performed to search the most similar strains to supplement the genomic data. The collection ended up with 488 unique records of influenza A virus information with corresponding complete genome strains. To further validate the robustness of the proposed model, the virulence is predicted based on the individual subtypes. These influenza sequences were divided into H1N1, H3N2, H5N1 and other subtypes, with 113, 54, 168 and 138 samples, respectively.

5.2.3 Feature transformation

The collected data of each influenza A virus protein type have been aligned with Multiple Alignment Using Fast Fourier Transform (MAFFT) [201]. The protein sequence from the same genome was concatenated using BioEdit [202] with virulence label. Each concatenated alignment of the genomic sequence was 4871 in length. Feature transformation in Chapter 4.2.3 was leveraged to efficiently encode these genome sequences. Seven physicochemical properties

(net charge, hydrophobicity, polarizability, normalized van der Waals, volume polarity, solvent accessibility and secondary structure) of amino acids were used to generate numerical vectors with AAindex that can be processed by machine learning algorithms. The transformation was performed by the CTD method and the details can be found in Chapter 4.2.3. Finally, the size of 147-dimensional vector was obtained for each genome sample.

5.2.4 Model construction

To tackle the issue of accurately predicting influenza virulence with interpretation, the posterior regularization technique [195] is utilized for the virulence prediction. The flowchart of the proposed framework for virulence prediction of influenza A viruses is presented in Figure 5.1. Given the input data $\boldsymbol{X}^n$=[$\boldsymbol{x}^1$, $\boldsymbol{x}^2$, ... ,$\boldsymbol{x}^n$] that each sample is the concatenated genome of all 8 segments, the true label vector is denoted as $\boldsymbol{Y}^n$=[$\boldsymbol{y}^1$, $\boldsymbol{y}^2$, ... ,$\boldsymbol{y}^n$]. The prediction model can be leveraged to acquire the predictive probability vector $\hat{\boldsymbol{Y}}^n$ = P($\boldsymbol{Y}^n|\boldsymbol{X}^n$; $\boldsymbol{\theta}$). The primary contribution of this model is to incorporate prior viral information into the existing model by utilizing the desired distribution q($\boldsymbol{y}^n$) with posterior regularization.

Posterior regularization [195] contains indirect supervision, i.e., prior mutation and reassortment information, via structural constrain on posterior distributions of latent variables [195]. The goal of the posterior regularization framework is to restrict the space of the model posteriors using prior information to guide the model towards desired parameter distributions. It has been widely applied to the prediction tasks, such as sentimental analysis [203] and risk prediction on electronic health records [204]. In this case, posterior information is denoted with sets $\boldsymbol{Q}$ of allowed distributions over hidden variables $\boldsymbol{Y}$. Assume $\boldsymbol{q}(\boldsymbol{y}^n)$ as the

desired distribution of sample n, the posterior regularized loss function is defined as

$$F(\theta, q) = L(\theta) + \alpha \frac{1}{|N|} \sum_{n=1}^{|N|} \min_{q \in Q} KL(q(y^n) || P(y^n | x^n; \theta)) \tag{5.1}$$

where α is a hyperparamter that makes a balance between the loss of predictive model and posterior regularization. KL$(\cdot||\cdot)$ represents the Kullback-Leibler divergence [205] to measure the difference between the desired distribution and the posterior distribution of the prediction model. N is the set of all input sequences and let $|N|$ denote the number of input sequences, where $n \in |N|$. Q represents the set of constraints for posterior information which is defined as:

$$Q = \{q(y^n) : \mathbb{E}_q[\phi(x^n, y^n)] \leq b\} \tag{5.2}$$

where $\phi(x^n, y^n)$ is the constraint features on sample n and b denotes the bounds on the desired expected values of constraint features. To specify the bound b

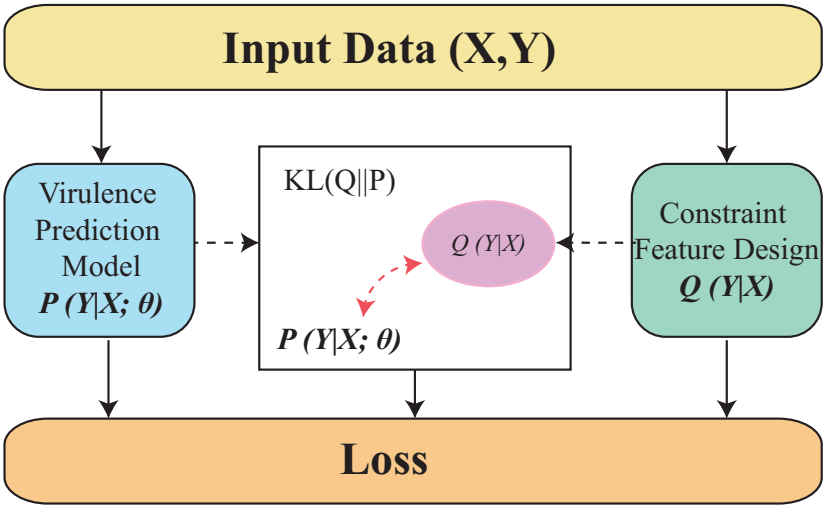


Figure 5.1: The flowchart of the proposed model for virulence prediction of influenza A viruses incorporating prior knowledge.

for different constraint features, a log-linear model is leveraged to represent the desired distribution $\boldsymbol{Q}$ that the objective function can be formulated as follows:

$$\boldsymbol{J(\theta,\tau,w) = L(\theta) + \alpha\frac{1}{|N|}\sum_{n=1}^{|N|} KL(\tilde{y}^n||P(y^n|x^n;\theta)) + \beta L^{'}(\tau,w)} \tag{5.3}$$

where the desired distribution $\tilde{\boldsymbol{y}}^n = \boldsymbol{Q}(y^n|x^n;\tau,w)$, which encodes prior mutation and reassortment information. The mathematical equation of $\tilde{\boldsymbol{y}}^n$ is defined below:

$$\boldsymbol{Q(y^n|x^n;\tau,w) = \frac{exp\{\tau\cdot\phi(x^n,y^n;w)\}}{\sum_{(y^n)'} exp\{\tau\cdot\phi(x^n,(y^n)';w)\}}} \tag{5.4}$$

where $\boldsymbol{\tau}$ is the updated confidence matrix for different categories of constraint features. The parameter $\boldsymbol{w}$ is added to successfully distinguish the difference among multiple pieces of the same categorical prior information. $\boldsymbol{\beta}$ is the hyper-parameter and $\boldsymbol{L^{'}(\tau,w)}$ stands for the average cross entropy between the ground truth $\boldsymbol{y^n}$ and desired distribution $\tilde{\boldsymbol{y}}^n$. The equation of $\boldsymbol{L^{'}(\tau,w)}$ can be found:

$$\boldsymbol{L^{'}(\tau,w) = -\frac{1}{|N|}\sum_{n=1}^{|N|}(y^n log(\tilde{y}^n) + (1-y^n)log(1-\tilde{y}^n))} \tag{5.5}$$

From these equations, it is observed that prior knowledge is incorporated into the model for the virulence prediction. Moreover, it is easy to optimize the objective function from Eq.(5.3) using different optimizer. In the training process, the goal of the proposed model is to minimizing the objective function by learning a set of parameters that can be represented by

$$\boldsymbol{\hat{\theta},\hat{\tau},\hat{w} = \underset{\theta,\tau,w}{\mathrm{argmin}}\{\Omega(\theta,\tau,w)\}} \tag{5.6}$$

Provided with the updated parameters, the virulence label can be predicted for a new concatenated input sequence $\boldsymbol{x^n}$ according to

$$\boldsymbol{\hat{y}^n = \mathrm{argmax}\{P(y^n|x^n;\hat{\theta})\}} \tag{5.7}$$

Although the equation above enable us to make predictions for any given sequences, it ignores the effect of prior mutation and reassortment information of the input sequence and could weaken the predictive ability of the model. Here, following formulation is leveraged to predict the virulence of any given influenza strains by incorporating prior knowledge into the model:

$$\hat{\boldsymbol{y}}^n = \mathbf{argmax}\{\boldsymbol{P}(\boldsymbol{y}^n|\boldsymbol{x}^n;\hat{\boldsymbol{\theta}}) + \boldsymbol{Q}(\boldsymbol{y}^n|\boldsymbol{x}^n;\hat{\boldsymbol{\tau}},\hat{\boldsymbol{w}})\} \tag{5.8}$$

where $\boldsymbol{P}(\boldsymbol{y}^n|\boldsymbol{x}^n;\hat{\boldsymbol{\theta}})$ is the virulence prediction and $\boldsymbol{Q}(\boldsymbol{y}^n|\boldsymbol{x}^n;\hat{\boldsymbol{\tau}},\hat{\boldsymbol{w}})$ denotes prior knowledge prediction.

5.2.5 Constraint feature design

The proposed model allows learning from both raw data and general rules. Since previous work has demonstrated several elements that influence the virulence of influenza viruses. Here these biological factors are categorized into two types: mutations and reassortment information. The prior mutation knowledge is on sequence level while the prior reassortment knowledge provides genome information of influenza viruses. In the following, the constraint feature design is formally provided based on these two types of prior information for each class.

5.2.5.1 Mutation information

It is natural to consider mutation information into the study of influenza virulence. Chapter 2 has described numerous studies on the influence of influenza virulence due to the mutations. These mutations may occur at different sites from the exiting proteins. For example, several mutations on HA segment, i.e., S138A, D222G and Q226L are considered to be associated with virulence [52] [53] [56]. Therefore, it is beneficial to take mutation information into consideration as

constraint features in the predictive model. Table 5.1 summarizes some important mutations on distinct proteins for constructing the constraint features. These mutations are either from previous studies (Chapter 2) or the identified virulent sites (Chapter 3). According to the results, although almost half of the mutations occur on HA protein, it shows an obvious diversity that many other mutations are found at other different proteins.

Given the constraint mutation information $\boldsymbol{g^n}$ = [$(\boldsymbol{g_i^a})^n$, $(\boldsymbol{g_i^b})^n$] of sequence sample $\boldsymbol{n}$ at position $\boldsymbol{i}$, and the corresponding label is $\boldsymbol{y^n}$. The feature on amino acids in certain sites can be defined as follows:

$$\phi_i(\boldsymbol{x_i^n}, \boldsymbol{y_i^n}) = \begin{cases} \mathbf{1} & \boldsymbol{if}\ (\boldsymbol{g_i^a})^n = (\boldsymbol{g_i^b})^n \\ \mathbf{0} & \boldsymbol{otherwise} \end{cases} \tag{5.9}$$

where $(\boldsymbol{g_i^a})^n$ represents the residue information at site $\boldsymbol{i}$ of sample $\boldsymbol{n}$ and $(\boldsymbol{g_i^b})^n$ is the mutation residue at the same site. The value of $\boldsymbol{\phi_i}$ is either 1 or 0, which determines the constraint feature at a certain site, illustrating the mutation information on the input data. For most of the sequences, there is usually more than one mutation in the genomic sequence. The more mutations occurred, the more virulent it will be as the virus may escape from the immune system and cause severe symptoms on humans. To take full advantage of all the underlying

Table 5.1: The summary of virulence-associated mutations

Protein	Virulence-associated mutation
HA	S138A, K163E, D185S, G187D, E190D, D222G, G225D, Q226L, G228S
NA	T223I, H275Y
PB2	Q591K, E627K, D701N
PA	F35L, T97I, S224P, N383P
NS1	D92E

mutation information located at different influenza proteins, similar procedures are applied to the other important sites in Table 5.1 to extract the constraint features from mutations.

5.2.5.2 Reassortment probability

Reassortment is another key mechanism in the viral evolution, resulting in the emergence of numerous novel stains. Different combination of influenza segments from distinct parent strains has indicated numerous descendant strains. The reassortant strains have been responsible for several pandemics since the 19th century and infected numerous people. Thousands of people died because of pandemics. The underlying influence of reassortment is another critical factor for the increased virulence of influenza viruses. If influenza strains have been detected to be reassortant, then the probability of being virulent is higher than those that are not reassortant. To obtain the potential reassortment information of influenza viruses, a novel model named HopPER is proposed in Chapter 4 that aims at estimating influenza reassortment probabilities through host tropism prediction. The details of HopPER are elaborated previously and the function of reassortment estimation for genome $\boldsymbol{x}$ is $\boldsymbol{REprob(x)}$ (Eq.(4.7)). Based on $\boldsymbol{REprob(x)}$, the constraint feature of reassortment is defined as follows:

$$\phi_r(\boldsymbol{x}^n, \boldsymbol{y}^n) = \begin{cases} \mathbf{1} & \boldsymbol{if\ \ REprob(x^n) \geq 0.5} \\ \mathbf{0} & \boldsymbol{if\ \ REprob(x^n) < 0.5} \end{cases} \tag{5.10}$$

5.3 Experimental setup

To fairly evaluate the effectiveness, the proposed model is tested on the collected dataset with the MLD-based definition of virulence. The experimental

results indicate that it obtains better predictive performance by incorporating prior mutation and reassortment information. Furthermore, this model can also learn the importance of distinct prior determinants of virulence for the final prediction. Next, the experimental settings are described followed by the performance comparison between the proposed model and the baseline approaches.

5.3.1 Baseline approach

Three types of baselines are set up for prediction to measure the ability of the proposed model. The first baseline directly applies rule-based methods from [197] to detect the virulence of influenza A viruses. As three different learning approaches are leveraged from this work. Only the best results are shown as the benchmark for comparison. The second baseline is to compare the proposed model with three traditional classification approaches, including neural network (NN), support vector machine (SVM) and logistic regression (LR). The third baseline is using deep learning methods that contain three variants of CNN architecture VGG-19 (Visual Geometry Group) [206], ResNet-50 [207] and ResNext-50 [208]. Figure 5.2 shows the structures of three deep learning-based approaches. The proposed framework is based on three deep learning models by incorporating prior knowledge with the posterior regularization technique. They are denoted as VGG-19*, ResNet-50* and ResNext-50*, which use VGG-19, ResNet-50 and ResNext-50 as the base models. The settings of proposed models are the same as deep learning-based baselines. For all compared methods, feature transformation (Section 5.2.3) is performed on the concatenated input sequence in the training process.

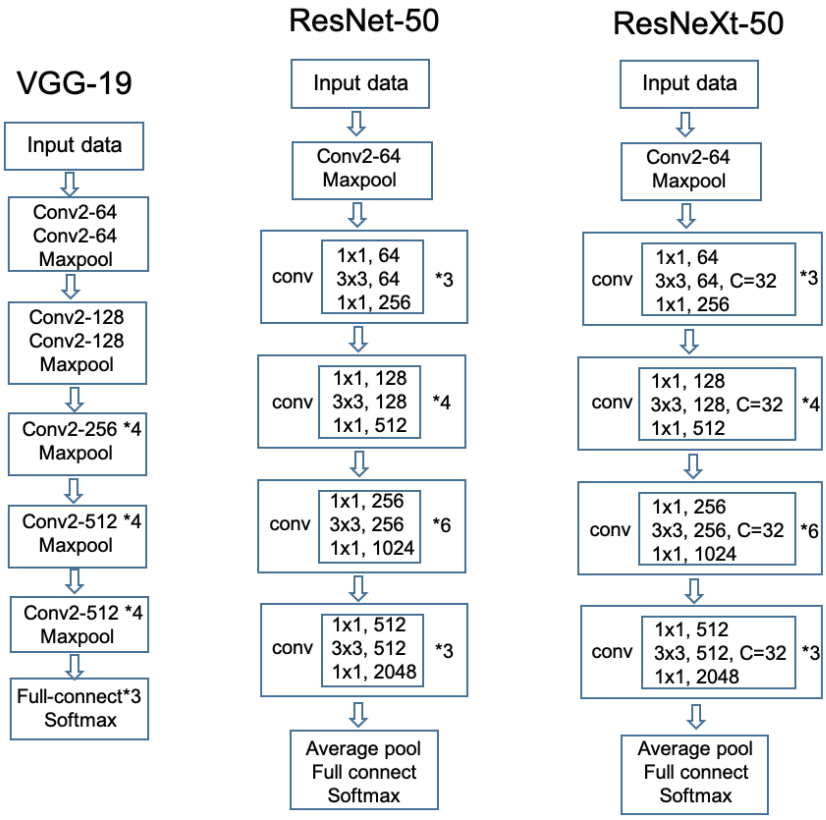


Figure 5.2: The structures of three different deep learning-based baselines for the virulence prediction.

5.3.2 Implementation and evaluation

All the approaches are implemented with Scikit-learn [209] and PyTorch [210]. The processed data is randomly divided into training and testing samples in a ratio of 0.8:0.2. The setting from [197] is utilized when implementing rule-based methods for virulence prediction. For traditional machine learning methods, the parameters for these models are set by default. For all deep learning-based models, stochastic gradient descent is applied with a minimum batch size of 4 for optimization. The learning rate is 0.001 and the size of the filter window is 10 with 130 filter maps. The L1 regularization and drop-out (rate = 0.5)

strategy are carried out for all deep learning approaches with 100 training epochs. The evaluation metrics comprise accuracy, precision, recall and F-score that are utilized in all comparing the methods in for virulence prediction.

5.4 Results and discussion

5.4.1 Comparative performance between the proposed model and other machine learning approaches

Table 5.2 shows the results of all the methods on the processed influenza dataset for virulence prediction. According to the table, the proposed model outperforms all the baselines by all the measures. In more detail, the baseline of rule-based method JRip achieves better performance than traditional classifiers LR, SVM and NN, whereas it is worse than deep learning-based models ResNet-50 and ResNext-50. This indicates that it is more effective for deep learning techniques to deal with the high dimensional feature space. As for the deep learning-based baselines, ResNet-50 performs better than VGG-16 and ResNext-50. Since CNN is constructing more complicated architecture, training such models needs abundant sequence samples. Even though, the performance of CNN is not as well as traditional image classification tasks. This is because a relatively small dataset is applied for virulence prediction, which significantly affects the predictive outcome. Nevertheless, CNN based models still outperform other baselines of rule-based and traditional machine learning approaches.

For the proposed models, ResNet-50* achieves the best performance than the other two methods in accuracy, precision and F-score on the testing set, which are 0.721, 0.711 and 0.816 respectively. The proposed model VGG-19* obtains the highest recall that the true positive rate is 1. Meanwhile, it is shown that

Table 5.2: Comparative performance of virulence prediction on influenza dataset.

Model		Training set				Testing set			
		Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
Baseline	JRip	0.722	0.725	0.733	0.729	0.684	0.682	0.695	0.688
Traditional classifier	LR	0.561	0.508	0.428	0.426	0.549	0.500	0.463	0.445
	SVM	0.627	0.369	0.487	0.379	0.617	0.323	0.490	0.388
	NN	0.554	0.330	0.521	0.413	0.554	0.371	0.506	0.380
Deep learning model	VGG-19	0.666	0.663	0.982	0.791	0.653	0.662	0.969	0.787
	ResNet-50	0.745	0.782	0.833	0.806	0.687	0.685	1.000	0.813
	ResNext-50	0.733	0.803	0.781	0.792	0.667	0.685	0.895	0.776
Deep learning model with prior knowledge	VGG-19*	0.692	0.667	0.986	0.796	0.690	0.655	1.000	0.792
	ResNet-50*	0.763	0.740	0.909	0.816	0.751	0.711	0.958	0.816
	ResNext-50*	0.757	0.773	0.886	0.826	0.720	0.694	0.905	0.785

the performance of all proposed models is better than the corresponding deep learning-based baselines. These observations provide strong evidence that the performance has been improved by incorporating prior mutation and reassortment information into the model. Since the number of training samples of influenza strains whose virulence is defined by MLD50 is relatively small, the rule-based approach achieves comparative performance in comparison with deep learning-based methods. Even though, on the simple data, it is concluded that integrating prior biological information into the existing model for virulence prediction of influenza viruses enhances the predictive performance.

5.4.2 Constraint feature analysis

One of the advantages of the proposed framework is that it can automatically learn the weights for constraint features. Figure 5.3 shows the weights learned by the proposed model ResNet-50* on virulence prediction of influenza viruses

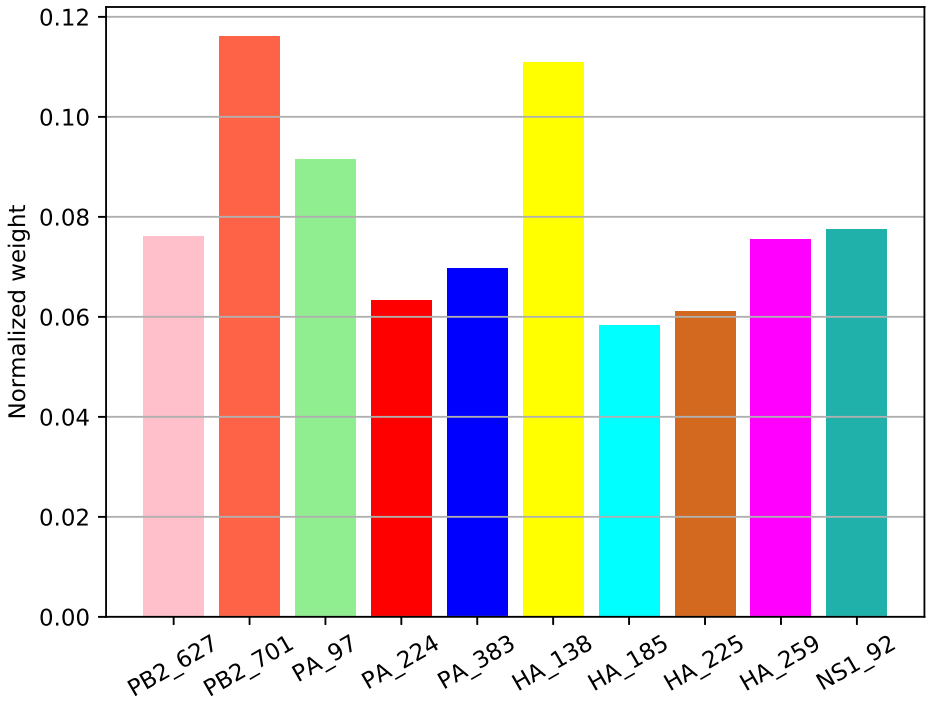


Figure 5.3: Learned weights by ResNet-50* for constraint features on the prediction of influenza virulence. The X-axis represents constraint mutation or reassortment information. As the values of the learned weights could be negative, the weight vector of constraint features is normalized with softmax function. Y-axis represents the weights after normalization. Only the top ten constraint features are shown with their weights.

for the constraint features. The softmax function is leveraged to normalize the learned weight vectors as they could be negative values. Only the top ten constraint features that are assigned larger values are presented in this figure. It is shown that all the constraint feature are mutation-based constraint features. For the virulence prediction, the mutations on site 701 of PB2 protein and site 138 of HA protein play more important roles than other selected sites. Interestingly, the constraint reassortment feature is not ranked within the top 10 constraint features, which suggests that it may not make a significant impact on the virulence in this dataset. One of the possible reasons is that plentiful constraint features generated from mutation information overwhelm transformed reassortment constraint features. However, it does not mean that the reassortment or other mutation sites that are out of the top ten are not the factors for the prediction of influenza virulence. Figure 5.3 shows the learned weights for virulence prediction of the proposed model, which proves its ability for different constraint features according to the characteristics of the input sequence.

5.4.3 Model evaluation on individual subtypes

To further validate the utility and robustness of the proposed framework, experiments on individual influenza subtypes are performed for virulence prediction. Table 5.3 shows the predictive performance on each type. Compared with the values in Table 5.2, it presents worse results on all measures. The reason is that the number of samples used for training of each subtype is smaller than those in Table 5.2, which is directly related to the predictive performance. Nevertheless, the proposed framework shows better performance over the baseline of deep learning models on the values of most measures. The difference is that ResNet-50* only achieves the highest accuracy in the H3N2 subtype, while VGG-16*

Table 5.3: Performance of virulence prediction on individual subtypes of influenza A viruses.

Subtype	Model		Accuracy	Precision	Recall	F-score
H1N1	Deep learning models	VGG-19	0.500	0.687	0.460	0.551
		ResNet-50	0.647	0.664	0.950	0.782
		ResNext-50	0.640	0.667	0.920	0.773
	Deep learning models with prior knowledge	VGG-19*	0.620	0.652	0.920	0.763
		ResNet-50*	0.667	0.667	1.000	0.800
		ResNext-50*	0.693	0.671	1.000	0.803
H3N2	Deep learning models	VGG-19	0.571	0.575	0.958	0.719
		ResNet-50	0.571	0.588	0.833	0.690
		ResNext-50	0.619	0.600	1.000	0.750
	Deep learning models with prior knowledge	VGG-19*	0.548	0.564	0.917	0.698
		ResNet-50*	0.705	0.720	0.750	0.750
		ResNext-50*	0.667	0.708	0.708	0.708
H5N1	Deep learning models	VGG-19	0.654	0.672	0.898	0.769
		ResNet-50	0.556	0.627	0.755	0.685
		ResNext-50	0.627	0.636	0.980	0.771
	Deep learning models with prior knowledge	VGG-19*	0.713	0.698	0.918	0.793
		ResNet-50*	0.647	0.724	0.724	0.724
		ResNext-50*	0.667	0.667	0.959	0.787
Others	Deep learning models	VGG-19	0.664	0.675	0.912	0.776
		ResNet-50	0.587	0.638	0.813	0.715
		ResNext-50	0.629	0.638	0.967	0.769
	Deep learning models with prior knowledge	VGG-19*	0.695	0.677	0.967	0.796
		ResNet-50*	0.664	0.690	0.857	0.765
		ResNext-50*	0.671	0.669	0.956	0.787

proves superior results in H5N1 and the other subtypes. This is probably VGG-16* is more sensitive to the number of training samples, whereas ResNet-50* performs better on large scale training dataset. Nevertheless, it is indicated that the proposed framework outperforms other baseline approaches for virulence prediction on the dataset of individual subtypes. This has further demonstrated the utility and effectiveness of incorporating prior knowledge into existing prediction models for the improvement of predictive performance.

5.5 Chapter summary

In this chapter, a general framework is proposed on virulence prediction of influenza A viruses using all 8 segments. By integrating prior mutation and reassortment knowledge into all the existing models, the proposed framework can improve the performance of virulence prediction. Specifically, three CNN architectures are employed as basic predictive models. To add discrete and heterogeneous prior influenza information, the posterior regularization technique is applied to the proposed model. Experimental results indicate that it outperforms existing machine learning methods on all measures. Moreover, it can provide the learned weights for different prior constraint features. The analysis of the learned weights indicates the importance of mutation on different sites and reassortment on influenza virulence, which can provide novel insights into the formation of virulence and the support for the inference of the lethality of influenza viruses.

Chapter 6

IAV-CNN: a 2D convolutional neural network model to predict antigenic variants of influenza A virus

In previous chapters, the framework of virulence prediction is constructed using prior mutation and reassortment information, which could be utilized to analyze the lethality of viruses. Timely determination of antigenicity is another critical element associated with viral lethality. Empirical experimental methods like hemagglutination inhibition (HI) assays are time-consuming and labor-intensive, requiring live viruses. Recently, many computational models have been developed to predict the antigenic variants without considerations of explicitly modeling the interdependencies between the channels of feature maps. Moreover, the influenza sequences consisting of similar distribution of residues will have high degrees of similarity and will affect the prediction outcome. Consequently, it is challenging but vital to determine the importance of different residue sites and enhance the predictive performance of influenza antigenicity. A 2D convolutional neural network (CNN) model is proposed to infer influenza antigenic variants (IAV-CNN). Specifically, a new distributed representation of amino acids, named ProtVec is introduced that can be applied to a variety of downstream proteomic machine learning tasks. After splittings and embeddings of influenza strains, a 2D squeeze-and-excitation CNN architecture is constructed that enables networks to focus on informative residue features by fusing both spatial and channel-wise information with local receptive fields at each layer. Experimental results on three

influenza datasets show IAV-CNN achieves state-of-the-art performance combing the new distributed representation with the proposed architecture. It outperforms both traditional machine algorithms with the same feature representations and the majority of existing models in the independent test data, which can be served as a reliable and robust tool for the prediction of antigenic variants.

6.1 Introduction

Seasonal influenza seriously threats public health and the global economy [211]. Influenza A/H1N1, A/H3N2, A/H5N1 are the principal subtypes circulating in humans [212]. Vaccination is the most effective way to prevent infection and severe outcomes caused by influenza viruses [29]. The components of vaccines have to be updated regularly to ensure its efficacy [213]. The influenza virus surface glycoproteins hemagglutinin (HA) is the main target for host immunity [214]. However, as stated in Chapter 3, the accumulation of mutations on HA proteins results in the emergence of novel antigenic variants that can not be effectively inhibited by antibodies, posing great challenges for vaccine design. Developing rapid and robust methods to determine influenza antigenicity is critical to influenza vaccine design and lethality analysis.

Hemagglutinin inhibition (HI) assay is the primary method to evaluate the antigenicity of influenza viruses by measuring the ability of antisera to block the HA of the antigen from agglutinating red blood cells [215]. Smith *et al.* created an antigenic map using HI assay data and determined the antigenic evolution of influenza A H3N2 virus from 1968 to 2003 [32]. Liu *et al.* developed PREDAC-H1 that systematically depicted the antigenic patterns and evolution of human influenza A H1N1 viruses [99]. By utilizing 1572 HA sequences and 197 pairs of HA sequences with HI assays data, Huang *et al.* presented the entropy and

likelihood ratio to model amino acid diversity and antigenic variant score [216]. Ren *et al.* employed random forest regression and support vector regression to identify antigenicity-associated sites on HA of A/H1N1 seasonal influenza virus [113]. Richard Neher *et al.* showed a web-based application to interpret measured antigenic data and predict the properties of viruses [78]. Harvery *et al.* analyzed the sequence and 3-D structure information of HA, together with corresponding HI assay data to identify the high- and low-impact amino acid substitutions that drive the antigenic drift of influenza H1N1 viruses [217].

Numerous studies have been conducted to timely predict the antigenic variants or antigenicity of influenza viruses. Lee and Chen investigated 70 mouse monoclonal antibody binding sites for predicting antigenic variants of influenza A/H3N2 with 83% agreement [116]. Sun *et al.* provided a novel method for quantifying antigenic distance and identifying antigenic variants using sequence alone [114]. Additionally, Yin *et al.* presented a stacking model to predict antigenic variants of the H1N1 influenza virus based on epidemics and pandemics [218]. A universal computational model was integrated to predict the antigenic variants for all HA subtypes of influenza A viruses through conserved antigenic structures [106]. Regarding the prediction models on antigenicity, there are several different works to infer the influenza antigenicity with computational models. Qiu *et al.* built an antigenicity prediction model for influenza A/H3N2 viruses by incorporating the structural context of HA protein [107]. Moreover, Yao *et al.* applied a joint random forest method to human H3N2 seasonal influenza data for predicting antigenicity [105]. Wang *et al.* developed a novel low-rank matrix complete model to infer antigenic distances between antigens and antisera [112]. This model exploited the correlations of the viruses and vaccines in serological tests in predicting influenza antigenicity.

Recently, with the successful application of deep neural networks in a variety

of areas including bioinformatics. CNN is one of the most popular approaches applied to solve bioinformatics problems, for example, classification of efflux proteins from membrane [219], human leukocyte antigen class I-peptide binding prediction [220], prediction of protein secondary structure [221] and prediction of protein-protein interaction [222]. In this chapter, deep learning techniques from the natural language processing (NLP) domain are leveraged to tackle the problem of antigenic variants prediction of influenza A viruses. Specifically, a new distributed representation amino acids, named ProtVec, is introduced that maps a 3-grams (three consecutive amino acids) to a 100-dimensional vector space. Then an approach is proposed that combines the 2D CNN model with squeeze-and-excitation mechanisms, named IAV-CNN, for the task of antigenic variants prediction. The flowchart of IAV-CNN is presented in Figure 6.1. The experimental results demonstrate that it achieves better performance compared with several state-of-the-art approaches on three different influenza datasets.

6.2 Materials and methods

6.2.1 Dataset

In the experiment, the antigenic data and sequence data are adopted including subtypes H1N1, H3N2 and H5N1. The antigenic data obtained by hemagglutination inhibition (HI) assays is collected from reports of international organizations and published papers. In total, 1562, 1249 and 666 distinct pairs of antigenic data are collected for influenza A/H1N1, A/H3N2 and A/H5N1, respectively. Correspondingly, the protein sequences of HA are derived from IVR [198] and GISAID [155]. The sequences are selected by full-length strains with the human host and the duplicate sequences are eliminated from the collection. Finally, it is

ended up with 294, 697 and 260 unique HA sequences for subtypes H1N1, H3N2 and H5N1.

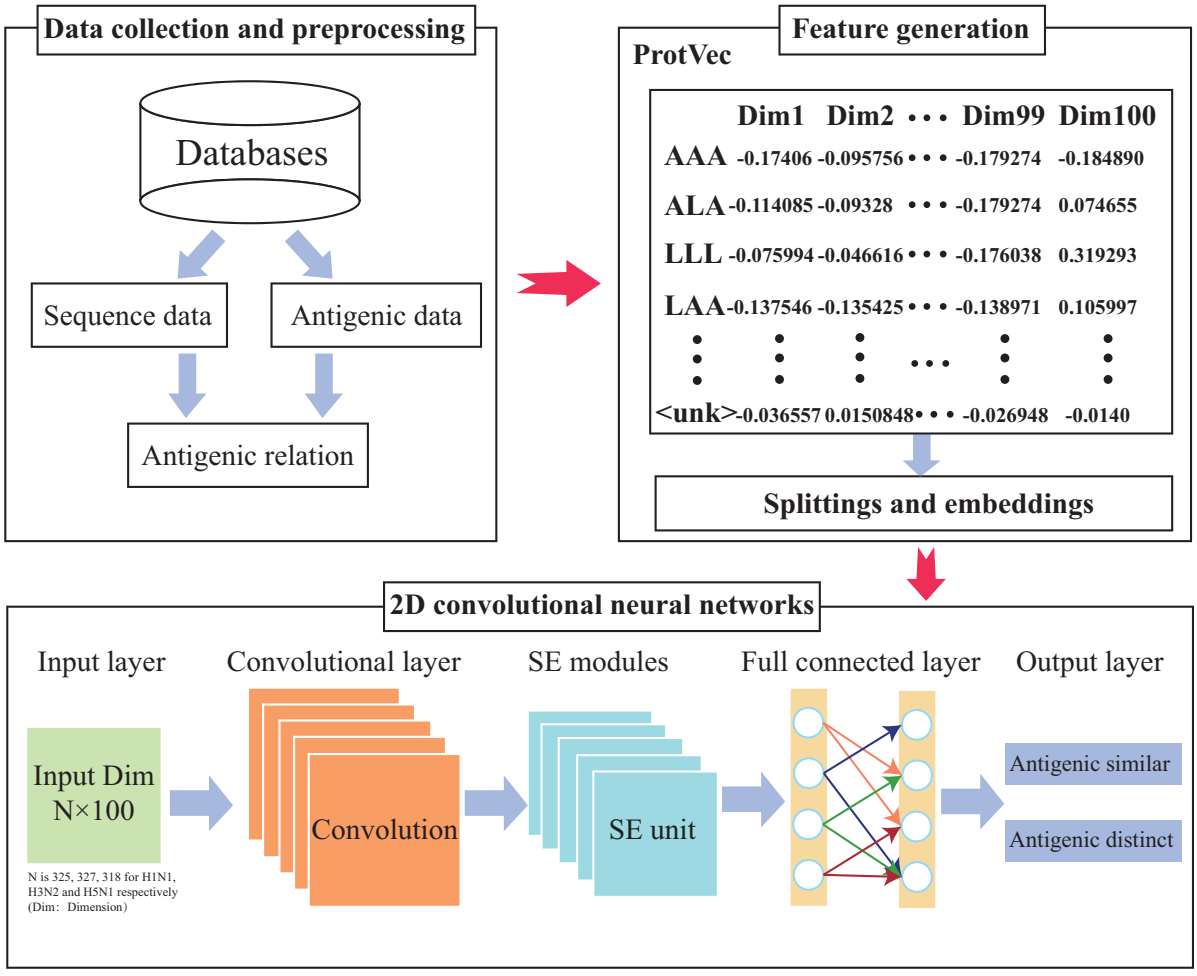


Figure 6.1: The flowchart of IAV-CNN for antigenic variants prediction using two-dimensional convolutional neural networks with squeeze-and-excitation modules.

6.2.2 Preprocessing

The antigenic distance $\boldsymbol{D_{ij}}$ between two strains is defined by Archetti-Horsfall distance [223] as follows:

$$\boldsymbol{D_{ij}} = \sqrt{\frac{\boldsymbol{H_{ii}} * \boldsymbol{H_{jj}}}{\boldsymbol{H_{ij}} * \boldsymbol{H_{ji}}}} \tag{6.1}$$

where the HI titer $\boldsymbol{H_{ij}}$ is the maximum dilution of antisera raised in strain $\boldsymbol{i}$ to inhibit cell agglutination caused by strain $\boldsymbol{j}$. If the antigenic distance $\boldsymbol{D_{ij}}$ is equal or greater than 4, strain $\boldsymbol{i}$ and strain $\boldsymbol{j}$ are antigenic distinct [115]. Otherwise, the pair of strains are regarded as antigenic similar. For the repetitive strain pairs where the HI titer is measured by different independent institutions, the median titer value is utilized to calculate the antigenic distance [102]. As a result, 937, 606, 409 antigenic distinct pairs and 625, 643, 257 antigenic similar pairs of A/H1N1, A/H3N2 and A/H5N1 strains are acquired.

For HA sequence data, only the HA1 proteins are kept for each subtype and the signal peptide is removed from the collected HA1 sequences. As a result, the HA1 lengths of subtype H1N1, H3N2, H5N1 are 327, 329 and 320, respectively. Multiple sequence alignment is performed using the software MAFFT [201] on HA1 proteins for each subtype. Finally, 294 unique sequences are obtained for H1N1, 697 for H3N2 and 260 for H5N1 in the experiment.

6.2.3 Feature generation

The representation of biological sequences is one of the most important problems expressing the biological information with a discrete model or a vector that keeps key pattern characteristics. This is because all the existing machine learning models can only handle vectors but not sequences as elucidated in a

comprehensive review [224]. Distributed representation has displayed significant success in NLP to train word embeddings, the mapping of words to numerical vector space [225], [226]. Recently, it has been explored for bioinformatics applications such as protein classification [227] and structure prediction [228]. To convert the protein sequence information into feature sets that can be managed by neural networks, ProtVec is introduced to encode proteins through distributed representation that each trigram (sequence of three amino acids) protein is embedded in a size of 100-dimension vector [227].

To preserve the sequence pattern information, protein sequences are split into shifted overlapping residues in the window size of 3 (3-grams). The splittings and embeddings are shown in Figure 6.2. Here the subtype H1N1 is used as an example to describe the process that a pair of influenza HA1 proteins are represented by 325 pairs of trigrams. The subtraction of a pair of trigrams characterizes the distinction between two strains at certain positions that can be denoted by a difference vector. The difference vectors $\boldsymbol{V} = [\boldsymbol{v_1}, \boldsymbol{v_2}, ..., \boldsymbol{v_{325}}]$ are derived from ProtVec embeddings. For each vector, i.e., $\boldsymbol{v_1} = \boldsymbol{ProtVec(trigram1)} - \boldsymbol{ProtVec(trigram2)}$, where $\boldsymbol{ProtVec(trigram)}$ is the distributed representation of a trigram in 100-dimension vector space, mapping from ProtVec. Therefore, the antigenic relationship between two HA1 strains is represented in a $\boldsymbol{325*100}$ dimensional vector space. The trigram that contains '-' at any positions will be assigned the 'unknown' embedding from ProtVec.

6.2.4 CNN structure

CNNs have been applied in a variety of fields with impressive results, especially in computer vision when the input is generally a 2D image. Much of the recent fervor has been spurred by accessibility to large training datasets and

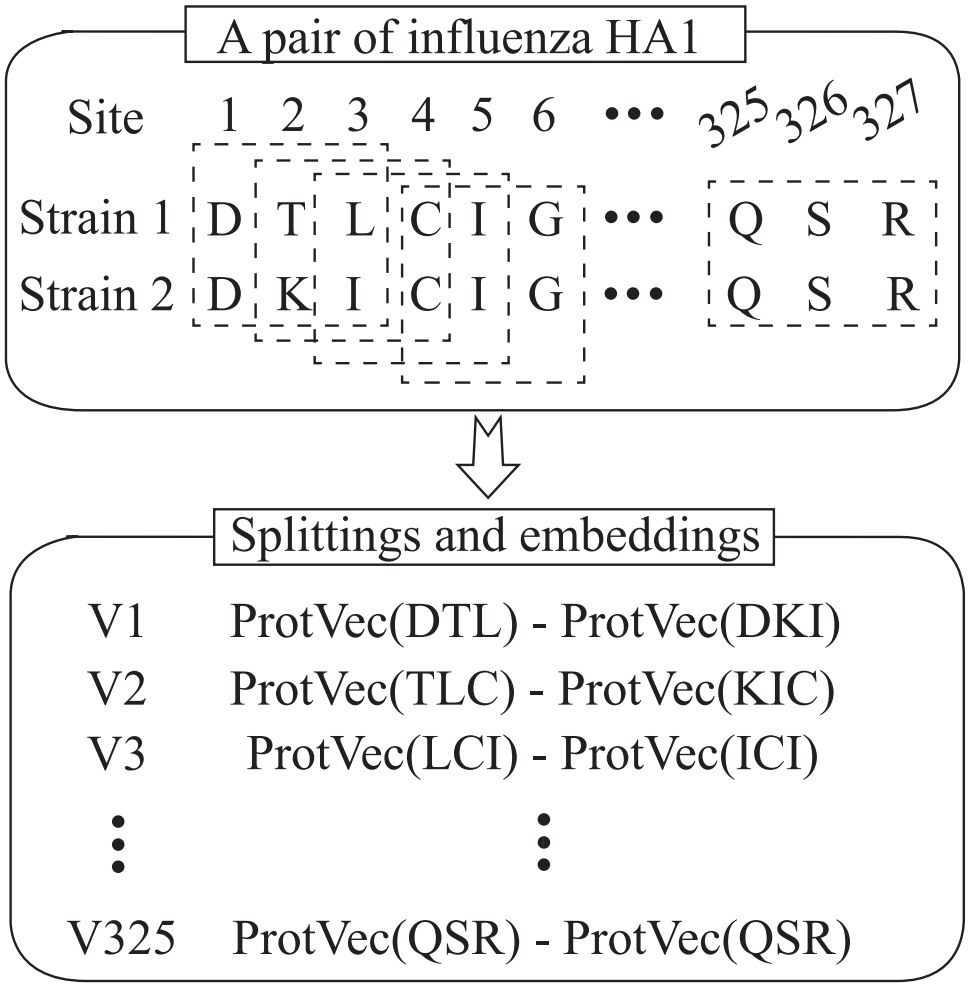


Figure 6.2: The procedure of splittings and embeddings of a pair of influenza H1N1 HA1 proteins. Each pair is embedded in a $\mathbf{325 * 100*}$ dimensional vector space to represent the information of antigenic distance. Strain 1: A/California/07/2009, Strain 2: A/Ohio/9/2015.

advances in cheap computing power to train deep neural networks in an affordable amount of time. Although originally proposed for image classification [188], [206], [229], CNN have been found work well for biological sequence data such as protein classification [219], [230], [231] and prediction of protein function [232], [233]. Encouraged by the successful application of CNN, I take advantage

of the CNN architecture applied to 2D image classification and conveniently generate similar 2D inputs of the ProtVec-based matrix that explores the antigenic relationship between two influenza strains. It is with this insight that the proposed IAV-CNN aims at the task of antigenic variants prediction with CNNs.

Regarding the way of constructing IAV-CNN, the fundamental CNN architecture is first leveraged. To enhance the representational power of the network and boost meaningful sites of strains, while suppressing weak ones, the Squeeze-and-Excitation (SE) block [234] is introduced in the CNN architecture. The SE block squeezes along the spatial domain and reweighs along the channels. The attention and gating mechanisms are activated by modeling the interdependencies between the channels of feature maps. The main idea is to add parameters to each channel of a convolutional block so that the network can adaptively adjust the weighting of each feature map and emphasize useful channels. Hence, it is capable of biasing the allocation of available computational resources towards the most informative residues of strains through SE blocks. The illustration of the SE block is shown in Figure 6.3.

Assume an input $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{H'} \times \boldsymbol{W'} \times \boldsymbol{K'}}$ that passes through a transformation $\boldsymbol{F_{tr}}$, a convolutional operator, to generate output feature map $\boldsymbol{U} \in \mathbb{R}^{\boldsymbol{H} \times \boldsymbol{W} \times \boldsymbol{K}}$. Here $\boldsymbol{H'}$ and $\boldsymbol{W'}$, $\boldsymbol{H}$ and $\boldsymbol{W}$ are the spatial height and width before and after transformation, with $\boldsymbol{K'}$ and $\boldsymbol{K}$ being the input and output channels. The vector $\boldsymbol{V} = [\boldsymbol{v_1}, \boldsymbol{v_2}, ..., \boldsymbol{v_K}]$ represents the learned set of filter kernels, where $\boldsymbol{v_k}$ stands for the parameters of the $\boldsymbol{k}$-th filter. The output is denoted as $\boldsymbol{U} = [\boldsymbol{u_1}, \boldsymbol{u_2}, ..., \boldsymbol{u_k}]$. For each $\boldsymbol{u_k}$, it is formulated by

$$\boldsymbol{u_k} = \sum_{\boldsymbol{n=1}}^{\boldsymbol{K}} \boldsymbol{v_k^n} * \boldsymbol{x^n} \tag{6.2}$$

where $*$ denotes convolution and $\boldsymbol{u_k} \in \mathbb{R}^{\boldsymbol{H} \times \boldsymbol{W}}$. $\boldsymbol{u_k^n}$ is a 2D spatial kernel denoting a single channel of $\boldsymbol{v_k}$ that acts on the corresponding channel of input

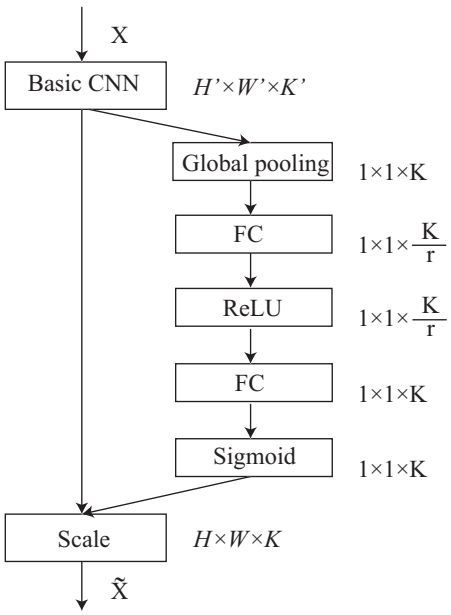


Figure 6.3: The schematic overview of squeeze-and-excitation unit with fundamental CNN module.

$\boldsymbol{X}$. To tackle the issue of exploiting channel dependencies, the global spatial information is squeezed into a channel descriptor by using global average pooling to generate channel-wise statistics. Consequently, a statistic $\boldsymbol{z}$ is obtained by squeezing $\boldsymbol{U}$ through its spatial dimensions $\boldsymbol{H} \times \boldsymbol{W}$. The $\boldsymbol{k}$-th element of $\boldsymbol{z}$ is formulated by

$$\boldsymbol{z_k} = \frac{\mathbf{1}}{\boldsymbol{H} \times \boldsymbol{W}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{H}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{W}} \boldsymbol{u_k}(\boldsymbol{i}, \boldsymbol{j}) \tag{6.3}$$

To capture channel-wise dependencies and make full use of information aggregated in the squeeze operation, a gating mechanism is employed with a sigmoid activation. The equation is described below, where $\boldsymbol{\delta}$ refers to the ReLU function [235], $\boldsymbol{W_1} \in \mathbb{R}^{\frac{\boldsymbol{K}}{\boldsymbol{r}} \times \boldsymbol{K}}$ and $\boldsymbol{W_2} \in \mathbb{R}^{\boldsymbol{K} \times \frac{\boldsymbol{K}}{\boldsymbol{r}}}$. The gate mechanism consists of two full-connected (FC) layers around the non-linearity, i.e., a ReLU and then follow by sigmoid activation, which returns to the channel dimension of

the transformation output $\boldsymbol{U}$. The hyperparameter $\boldsymbol{r}$ is the reduction ratio that allows us to adjust the computational cost and capacity of the SE modules in the network [234].

$$\boldsymbol{s} = \boldsymbol{\sigma}(\boldsymbol{W_2}\boldsymbol{\delta}(\boldsymbol{W_1}\boldsymbol{z})) \tag{6.4}$$

The output of the SE module is finally obtained by rescaling $\boldsymbol{U}$ with the activation $\boldsymbol{s}$

$$\tilde{\boldsymbol{x}}_k = \boldsymbol{F}_{scale}(\boldsymbol{u}_k, \boldsymbol{s}_k) \tag{6.5}$$

Where $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, ..., \tilde{\boldsymbol{x}}_k]$ and $\boldsymbol{F}_{scale}(\boldsymbol{u}_k, \boldsymbol{s}_k)$ is the multiplication between the scalar $\boldsymbol{s}_k$ and the feature map $\boldsymbol{u}_k$. In this regard, the squeeze operator compresses global spatial information into local descriptors and the excitation operator maps these specific descriptors into a set of channel weights. Consequently, SE modules present a global understanding of each channel by squeezing the feature maps to a single numeric value and dynamically change it by adding a content-aware mechanism to weight each channel. Algorithm 2 clarifies the detailed steps of the proposed model for predicting influenza antigenic relationships using 2D CNNs based on ProtVec.

6.2.5 Baseline approaches

Two types of baselines are set to evaluate IAV-CNN. The first baseline is to compare the proposed model with several traditional machine learning algorithms using the same feature space for the prediction tasks. The classifiers include logistic regression (LR), K-nearest neighbor (KNN), support vector machine (SVM), random forest (RF) and neural network (NN). The second baseline is to apply several art-of-the-state approaches. Liao et al. proposed a method by incorporating scoring and regression methods to predict antigenic variants [115]. Yao et al. developed a joint random forest regression algorithm, cooperatively consider top

Algorithm 2 The IAV-CNN algorithm for predicting influenza antigenic variants through ProtVec.

Input: A pair of influenza HA1 sequences $\boldsymbol{a}$ and $\boldsymbol{b}$

Output: Antigenic relationship between $\boldsymbol{a}$ and $\boldsymbol{b}$

Feature generation (Section 2.3)

$\boldsymbol{n} \leftarrow$ The length of HA1 protein

for $\boldsymbol{i}$ = 1 to $\boldsymbol{n}$ **do**

  $\boldsymbol{a_i}, \boldsymbol{b_i} \leftarrow$ Splittings for strains $\boldsymbol{a}$ and $\boldsymbol{b}$

  ProtVec($\boldsymbol{a_i}$), ProtVec($\boldsymbol{b_i}$) $\leftarrow$ Embedding vectors for HA1 subsequences $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$

  $\boldsymbol{v_i}$ = ProtVec($\boldsymbol{a_i}$) - ProtVec($\boldsymbol{b_i}$) $\leftarrow$ The difference vector for two subsequences $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$

end for

$\boldsymbol{V} = [\boldsymbol{v_1}, ..., \boldsymbol{v_n}] \leftarrow$ The representation of two stains $\boldsymbol{a}$ and $\boldsymbol{b}$

$\boldsymbol{X}, \boldsymbol{Y} \leftarrow$ The training samples through feature space $\boldsymbol{V}$

for $\boldsymbol{i}$ = 1 to epoch **do**

  Do initialization

  $\boldsymbol{net} = \boldsymbol{train}$(IAV-CNN, parameters)

  for $\boldsymbol{j}$ = 1 to numbatches **do**

    $\boldsymbol{X_{batch}} = \boldsymbol{X}[\boldsymbol{j} : \boldsymbol{j} + \boldsymbol{batchsize}, :, :, :]$

    $\boldsymbol{Y_{batch}} = \boldsymbol{Y}[\boldsymbol{j} : \boldsymbol{j} + \boldsymbol{batchsize}]$

    $\boldsymbol{scores} = \boldsymbol{net}(\boldsymbol{X_{batch}})$

    $\boldsymbol{loss} = \boldsymbol{CrossEntropyLoss}(\boldsymbol{scores}, \boldsymbol{Y_{batch}})$

    $\boldsymbol{optimizer.step}()$

    $\boldsymbol{Predictions} = \boldsymbol{Output}(\boldsymbol{scores})$

  end for

end for

return $\boldsymbol{Predictions}$

substitution matrices that can improve the prediction performance [105]. Lee and Chen used the number of amino acid changes located on the five epitope regions for the antigenic variants prediction [116]. Lees et al. provided an update for the frequently referenced five antigenic sites and increase additional assignments to establish five canonical regions [117]. They constructed a range of linear models based on banded changes for the prediction. Peng et al. built a universal model for the antigenic variation prediction of influenza A virus H1N1, H3N2 and H5N1 using conserved antigenic structures [106]. The reconstruction of these models is implemented to predict antigenic variants on three influenza datasets in comparison with the proposed algorithm.

6.2.6 Implementation and evaluation

All the approaches are implemented with Scikit-learn [209] and PyTorch [210]. The antigenic distinct is labeled as '1' and antigenic similar is '0' for the relationship of two strains. The influenza datasets of each subtype are randomly divided into independent training and testing set with a ratio of 0.8:0.2. The training process is evaluated on the training dataset and the independent testing dataset is used to assess its capability in predicting relations of novel antigenic variants. For CNN-based models, several algorithms are applied with a minimum batch size of 32 for optimization. The one that achieves the best performance of the experimental results will be selected. The drop-out (rate=0.5) strategy is carried out with the learning rate of 0.001 and all the models are iterated for 100 training epochs. Five different metrics are adopted including accuracy, precision, recall, f-score and Matthews's Correlation Coefficient (MCC) to evaluate the predictive performance of the models.

6.3 Results and discussion

The quality and reproducibility of the model is a crucial factor for the study. In the experiments, the effect of using different optimizers is investigated, including Adaptive Moment Estimation (Adam) [236], Adadelta [237], Adaptive Gradient Algorithm (AdaGrad) [238], Root Mean Square Propagation (RMSProp) [239] and Stochastic Gradient Descent (SGD) [240]. Next, IAV-CNN is described and how it made use of the new distributed representation of amino acids to solve the problem of antigenicity prediction over other traditional classifiers. Finally, the experimental results obtained by IAV-CNN were presented against the recently developed computational prediction methods.

6.3.1 The performance of IAV-CNN with different optimizers

Table 6.1 shows the predictive performance of IAV-CNN with different optimizers on the testing data of three influenza subtypes. The best results for each dataset are highlighted in bold. It is observed from the table that by using SGD optimizer, IAV-CNN achieves the best performance of 0.876, 0.897, 0.851 0.843 and 0.735 in terms of accuracy, precision, recall, F-score and MCC on H3N2 influenza data. Similarly, when applied SGD optimizer in the other two datasets, the proposed model also displays the best performance in all of the metrics except recall. Therefore, the SGD algorithm is used as the optimizer on subsequent experiments in comparison with other approaches for antigenicity prediction. However, varied performance is observed for different datasets, for instance, H1N1 displays an overall more desirable outcome than the other two types, This may largely owe to the inconsistent sample size that the model on H1N1 dataset presents better predictive results compared with H3N2 and H5N1.

Table 6.1: Performance comparison of IAV-CNN model with different optimizers on H1N1, H3N2 and H5N1 datasets.

Dataset	Optimizer	Accuracy	Precision	Recall	F-score	MCC
H1N1	Adam	0.850	0.857	0.914	0.885	0.663
	Adadelta	0.856	0.928	0.824	0.873	0.716
	AdaGrad	0.885	0.896	0.915	0.906	0.759
	RMSProp	0.872	0.871	**0.933**	0.901	0.725
	SGD	**0.917**	**0.928**	0.915	**0.924**	**0.806**
H3N2	Adam	0.796	0.866	0.729	0.792	0.603
	Adadelta	0.806	0.831	0.737	0.787	0.586
	AdaGrad	0.828	0.851	0.759	0.785	0.622
	RMSProp	0.784	0.819	0.724	0.763	0.598
	SGD	**0.876**	**0.897**	**0.851**	**0.843**	**0.735**
H5N1	Adam	0.836	0.868	0.846	0.857	0.665
	Adadelta	0.813	0.863	0.808	0.834	0.623
	AdaGrad	0.851	0.822	0.949	0.881	0.696
	RMSProp	0.836	0.826	**0.910**	0.866	0.661
	SGD	**0.881**	**0.908**	0.885	**0.896**	**0.756**

6.3.2 Comparative performance between IAV-CNN and traditional classifiers on ProtVec-based features

Table 6.2 summarizes the comparative results of IAV-CNN and other traditional machine learning methods including logistic regression, k-nearest neighbor, support vector machine, random forest and neural network. For a fair comparison, the optimal parameters are utilized for the classifiers in all experiments. Specifically, for all subtypes of influenza data, random forest and neural networks

perform higher accuracy than the proposed model on the training data, whereas IAV-CNN has demonstrated better predictive results on the testing data. This is probably the overfitting problem that the classic algorithms fit too well with the training data, but it is difficult for them to well generalize on new samples that are not in the training set. IAV-CNN overcomes this issue by applying the dropout mechanism that randomly sets activation to zero during the training process to avoid overfitting. The results present the accuracy of 0.917, 0.876 and 0.881 for three subtypes. The results are 5.4%, 4.8% and 5.3% higher than the best traditional classifiers, which only achieve 0.895, 0.824 and 0.833, respectively. Besides, it is noticed that a simple SVM algorithm is not suitable for the prediction of small-scale H5N1 data. The SVM finds a maximum edge hyperplane for classification. Since there is no large number of the iterative process, the prediction ability is limited and the accuracy is low.

6.3.3 Comparative performance between IAV-CNN and other methods

To demonstrate the effectiveness, IAV-CNN is further compared with several state-of-the-art methods on the prediction of influenza antigenicity on three datasets. Here, 5-fold cross-validation is adopted in the training data, which has been utilized by many investigators to construct the predictive models. According to the experimental results in Table 6.1 and 6.2, SGD has been chosen as the best optimizer for IAV-CNN. The parameters including the learning rate (0.001) with a dropout (0.5) mechanism remain the same in the experiments for CNN-based models. Furthermore, independent testing data is used to evaluate the ability of the model to predict new sample data with robustness. Figure 6.4 shows the performance comparison of IAV-CNN with other computational methods on

Table 6.2: Comparative performance between IAV-CNN and other machine learning methods using ProtVec features on training and testing data of three influenza subtypes. Acc: Accuracy; Pre: Precision; Rec: Recall; F1: F-score

Subtype	Model	Training data					Testing data				
		Acc	Pre	Rec	F1	MCC	Acc	Pre	Rec	F1	MCC
H1N1	LR	0.817	0.816	0.892	0.853	0.616	0.722	0.752	0.826	0.787	0.392
	KNN	0.901	0.956	0.873	0.913	0.803	0.815	0.915	0.774	0.839	0.637
	SVM	0.594	0.594	1.000	0.745	0.409	0.623	0.623	1.000	0.768	0.423
	RF	0.987	0.993	0.985	0.989	0.974	0.863	0.884	0.897	0.891	0.706
	NN	0.998	0.997	0.999	0.998	0.995	0.859	0.895	0.877	0.886	0.703
	IAV-CNN	0.968	0.976	0.972	0.974	0.937	0.917	0.928	0.915	0.924	0.806
H3N2	LR	0.847	0.872	0.793	0.831	0.694	0.696	0.761	0.624	0.686	0.404
	KNN	0.863	0.893	0.808	0.848	0.727	0.728	0.804	0.647	0.717	0.471
	SVM	0.473	0.473	1.000	0.643	0.403	0.532	0.532	1.000	0.695	0.429
	RF	0.962	0.968	0.951	0.959	0.924	0.776	0.824	0.737	0.778	0.557
	NN	0.973	0.967	0.977	0.972	0.946	0.772	0.817	0.737	0.775	0.548
	IAV-CNN	0.953	0.928	0.911	0.937	0.909	0.876	0.897	0.851	0.843	0.735
H5N1	LR	0.889	0.902	0.921	0.912	0.763	0.813	0.863	0.808	0.834	0.623
	KNN	0.883	0.930	0.879	0.904	0.758	0.799	0.849	0.795	0.821	0.593
	SVM	0.378	0.000	0.000	0.000	0.000	0.418	0.000	0.000	0.000	0.000
	RF	0.976	0.991	0.970	0.980	0.949	0.828	0.867	0.833	0.850	0.651
	NN	0.981	0.997	0.973	0.985	0.961	0.828	0.867	0.833	0.850	0.651
	IAV-CNN	0.955	0.997	0.990	0.994	0.983	0.881	0.908	0.885	0.896	0.756

independent testing data, as detailed below.

The x-axis represents the different methods applied to the prediction and the

y-axis shows the values of all metrics. It is shown that IAV-CNN achieves a remarkable higher performance than compared methods. In more detail, it can obtain an accuracy of 0.920, 0.873 and 0.889 for independent H1N1, H3N2 and H5N1 testing data, respectively. The results are 13.9%, 8% and 6.7% higher

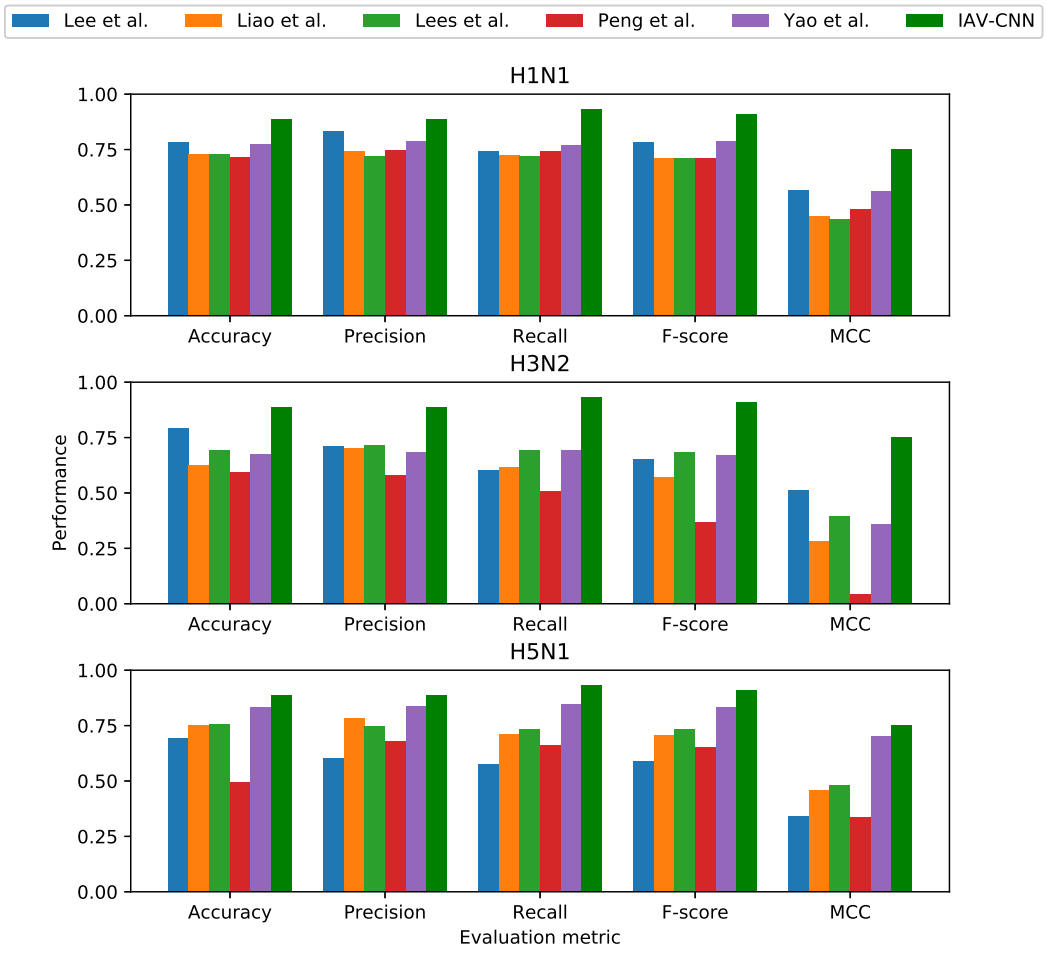


Figure 6.4: The comparative performance between IAV-CNN and other advanced methods for predicting influenza antigenic variants on independent testing data of three influenza subtypes.

than the best performance of compared methods. Regarding other evaluation metrics, the results also indicate that IAV-CNN outperforms the current methods on all datasets. Overall, it is demonstrated that IAV-CNN can accurately predict influenza antigenic variants on selected subtypes with feasibility and robustness. It may also be applicable to predict the antigenicity of a wide range of viruses and drive the development of personalized medicine for infectious diseases.

6.3.4 Interpretation

The prediction of influenza antigenicity is critical for the study of viral evolution and vaccine selection. Although many methods have been proposed to predict novel influenza variants using diverse feature representations, i.g. epitope and physicochemical properties, when establishing the machine learning models, the correlation between features are never taken into consideration. The proposed IAV-CNN is an important type of 2D CNN model consisting of a convolutional kernel, squeeze-and-excitation module and a full-connected layer. It intends to capture meaningful residue sites and even hidden features by scanning the sequences of pair of strains. The introduction of SE modules helps us to focus on the sites with different residues that are given larger weights in the training process. The results prove that IAV-CNN can enhance the predictive performance over other existing machine learning approaches by capturing important residue sites of the compared strains. As a result, the proposed model can be served as a reliable tool for the prediction of influenza antigenicity, which assists biologists to gain a better view to judge the lethality of influenza viruses and facilitate vaccine selection for seasonal epidemics.

6.4 Chapter summary

In this chapter, it has been demonstrated that how machine learning techniques from NLP domain could be leveraged to solve bioinformatic problems, specifically the antigenicity prediction of influenza A viruses. IAV-CNN is proposed to extract a vector space with a distributed representation of amino acids through ProtVec and predict the influenza antigenic variants, using a 2D CNN architecture with squeeze-and-excitation mechanisms. Compared with other traditional machine learning algorithms, IAV-CNN produces superior predictive efficacy with the same feature representations on three different influenza datasets. Moreover, further experiments demonstrate that the proposed model achieves state-of-the-art antigenicity prediction results on the majority of test sets over existing models. This presents an exciting prospect for the estimation of viral lethality where it is possible to detect potential dangerous strains that may infect humans.

Chapter 7

Conclusion and future work

Despite increasingly available biological sequences and progress in bioinformatic tools for the study of influenza viruses in recent years, novel influenza strains are still emerging to cause severe human infections. Current flu surveillance mainly depends on traditional experiments to estimate the degree of lethality of influenza viruses, which is costly and time-lagged. Therefore, there is a pressing need to develop a timely efficient framework with the capability to analyze the lethality of emerging influenza strains and promote flu surveillance.

7.1 Thesis summary

This thesis aims at the lethality analysis of influenza strains through their virulence and antigenicity. It first introduces the background, purpose and objectives in Chapter 1. The literature review is provided to present an overview of the previous research that has been done as well as the challenges left to be solved in Chapter 2. Then, computational methods are systematically explored to address the problem of predicting the virulence and antigenicity of influenza A viruses, respectively. This will show significant help with the increasing efforts in flu surveillance and a large amount of available sequence data. The application of machine learning methods on the existing viral sequences allows us to establish computational models and achieve final prediction goals. This is implemented through four phases that are illustrated in Chapters 3 to 6.

This first phase is the identification of potential critical virulent sites through the hemagglutinin of influenza A virus in Chapter 3. By labeling the influenza strains into two categories (i.e., pandemic and non-pandemic) based on the information of previous pandemics, it is successfully converted into a binary classification problem. Three rule-based methods are applied to classify pandemic strains as well as extract corresponding rules to identify potential virulent sites (Table 3.3). Experimental results indicate high performance with over 90% in accuracy (Table 3.2). Meanwhile, 16 different sites on HA protein are detected as potential virulent sites, where 14 of them are experimentally validated as antigenic sites or make an impact on the virulence (Figure 3.2).

Next, the detection of reassortment of influenza A viruses using host tropism prediction is performed at the second phase (Chapter 4). The host tropism of the influenza genome is first examined to understand the reassortment process by exchanging the segments between different viruses. Eight individual models are constructed to predict host tropism independently based on the physicochemical properties of amino acids (Table 4.2). The predictive results range from 86.5% to 96.5% in accuracy on the testing data for different protein models (Table 4.3). Based on these host tropism prediction models, HopPER is further proposed to effectively estimate the probabilities of influenza reassortment. The experiments on real and synthetic datasets show comparative predictive performance over existing state-of-the-art methods on reassortment detection (Table 4.5). It has also been demonstrated that HopPER is effective in both complete and incomplete genomes (Table 4.7). Moreover, the analysis of evolutionary success through reassortment across different years offers novel insights for the evolution path of influenza viruses (Figure 4.3).

The identification of virulent sites and detection of reassortment for influenza viruses lead to the construction of an integrative framework for the prediction of

influenza virulence in the third phase (Chapter 5). The mutation and reassortment information of influenza viruses are used as prior knowledge and added to the feature space using posterior regularization, which makes the training process more efficient using deep learning models. The proposed model with prior knowledge outperforms other machine learning approaches on the processed dataset for virulence prediction (Table 5.2). Besides, the model is capable of automatically learning the weights for different constraint features (Figure 5.3), which indicates the importance of mutations from different sites and reassortment that have contributed to influenza virulence. The experiments on individual subtypes further prove the utility of the proposed model on the virulence prediction of influenza viruses.

Finally, a 2D CNN model is developed to predict antigenic variants of influenza A virus in the fourth phase (Chapter 6). Specifically, sequence data is used to generate features in the training process and the label is produced by calculating the antigenic distance between two strains. A new distributed amino acid representation named ProtVec is introduced to encode biological sequences. To deal with long-length influenza strains, the raw sequences are split into subsequences of 3-grams. Squeeze-and-excitation module is added into the fundamental CNN architecture, which allows us to focus on informative residue features. Compared with existing machine learning approaches, the proposed model achieves a significantly better performance on the testing data (Table 6.2 and Figure 6.4). This model can be used to analyze the antigenicity of novel emerging strains compared with exiting known viruses. Hence, this model will facilitate the rapid determination of antigenicity and influenza surveillance.

In conclusion, the primary goal of meta-analysis for analyzing the lethality of influenza A viruses concerning virulence and antigenicity using machine learning approaches has been accomplished. The detection of the potential virulent sites

and the estimation of influenza reassortment in Chapters 3 and 4 pave the path for the construction of a novel virulence prediction model in Chapter 5, which takes prior biological information into consideration. A 2D CNN built in Chapter 6 on the prediction of antigenicity of influenza viruses reflects the capacity of viral antigen to bind to or interact with the host immunity system, which further describes the lethality of the potential damage the virus will cause on hosts.

7.2 Strategies for lethality estimation

The development of computational models in Chapters 3 to 6 for virulence and antigenicity prediction aims at profiling the lethality of influenza viruses, as well as improving the current influenza surveillance system. With the increasing biological data and the improvement of sequencing technology, the surveillance of influenza viruses has been already created. Existing risk assessment still depends on genomic markers to detect the lethality of new emerging influenza strains that may infect humans [241]. While other evaluation metrics consist of the detection of receptor binding property, antiviral treatment resistance and lab animal transmission. All of these methods require a large amount of time and are label-intensive. Thus they will not meet the requirement for a timely determination regarding the risk of novel emerging influenza strains, which is critical for the prevention of the outbreak of catastrophic epidemics or pandemics.

Therefore, lethality is leveraged as one of the indicators to complement the existing influenza virologic surveillance. By constructing machine learning tools to predict the virulence and antigenicity of influenza viruses, it can infer the lethality of the virus to timely predict the risk of the influenza strains. As an example, during the flu outbreaks, the viral samples can be collected and obtained using sequencing technology. On the one hand, the virulence of the collected

strains can be rapidly estimated through the models developed in Chapter 5. On the other hand, the methods in Chapter 6 would help us to identify its antigenicity by comparing it with previous existing strains. The results are integrated to determine the lethality and estimate the risks when the viruses spread to humans, which enables the authorities to make a more informed decision swiftly in the event of any potential widespread outbreaks.

7.3 Future directions

This thesis is just the beginning of leveraging machine learning approaches to investigate the lethality of emerging influenza A viruses. It analyzes the degree of lethality through virulence and antigenicity of influenza virus using primarily protein sequences and antigenic data. Several potential directions are proposed for future work.

One important direction is the increase of virulent and antigenic influenza data, which significantly influences the performance of computational models. Besides, the integration of heterogeneous biological data, such as spatiotemporal and medical data can strengthen the diversity of data sources. For example, Incorporating accurate timing information, host information and geographical location of each strain would provide more detailed information about the outbreak of flu, enabling more comprehensive studies on specific outbreaks and identify the lethality of the viruses. Similarly, the inclusion of ecological data is useful to track the path of the virus and possibly predict future transmission paths [242]. Collectively, the increase and integration of data from multidisciplinary sources can provide more information on the virus, allowing for comprehensive analysis for the estimation of the lethality of viruses.

Another possible direction is to accommodate more virulence-associated factors into virulence prediction. In this work, only mutations and reassortment are taken into consideration as prior information for the construction of the virulence prediction model. In reality, it is a complex process for the formation of virulence. A single one or two factors may not provide enough confidence to determine the virulence. Therefore, uncovering more factors that are responsible for the virulence of viral infections and converting them into the computational model can provide a more comprehensive analysis and higher predictive accuracy. Hence, the virulence prediction model developed in Chapter 5 can be further expanded by incorporating more constraint viral features such as the co-occurring mutations [110], host immune system [243] and environmental elements [244] to improve the predictive performance.

Moreover, the objective of this thesis is to analyze the lethality of influenza viruses using machine learning approaches, which is decomposed by the prediction of viral virulence and antigenicity through computational models. However, the direct inference of viral lethality is a highly complex task and poses significant challenges to quantitatively identify lethality based on influenza sequences. Though several models have been proposed for the prediction of virulence and antigenicity, the determination of lethality not only depends on these two measures but is closely related to the information of host infections, such as the immune system and health status. Besides, there is a lack of qualitative models to infer the lethality of influenza viruses. As a result, more efforts are needed to address this issue in future studies.

Finally, the ultimate goal of this research is accurate influenza surveillance that includes the ability to predict future epidemics or pandemics, the detection of viral lethality, etc. This is, in truth, a huge project that may not be achieved in a short period with the current state of the technological and available biological

data. Even though the methodology and framework proposed in this thesis will contribute to the final goal of influenza research and can be applied to the lethality analysis of other RNA viruses such as Ebola virus [245], Henipavirus [246] and Dengue virus [247] and the recent emerging novel coronavirus [248]. Of particular interests, this new coronavirus named COVID-19 has rapidly spread to 216 countries or territories on May 21, 2020, leading to 4,904,413 confirmed cases with 323,412 deaths according to the World Health Organization [249]. It has caused great morbidity and mortality, and unfortunately, the fast and untraceable virus mutations take the lives of people before the immune system can produce the inhibitory antibody [250]. No miracle drug or vaccines are available to treat or prevent the humans infected by coronaviruses [251] [252]. Therefore, there is a desperate need for developing therapeutics to defeat novel coronavirus. I believe that the proposed framework in this thesis can shed light on the inference of the lethality of coronaviruses using machine learning approaches. Through the effective representation of the molecular structure of amino acids and analysis of antigen-antibody interaction, it could even facilitate the discovery of neutralizing antibodies for potential novel coronavirus, and further assist the development of vaccines design and enables other scientists to gain a better understanding of coronavirus. That, in turn, will improve public health and provide effective preparedness for potential future epidemics or pandemics.

List of Publications

Journal Publications

• **Yin R,** Luo Z, Kwoh CK, et al. A weighted ensemble CNN model for the virulence prediction of influenza A viruses using all 8 segments. Submitted to Journal of Bioinformatics (**under review**)

• **Yin R,** Thwin N, Zhuang P, et al. IAV-CNN: a 2D convolutional neural network model to predict antigenic variants of influenza A virus. Submitted to IEEE/ACM Transactions on Computational Biology and Bioinformatics (**under review**)

• **Yin R**, Luusua E, Dabrowski J, et al. Tempel: Time-series Mutation Prediction of Influenza A Viruses via Attention-based Recurrent Neural Networks[J]. Bioinformatics, 2020.

• **Yin R,** Zhou X, Rashid S, et al. HopPER: an adaptive model for probability estimation of influenza reassortment through host prediction[J]. BMC Medical Genomics, 2020, 13(1): 9.

• Zhou X, **Yin R,** Zheng J, et al. An encoding scheme capturing generic priors and properties of amino acids improves protein classification[J]. IEEE Access, 2019, 7: 7348-7356.

• **Yin R,** Tran V H, Zhou X, et al. Predicting antigenic variants of H1N1 influenza virus based on epidemics and pandemics using a stacking model[J]. PloS one, 2018, 13(12): e0207777.

• Zhou X, **Yin R,** Kwoh C K, et al. A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza A viruses[J]. BMC genomics, 2018, 19(10): 936.

• **Yin R,** Zhou X, Zheng J, et al. Computational identification of physicochemical

signatures for host tropism of influenza A virus[J]. Journal of bioinformatics and computational biology, 2018: 1840023-1840023.

• Ding P, **Yin R,** Luo J, et al. Ensemble Prediction of Synergistic Drug Combinations Incorporating Biological, Chemical, Pharmacological and Network Knowledge[J]. IEEE journal of biomedical and health informatics, 2018.

Conference Proceedings

• **Yin R,** Zhang Y, Zhou X, et al. Time series prediction of vaccine selection for influenza A H3N2 with recurrent neural networks. The 30th International Conference on Genome informatics (**accepted**)

• **Yin R,** Tan J, Akhila D, et al. Inference of Sequence Homology by BLAST visualization of Influenza Genome set[C] Proceedings of the 9th International Conference on Computational Systems-Biology and Bioinformatics. ACM, 2018: 5.

• Ivan F X, Zhou X, Deshpande A, **Yin R**, et al. Phylogenetic Tree based Method for Uncovering Co-mutational Site-pairs in Influenza Viruses[C]//Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM, 2017: 21-26.

• **Yin R,** Zhou X, Ivan F X, et al. Identification of Potential Critical Virulent Sites Based on Hemagglutinin of Influenza a Virus in Past Pandemic Strains[C] Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science. ACM, 2017: 30-36.

Book Chapter

• **Yin R,** Chee Keong Kwoh, and Jie Zheng. Whole Genome Sequencing Analysis: Computational Pipelines and Workflows in Bioinformatics, Encyclopedia of Bioinformatics and Computational Biology, 176-183, 2019.

References

[1] A. S. Monto, S. Gravenstein, M. Elliott, M. Colopy, and J. Schweinle, "Clinical signs and symptoms predicting influenza infection," *Archives of internal medicine*, vol. 160, no. 21, pp. 3243–3247, 2000.

[2] M. C. Zambon, "Epidemiology and pathogenesis of influenza," *Journal of Antimicrobial Chemotherapy*, vol. 44, no. suppl 2, pp. 3–9, 1999.

[3] M. F. Ducatez, C. Pelletier, and G. Meyer, "Influenza d virus in cattle, france, 2011–2014," *Emerging infectious diseases*, vol. 21, no. 2, p. 368, 2015.

[4] G. Boivin, I. Hardy, G. Tellier, and J. Maziade, "Predicting influenza infections during epidemics with use of a clinical case definition," *Clinical infectious diseases*, vol. 31, no. 5, pp. 1166–1169, 2000.

[5] W. H. Organization *et al.*, "Fact sheet no 211: Influenza (seasonal)," *WHO: Geneva, Switzerland, April*, 2009.

[6] J. K. Taubenberger and D. M. Morens, "1918 influenza: the mother of all pandemics," *Rev Biomed*, vol. 17, pp. 69–79, 2006.

[7] G. J. Smith, D. Vijaykrishna, J. Bahl, S. J. Lycett, M. Worobey, O. G. Pybus, S. K. Ma, C. L. Cheung, J. Raghwani, S. Bhatt *et al.*, "Origins and evolutionary genomics of the 2009 swine-origin h1n1 influenza a epidemic," *Nature*, vol. 459, no. 7250, pp. 1122–1125, 2009.

[8] C. Ke, C. K. P. Mok, W. Zhu, H. Zhou, J. He, W. Guan, J. Wu, W. Song, D. Wang, J. Liu *et al.*, "Human infection with highly pathogenic avian influenza a (h7n9) virus, china," *Emerging infectious diseases*, vol. 23, no. 8, p. 1332, 2017.

[9] S. SacristAN and F. GARCÍA-ARENAL, "The evolution of virulence and pathogenicity in plant pathogen populations," *Molecular Plant Pathology*, vol. 9, no. 3, pp. 369–384, 2008.

[10] F. Krammer, G. J. Smith, R. A. Fouchier, M. Peiris, K. Kedzierska, P. C. Doherty, P. Palese, M. L. Shaw, J. Treanor, R. G. Webster *et al.*, "Influenza

(primer)," *Nature Reviews: Disease Primers*, 2018.

[11] M. Khanna, P. Kumar, K. Choudhary, B. Kumar, and V. Vijayan, "Emerging influenza virus: a global threat," *Journal of biosciences*, vol. 33, no. 4, pp. 475–482, 2008.

[12] R. J. Russell, P. S. Kerry, D. J. Stevens, D. A. Steinhauer, S. R. Martin, S. J. Gamblin, and J. J. Skehel, "Structure of influenza hemagglutinin in complex with an inhibitor of membrane fusion," *Proceedings of the National Academy of Sciences*, vol. 105, no. 46, pp. 17 736–17 741, 2008.

[13] S. Bertram, I. Glowacka, I. Steffen, A. Kühl, and S. Pöhlmann, "Novel insights into proteolytic cleavage of influenza virus hemagglutinin," *Reviews in medical virology*, vol. 20, no. 5, pp. 298–310, 2010.

[14] E. Brown, "Influenza virus genetics," *Biomedicine & Pharmacotherapy*, vol. 54, no. 4, pp. 196–209, 2000.

[15] E. J. Schrauwen, M. de Graaf, S. Herfst, G. F. Rimmelzwaan, A. D. Osterhaus, and R. A. Fouchier, "Determinants of virulence of influenza A virus," *European journal of clinical microbiology & infectious diseases*, vol. 33, no. 4, pp. 479–490, 2014.

[16] Y. Sugita, H. Sagara, T. Noda, and Y. Kawaoka, "Configuration of viral ribonucleoprotein complexes within the influenza a virion," *Journal of virology*, vol. 87, no. 23, pp. 12 879–12 884, 2013.

[17] T. Deng, J. L. Sharps, and G. G. Brownlee, "Role of the influenza virus heterotrimeric rna polymerase complex in the initiation of replication," *Journal of General Virology*, vol. 87, no. 11, pp. 3373–3377, 2006.

[18] B. G. Hale, R. E. Randall, J. Ortín, and D. Jackson, "The multifunctional ns1 protein of influenza a viruses," *Journal of general virology*, 2008.

[19] D. Paterson and E. Fodor, "Emerging roles for the influenza a virus nuclear export protein (nep)," *PLoS pathogens*, vol. 8, no. 12, p. e1003019, 2012.

[20] S. Krauss, C. A. Obert, J. Franks, D. Walker, K. Jones, P. Seiler, L. Niles, S. P. Pryor, J. C. Obenauer, C. W. Naeve *et al.*, "Influenza in migratory birds and evidence of limited intercontinental virus exchange," *PLoS pathogens*, vol. 3, no. 11, p. e167, 2007.

[21] T. Kuiken, E. C. Holmes, J. McCauley, G. F. Rimmelzwaan, C. S. Williams, and B. T. Grenfell, "Host species barriers to influenza virus infections," *Science*, vol. 312, no. 5772, pp. 394–397, 2006.

[22] G. Neumann, T. Noda, and Y. Kawaoka, "Emergence and pandemic potential of swine-origin h1n1 influenza virus," *Nature*, vol. 459, no. 7249, p. 931, 2009.

[23] N. Cox and K. Subbarao, "Global epidemiology of influenza: past and present," *Annual review of medicine*, vol. 51, no. 1, pp. 407–421, 2000.

[24] A. Gambotto, S. M. Barratt-Boyes, M. D. de Jong, G. Neumann, and Y. Kawaoka, "Human infection with highly pathogenic h5n1 influenza virus," *The Lancet*, vol. 371, no. 9622, pp. 1464–1475, 2008.

[25] R. A. Fouchier, P. M. Schneeberger, F. W. Rozendaal, J. M. Broekman, S. A. Kemink, V. Munster, T. Kuiken, G. F. Rimmelzwaan, M. Schutten, G. J. van Doornum *et al.*, "Avian influenza a virus (h7n7) associated with human conjunctivitis and a fatal case of acute respiratory distress syndrome," *Proceedings of the National Academy of Sciences*, vol. 101, no. 5, pp. 1356–1361, 2004.

[26] S. Luo, Z. Xie, Z. Xie, L. Xie, L. Huang, J. Huang, X. Deng, T. Zeng, S. Wang, Y. Zhang *et al.*, "Surveillance of live poultry markets for low pathogenic avian influenza viruses in guangxi province, southern china, from 2012–2015," *Scientific reports*, vol. 7, no. 1, p. 17577, 2017.

[27] N. A. Tuan, P. H. My, T. T. K. Huong, N. T. Y. Chi, T. T. H. Thu, J. Carrique-Mas, M. T. Duong, N. D. Tho, N. D. Hoang, T. L. Thanh *et al.*, "Assessing evidence for avian-to-human transmission of influenza a/h9n2 virus in rural farming communities in northern vietnam," *Journal of General Virology*, vol. 98, no. 8, pp. 2011–2016, 2017.

[28] A. Khatua, A. Khatua, and E. Cambria, "A tale of two epidemics: Contextual word2vec for classifying twitter streams during outbreaks," *Information Processing & Management*, vol. 56, no. 1, pp. 247–257, 2019.

[29] Z.-Y. Yang, C.-J. Wei, W.-P. Kong, L. Wu, L. Xu, D. F. Smith, and G. J. Nabel, "Immunization by avian h5 influenza hemagglutinin mutants with

altered receptor binding specificity," *Science*, vol. 317, no. 5839, pp. 825–828, 2007.

[30] J.-S. Yeom, T. Kostova-Vassilevska, P. D. Barnes, D. R. Jefferson, and T. Oppelstrup, "Exploratory modeling and simulation of the evolutionary dynamics of single-stranded rna virus populations," in *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2017, pp. 263–272.

[31] R. Chen and E. C. Holmes, "Avian influenza virus exhibits rapid evolutionary dynamics," *Molecular biology and evolution*, vol. 23, no. 12, pp. 2336–2341, 2006.

[32] D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. Osterhaus, and R. A. Fouchier, "Mapping the antigenic and genetic evolution of influenza virus," *science*, vol. 305, no. 5682, pp. 371–376, 2004.

[33] D. Fleury, S. A. Wharton, J. J. Skehel, M. Knossow, and T. Bizebard, "Antigen distortion allows influenza virus to escape neutralization," *Nature structural biology*, vol. 5, no. 2, p. 119, 1998.

[34] A. C.-C. Shih, T.-C. Hsiao, M.-S. Ho, and W.-H. Li, "Simultaneous amino acid substitutions at antigenic sites drive influenza a hemagglutinin evolution," *Proceedings of the National Academy of Sciences*, vol. 104, no. 15, pp. 6283–6288, 2007.

[35] X. Du, Z. Wang, A. Wu, L. Song, Y. Cao, H. Hang, and T. Jiang, "Networks of genomic co-occurrence capture characteristics of human influenza a (h3n2) evolution," *Genome research*, vol. 18, no. 1, pp. 178–187, 2008.

[36] W. Zhang, Y. Shi, X. Lu, Y. Shu, J. Qi, and G. F. Gao, "An airborne transmissible avian influenza h5 hemagglutinin seen at the atomic level," *Science*, vol. 340, no. 6139, pp. 1463–1467, 2013.

[37] R. P. De Vries, W. Peng, O. C. Grant, A. J. Thompson, X. Zhu, K. M. Bouwman, A. T. T. de la Pena, M. J. van Breemen, I. N. A. Wickramasinghe, C. A. de Haan *et al.*, "Three mutations switch h7n9 influenza to human-type receptor specificity," *PLoS pathogens*, vol. 13, no. 6, p. e1006390, 2017.

[38] E. Visher, S. E. Whitefield, J. T. McCrone, W. Fitzsimmons, and A. S. Lauring, "The mutational robustness of influenza a virus," *PLoS pathogens*, vol. 12, no. 8, p. e1005856, 2016.

[39] M. D. Pauly, M. C. Procario, and A. S. Lauring, "A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza a viruses," *Elife*, vol. 6, p. e26437, 2017.

[40] D. M. Lyons and A. S. Lauring, "Mutation and epistasis in influenza virus evolution," *Viruses*, vol. 10, no. 8, p. 407, 2018.

[41] C. Li and H. Chen, "Enhancement of influenza virus transmission by gene reassortment," in *Influenza Pathogenesis and Control-Volume I*. Springer, 2014, pp. 185–204.

[42] J. Steel and A. C. Lowen, "Influenza a virus reassortment," in *Influenza Pathogenesis and Control-Volume I*. Springer, 2014, pp. 377–401.

[43] N. S.-O. I. A. H. V. I. Team, "Emergence of a novel swine-origin influenza a (h1n1) virus in humans," *New England journal of medicine*, vol. 360, no. 25, pp. 2605–2615, 2009.

[44] S. E. Lindstrom, N. J. Cox, and A. Klimov, "Genetic analysis of human h2n2 and early h3n2 influenza viruses, 1957–1972: evidence for genetic divergence and multiple reassortment events," *Virology*, vol. 328, no. 1, pp. 101–119, 2004.

[45] G. J. Smith, J. Bahl, D. Vijaykrishna, J. Zhang, L. L. Poon, H. Chen, R. G. Webster, J. M. Peiris, and Y. Guan, "Dating the emergence of pandemic influenza viruses," *Proceedings of the National Academy of Sciences*, vol. 106, no. 28, pp. 11 709–11 712, 2009.

[46] K. B. Westgeest, C. A. Russell, X. Lin, M. I. Spronken, T. M. Bestebroer, J. Bahl, R. van Beek, E. Skepner, R. A. Halpin, J. C. de Jong *et al.*, "Genomewide analysis of reassortment and evolution of human influenza a (h3n2) viruses circulating between 1968 and 2011," *Journal of virology*, vol. 88, no. 5, pp. 2844–2857, 2014.

[47] A. H. Reid and J. K. Taubenberger, "The origin of the 1918 pandemic influenza virus: a continuing enigma," *Journal of general virology*, vol. 84,

no. 9, pp. 2285–2292, 2003.

[48] W. Liu, H. Fan, J. Raghwani, T. T.-Y. Lam, J. Li, O. G. Pybus, H.-W. Yao, Y. Wo, K. Liu, X.-P. An *et al.*, "Occurrence and reassortment of avian influenza a (h7n9) viruses derived from coinfected birds in china," *Journal of virology*, vol. 88, no. 22, pp. 13 344–13 351, 2014.

[49] M. Gu, L. Xu, X. Wang, and X. Liu, "Current situation of h9n2 subtype avian influenza in china," *Veterinary research*, vol. 48, no. 1, p. 49, 2017.

[50] X. Zhou, J. Zheng, F. X. Ivan, R. Yin, S. Ranganathan, V. T. Chow, and C.-K. Kwoh, "Computational analysis of the receptor binding specificity of novel influenza a/h7n9 viruses," *BMC genomics*, vol. 19, no. 2, p. 88, 2018.

[51] R. Yin, C. K. Kwoh, and J. Zheng, "Whole genome sequencing analysis," 2019.

[52] H.-L. Yen and J. M. Peiris, "Mapping antibody epitopes of the avian h5n1 influenza virus," *PLoS medicine*, vol. 6, no. 4, p. e1000064, 2009.

[53] R. Seyer, E. R. Hrincius, D. Ritzel, M. Abt, A. Mellmann, H. Marjuki, J. Kühn, T. Wolff, S. Ludwig, and C. Ehrhardt, "Synergistic adaptive mutations in the hemagglutinin and polymerase acidic protein lead to increased virulence of pandemic 2009 h1n1 influenza a virus in mice," *Journal of Infectious Diseases*, vol. 205, no. 2, pp. 262–271, 2011.

[54] R. Yin, X. Zhou, F. X. Ivan, J. Zheng, V. T. Chow, and C. K. Kwoh, "Identification of potential critical virulent sites based on hemagglutinin of influenza a virus in past pandemic strains," in *Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science*. ACM, 2017, pp. 30–36.

[55] M. Imai and Y. Kawaoka, "The role of receptor binding specificity in interspecies transmission of influenza viruses," *Current opinion in virology*, vol. 2, no. 2, pp. 160–167, 2012.

[56] A. Gambaryan, N. Lomakina, E. Boravleva, L. Mochalova, G. Sadykova, A. Prilipov, T. Matrosovich, and M. Matrosovich, "Mutations in hemagglutinin and polymerase alter the virulence of pandemic a (h1n1) influenza

virus," *Molecular Biology*, vol. 52, no. 4, pp. 556–569, 2018.

[57] H. Song, J. Qi, H. Xiao, Y. Bi, W. Zhang, Y. Xu, F. Wang, Y. Shi, and G. F. Gao, "Avian-to-human receptor-binding adaptation by influenza a virus hemagglutinin h4," *Cell reports*, vol. 20, no. 5, pp. 1201–1214, 2017.

[58] E. de Vries, R. P. de Vries, M. J. Wienholts, C. E. Floris, M.-S. Jacobs, A. van den Heuvel, P. J. Rottier, and C. A. de Haan, "Influenza a virus entry into cells lacking sialylated n-glycans," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7457–7462, 2012.

[59] J. LeGoff, D. Rousset, G. Abou-Jaoude, A. Scemla, P. Ribaud, S. Mercier-Delarue, V. Caro, V. Enouf, F. Simon, J.-M. Molina *et al.*, "I223r mutation in influenza a (h1n1) pdm09 neuraminidase confers reduced susceptibility to oseltamivir and zanamivir and enhanced resistance with h275y," *PLoS One*, vol. 7, no. 8, p. e37095, 2012.

[60] G. Gabriel, B. Dauber, T. Wolff, O. Planz, H.-D. Klenk, and J. Stech, "The viral polymerase mediates adaptation of an avian influenza virus to a mammalian host," *Proceedings of the National Academy of Sciences*, vol. 102, no. 51, pp. 18 590–18 595, 2005.

[61] S. Yamada, M. Hatta, B. L. Staker, S. Watanabe, M. Imai, K. Shinya, Y. Sakai-Tagawa, M. Ito, M. Ozawa, T. Watanabe *et al.*, "Biological and structural characterization of a host-adapting amino acid in influenza virus," *PLoS pathogens*, vol. 6, no. 8, p. e1001034, 2010.

[62] R. P. Kamal, J. M. Katz, and I. A. York, "Molecular determinants of influenza virus pathogenesis in mice," in *Influenza Pathogenesis and Control-Volume I*. Springer, 2014, pp. 243–274.

[63] M.-S. Song, P. N. Q. Pascua, J. H. Lee, Y. H. Baek, O.-J. Lee, C.-J. Kim, H. Kim, R. J. Webby, R. G. Webster, and Y. K. Choi, "The polymerase acidic protein gene of influenza a virus contributes to pathogenicity in a mouse model," *Journal of virology*, vol. 83, no. 23, pp. 12 325–12 335, 2009.

[64] J. Song, J. Xu, J. Shi, Y. Li, and H. Chen, "Synergistic effect of s224p and n383d substitutions in the pa of h5n1 avian influenza virus contributes to mammalian adaptation," *Scientific reports*, vol. 5, p. 10510, 2015.

[65] S. H. Seo, E. Hoffmann, and R. G. Webster, "Lethal h5n1 influenza viruses escape host anti-viral cytokine responses," *Nature medicine*, vol. 8, no. 9, p. 950, 2002.

[66] G. M. Conenello, D. Zamarin, L. A. Perrone, T. Tumpey, and P. Palese, "A single mutation in the pb1-f2 of h5n1 (hk/97) and 1918 influenza a viruses contributes to increased virulence," *PLoS pathogens*, vol. 3, no. 10, p. e141, 2007.

[67] T. R. Maines, A. Jayaraman, J. A. Belser, D. A. Wadford, C. Pappas, H. Zeng, K. M. Gustin, M. B. Pearce, K. Viswanathan, Z. H. Shriver *et al.*, "Transmission and pathogenesis of swine-origin 2009 a (h1n1) influenza viruses in ferrets and mice," *Science*, vol. 325, no. 5939, pp. 484–487, 2009.

[68] T. Watanabe, K. Shinya, S. Watanabe, M. Imai, M. Hatta, C. Li, B. F. Wolter, G. Neumann, A. Hanson, M. Ozawa *et al.*, "Avian-type receptor-binding ability can increase influenza virus pathogenicity in macaques," *Journal of virology*, pp. JVI–00 859, 2011.

[69] S. Herfst, E. J. Schrauwen, M. Linster, S. Chutinimitkul, E. de Wit, V. J. Munster, E. M. Sorrell, T. M. Bestebroer, D. F. Burke, D. J. Smith *et al.*, "Airborne transmission of influenza a/h5n1 virus between ferrets," *science*, vol. 336, no. 6088, pp. 1534–1541, 2012.

[70] R. Zhou, P. Das, and A. K. Royyuru, "Single mutation induced h3n2 hemagglutinin antibody neutralization: a free energy perturbation study," *The Journal of Physical Chemistry B*, vol. 112, no. 49, pp. 15 813–15 820, 2008.

[71] R. Goya, M. G. Sun, R. D. Morin, G. Leung, G. Ha, K. C. Wiegand, J. Senz, A. Crisan, M. A. Marra, M. Hirst *et al.*, "Snvmix: predicting single nucleotide variants from next-generation sequencing of tumors," *Bioinformatics*, vol. 26, no. 6, pp. 730–736, 2010.

[72] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature methods*, vol. 12, no. 10, p. 931, 2015.

[73] D. Barash and A. Churkin, "Mutational analysis in rnas: comparing programs for rna deleterious mutation prediction," *Briefings in bioinformatics*, vol. 12, no. 2, pp. 104–114, 2010.

[74] L. Quan, Q. Lv, and Y. Zhang, "Strum: structure-based prediction of protein stability changes upon single-point mutation," *Bioinformatics*, vol. 32, no. 19, pp. 2936–2946, 2016.

[75] R. S. Mandal, S. Panda, and S. Das, "In silico prediction of drug resistance due to s247r mutation of influenza h1n1 neuraminidase protein," *Journal of Biomolecular Structure and Dynamics*, vol. 36, no. 4, pp. 966–980, 2018.

[76] A. Gillman, S. Muradrasoli, H. Söderström, F. Holmberg, N. Latorre-Margalef, C. Tolf, J. Waldenström, G. Gunnarsson, B. Olsen, and J. D. Järhult, "Oseltamivir-resistant influenza a (h1n1) virus strain with an h274y mutation in neuraminidase persists without drug pressure in infected mallards," *Appl. Environ. Microbiol.*, vol. 81, no. 7, pp. 2378–2383, 2015.

[77] M. A. Salama, A. E. Hassanien, and A. Mostafa, "The prediction of virus mutation using neural networks and rough set techniques," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2016, no. 1, p. 10, 2016.

[78] R. A. Neher, T. Bedford, R. S. Daniels, C. A. Russell, and B. I. Shraiman, "Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses," *Proceedings of the National Academy of Sciences*, vol. 113, no. 12, pp. E1701–E1709, 2016.

[79] M. Łuksza and M. Lässig, "A predictive fitness model for influenza," *Nature*, vol. 507, no. 7490, p. 57, 2014.

[80] W. Zhu, Y. Zhu, K. Qin, Z. Yu, R. Gao, H. Yu, J. Zhou, and Y. Shu, "Mutations in polymerase genes enhanced the virulence of 2009 pandemic h1n1 influenza virus in mice," *PloS one*, vol. 7, no. 3, p. e33383, 2012.

[81] Y. Peng, W. Zhu, Z. Feng, Z. Zhu, Z. Zhang, Y. Chen, S. Liu, A. Wu, D. Wang, Y. Shu *et al.*, "Identification of genome-wide nucleotide sites associated with mammalian virulence in influenza a viruses," *bioRxiv*, p. 416586, 2018.

[82] E. C. Holmes, E. Ghedin, N. Miller, J. Taylor, Y. Bao, K. St George, B. T. Grenfell, S. L. Salzberg, C. M. Fraser, D. J. Lipman *et al.*, "Whole-genome analysis of human influenza a virus reveals multiple persistent lineages and reassortment among recent h3n2 viruses," *PLoS biology*, vol. 3, no. 9, p. e300, 2005.

[83] A. I. Karasin, S. Carman, and C. 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*, vol. 44, no. 3, pp. 1123–1126, 2006.

[84] 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 research*, vol. 68, no. 1, pp. 71–85, 2000.

[85] A. I. Karasin, J. Landgraf, S. Swenson, G. Erickson, S. Goyal, M. Woodruff, G. Scherba, G. Anderson, and C. W. Olsen, "Genetic characterization of H1N2 influenza A viruses isolated from pigs throughout the united states," *Journal of clinical microbiology*, vol. 40, no. 3, pp. 1073–1079, 2002.

[86] C. Kingsford, N. Nagarajan, and S. L. Salzberg, "2009 swine-origin influenza a (H1N1) resembles previous influenza isolates," *Plos one*, vol. 4, no. 7, p. e6402, 2009.

[87] C. W. Olsen, A. I. Karasin, S. Carman, Y. Li, N. Bastien, D. Ojkic, D. Alves, G. Charbonneau, B. M. Henning, D. E. Low *et al.*, "Triple reassortant H3N2 influenza A viruses, canada, 2005," *Emerging infectious diseases*, vol. 12, no. 7, p. 1132, 2006.

[88] H. Khiabanian, V. Trifonov, and R. Rabadan, "Reassortment patterns in swine influenza viruses," *PloS one*, vol. 4, no. 10, p. e7366, 2009.

[89] U. C. de Silva, H. Tanaka, S. Nakamura, N. Goto, and T. Yasunaga, "A comprehensive analysis of reassortment in influenza A virus," *Biology open*, vol. 1, no. 4, pp. 385–390, 2012.

[90] N. Nagarajan and C. Kingsford, "Giraf: robust, computational identification of influenza reassortments via graph mining," *Nucleic acids research*, vol. 39, no. 6, pp. e34–e34, 2010.

[91] ——, "Uncovering genomic reassortments among influenza strains by enumerating maximal bicliques," in *Bioinformatics and Biomedicine, 2008. BIBM'08. IEEE International Conference on*. IEEE, 2008, pp. 223–230.

[92] V. Svinti, J. A. Cotton, and J. O. McInerney, "New approaches for unravelling reassortment pathways," *BMC evolutionary biology*, vol. 13, no. 1, p. 1, 2013.

[93] A. Yurovsky and B. M. Moret, "Flurf, an automated flu virus reassortment finder based on phylogenetic trees," in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2010, pp. 579–584.

[94] R. Rabadan, A. J. Levine, and M. Krasnitz, "Non-random reassortment in human influenza A viruses," *Influenza and other respiratory viruses*, vol. 2, no. 1, pp. 9–22, 2008.

[95] E. Ghedin, N. A. Sengamalay, M. Shumway, J. Zaborsky, T. Feldblyum, V. Subbu, D. J. Spiro, J. Sitz, H. Koo, P. Bolotov *et al.*, "Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution," *Nature*, vol. 437, no. 7062, p. 1162, 2005.

[96] J. De Jong, A. Palache, W. Beyer, G. Rimmelzwaan, A. Boon, and A. Osterhaus, "Haemagglutination-inhibiting antibody to influenza virus." *Developments in biologicals*, vol. 115, pp. 63–73, 2003.

[97] J. L. Barnett, J. Yang, Z. Cai, T. Zhang, and X.-F. Wan, "Antigenmap 3d: an online antigenic cartography resource," *Bioinformatics*, vol. 28, no. 9, pp. 1292–1293, 2012.

[98] A. Lorusso, A. L. Vincent, M. L. Harland, D. Alt, D. O. Bayles, S. L. Swenson, M. R. Gramer, C. A. Russell, D. J. Smith, K. M. Lager *et al.*, "Genetic and antigenic characterization of h1 influenza viruses from united states swine from 2008," *Journal of General Virology*, vol. 92, no. 4, pp. 919–930, 2011.

[99] M. Liu, X. Zhao, S. Hua, X. Du, Y. Peng, X. Li, Y. Lan, D. Wang, A. Wu, Y. Shu *et al.*, "Antigenic patterns and evolution of the human influenza a (h1n1) virus," *Scientific reports*, vol. 5, 2015.

[100] T. Bedford, S. Riley, I. G. Barr, S. Broor, M. Chadha, N. J. Cox, R. S. Daniels, C. P. Gunasekaran, A. C. Hurt, A. Kelso *et al.*, "Global circulation patterns of seasonal influenza viruses vary with antigenic drift," *Nature*, vol. 523, no. 7559, pp. 217–220, 2015.

[101] X. Du, L. Dong, Y. Lan, Y. Peng, A. Wu, Y. Zhang, W. Huang, D. Wang, M. Wang, Y. Guo *et al.*, "Mapping of h3n2 influenza antigenic evolution in china reveals a strategy for vaccine strain recommendation," *Nature communications*, vol. 3, p. 709, 2012.

[102] M. Zacour, B. J. Ward, A. Brewer, P. Tang, G. Boivin, Y. Li, M. Warhuus, S. A. McNeil, J. J. LeBlanc, and T. F. Hatchette, "Standardization of hemagglutination inhibition assay for influenza serology allows for high reproducibility between laboratories," *Clinical and Vaccine Immunology*, vol. 23, no. 3, pp. 236–242, 2016.

[103] P. Kitikoon and A. L. Vincent, "Microneutralization assay for swine influenza virus in swine serum," in *Animal Influenza Virus*. Springer, 2014, pp. 325–335.

[104] R. M. Lequin, "Enzyme immunoassay (eia)/enzyme-linked immunosorbent assay (elisa)," *Clinical chemistry*, vol. 51, no. 12, pp. 2415–2418, 2005.

[105] Y. Yao, X. Li, B. Liao, L. Huang, P. He, F. Wang, J. Yang, H. Sun, Y. Zhao, and J. Yang, "Predicting influenza antigenicity from hemagglutintin sequence data based on a joint random forest method," *Scientific Reports*, vol. 7, 2017.

[106] Y. Peng, D. Wang, J. Wang, K. Li, Z. Tan, Y. Shu, and T. Jiang, "A universal computational model for predicting antigenic variants of influenza a virus based on conserved antigenic structures," *Scientific Reports*, vol. 7, p. 42051, 2017.

[107] J. Qiu, T. Qiu, Y. Yang, D. Wu, and Z. Cao, "Incorporating structure context of ha protein to improve antigenicity calculation for influenza virus

a/h3n2," *Scientific reports*, vol. 6, p. 31156, 2016.

[108] X. Zhou, R. Yin, C.-K. Kwoh, and J. Zheng, "A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza a viruses," *BMC genomics*, vol. 19, no. 10, pp. 145–154, 2018.

[109] L. Han, L. Li, F. Wen, L. Zhong, T. Zhang, and X.-F. Wan, "Graph-guided multi-task sparse learning model: a method for identifying antigenic variants of influenza a (h3n2) virus," *Bioinformatics*, vol. 35, no. 1, pp. 77–87, 2018.

[110] H. Chen, X. Zhou, J. Zheng, and C.-K. Kwoh, "Rules of co-occurring mutations characterize the antigenic evolution of human influenza a/h3n2, a/h1n1 and b viruses," *BMC medical genomics*, vol. 9, no. 3, p. 69, 2016.

[111] Y. Li, D. L. Bostick, C. B. Sullivan, J. L. Myers, S. B. Griesemer, K. S. George, J. B. Plotkin, and S. E. Hensley, "Single hemagglutinin mutations that alter both antigenicity and receptor binding avidity influence influenza virus antigenic clustering," *Journal of virology*, pp. JVI–01 023, 2013.

[112] P. Wang, W. Zhu, B. Liao, L. Cai, L. Peng, and J. Yang, "Predicting influenza antigenicity by matrix completion with antigen and antiserum similarity," *Frontiers in microbiology*, vol. 9, p. 2500, 2018.

[113] X. Ren, Y. Li, X. Liu, X. Shen, W. Gao, and J. Li, "Computational identification of antigenicity-associated sites in the hemagglutinin protein of a/h1n1 seasonal influenza virus," *PloS one*, vol. 10, no. 5, p. e0126742, 2015.

[114] H. Sun, J. Yang, T. Zhang, L.-P. Long, K. Jia, G. Yang, R. J. Webby, and X.-F. Wan, "Using sequence data to infer the antigenicity of influenza virus," *MBio*, vol. 4, no. 4, pp. e00 230–13, 2013.

[115] Y.-C. Liao, M.-S. Lee, C.-Y. Ko, and C. A. Hsiung, "Bioinformatics models for predicting antigenic variants of influenza a/h3n2 virus," *Bioinformatics*, vol. 24, no. 4, pp. 505–512, 2008.

[116] M.-S. Lee and J. S.-E. Chen, "Predicting antigenic variants of influenza a/h3n2 viruses," *Emerging infectious diseases*, vol. 10, no. 8, p. 1385, 2004.

[117] W. D. Lees, D. S. Moss, and A. J. Shepherd, "A computational analysis of the antigenic properties of haemagglutinin in influenza a h3n2," *Bioinformatics*, vol. 26, no. 11, pp. 1403–1408, 2010.

[118] F. X. Ivan, X. Zhou, A. Deshpande, R. Yin, J. Zheng, and C. K. Kwoh, "Phylogenetic tree based method for uncovering co-mutational site-pairs in influenza viruses," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017, pp. 21–26.

[119] R. Yin, E. Luusua, J. Dabrowski, Y. Zhang, and C. K. Kwoh, "Tempel: time-series mutation prediction of influenza a viruses via attention-based recurrent neural networks," *Bioinformatics*, vol. 36, no. 9, pp. 2697–2704, 2020.

[120] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein secondary structure prediction using deep convolutional neural fields," *Scientific reports*, vol. 6, p. 18962, 2016.

[121] R. Yin, X. Zhou, S. Rashid, and C. K. Kwoh, "Hopper: an adaptive model for probability estimation of influenza reassortment through host prediction," *BMC medical genomics*, vol. 13, no. 1, p. 9, 2020.

[122] R. Yin, Y. Zhang, X. Zhou, and C. K. Kwoh, "Time series computational prediction of vaccines for influenza a h3n2 with recurrent neural networks," *Journal of Bioinformatics and Computational Biology*, vol. 18, no. 01, p. 2040002, 2020.

[123] W. W. Group, W. K. Ampofo, N. Baylor, S. Cobey, N. J. Cox, S. Daves, S. Edwards, N. Ferguson, G. Grohmann, A. Hay *et al.*, "Improving influenza vaccine virus selectionreport of a who informal consultation held at who headquarters, geneva, switzerland, 14–16 june 2010," *Influenza and other respiratory viruses*, vol. 6, no. 2, pp. 142–152, 2012.

[124] T. R. Klingen, S. Reimering, C. A. Guzmán, and A. C. McHardy, "In silico vaccine strain prediction for human influenza viruses," *Trends in microbiology*, vol. 26, no. 2, pp. 119–131, 2018.

[125] H. Xie, X.-F. Wan, Z. Ye, E. P. Plant, Y. Zhao, Y. Xu, X. Li, C. Finch, N. Zhao, T. Kawano *et al.*, "H3n2 mismatch of 2014–15 northern

hemisphere influenza vaccines and head-to-head comparison between human and ferret antisera derived antigenic maps," *Scientific reports*, vol. 5, p. 15279, 2015.

[126] S. E. Hensley, S. R. Das, A. L. Bailey, L. M. Schmidt, H. D. Hickman, A. Jayaraman, K. Viswanathan, R. Raman, R. Sasisekharan, J. R. Bennink *et al.*, "Hemagglutinin receptor binding avidity drives influenza a virus antigenic drift," *Science*, vol. 326, no. 5953, pp. 734–736, 2009.

[127] P. Doshi, "The elusive definition of pandemic influenza," *Bulletin of the World Health Organization*, vol. 89, pp. 532–538, 2011.

[128] C. T.-T. Su, X. Ouyang, J. Zheng, and C.-K. Kwoh, "Structural analysis of the novel influenza a (h7n9) viral neuraminidase interactions with current approved neuraminidase inhibitors oseltamivir, zanamivir, and peramivir in the presence of mutation r289k," *BMC bioinformatics*, vol. 14, no. 16, p. S7, 2013.

[129] G. Wu and S.-m. Yan, "Mutation trend of hemagglutinin of influenza a virus: a review from a computational mutation viewpoint," *Acta Pharmacologica Sinica*, vol. 27, no. 5, pp. 513–526, 2006.

[130] O. Miotto, A. Heiny, T. W. Tan, J. T. August, and V. Brusic, "Identification of human-to-human transmissibility factors in pb2 proteins of influenza a by large-scale mutual information analysis," *BMC bioinformatics*, vol. 9, no. 1, p. S18, 2008.

[131] 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," *Journal of virology*, vol. 82, no. 2, pp. 596–601, 2008.

[132] 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*, vol. 182, no. 2, pp. 475–485, 1991.

[133] D. F. Burke and D. J. Smith, "A recommended numbering scheme for influenza a ha subtypes," *PloS one*, vol. 9, no. 11, p. e112302, 2014.

[134] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen, "Signalp 4.0: discriminating signal peptides from transmembrane regions," *Nature methods*, vol. 8, no. 10, pp. 785–786, 2011.

[135] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez *et al.*, "Machine learning in bioinformatics," *Briefings in bioinformatics*, pp. 86–112, 2006.

[136] R. J. Urbanowicz and J. H. Moore, "Learning classifier systems: a complete introduction, review, and roadmap," *Journal of Artificial Evolution and Applications*, vol. 2009, p. 1, 2009.

[137] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[138] M. W. Deem and K. Pan, "The epitope regions of h1-subtype influenza a, with application to vaccine efficacy," *Protein Engineering, Design & Selection*, vol. 22, no. 9, pp. 543–546, 2009.

[139] S. J. Baigent and J. W. McCauley, "Influenza type a in humans, mammals and birds: Determinants of virus virulence, host-range and interspecies transmission," *Bioessays*, vol. 25, no. 7, pp. 657–671, 2003.

[140] J. C. Kash, L. Qi, V. G. Dugan, B. W. Jagger, R. J. Hrabal, M. J. Memoli, D. M. Morens, and J. K. Taubenberger, "Prior infection with classical swine h1n1 influenza viruses is associated with protective immunity to the 2009 pandemic h1n1 virus," *Influenza and other respiratory viruses*, vol. 4, no. 3, pp. 121–127, 2010.

[141] K. Bragstad, L. P. Nielsen, and A. Fomsgaard, "The evolution of human influenza a viruses from 1999 to 2006: a complete genome study," *Virology journal*, vol. 5, no. 1, p. 40, 2008.

[142] A. Piralla, E. Pariani, F. Rovida, G. Campanini, A. Muzzi, V. Emmi, G. A. Iotti, A. Pesenti, P. G. Conaldi, A. Zanetti *et al.*, "Segregation of virulent influenza a (h1n1) variants in the lower respiratory tract of critically ill patients during the 2010–2011 seasonal epidemic," *PLoS One*, vol. 6, no. 12, p. e28332, 2011.

[143] N. Tewawong, S. Prachayangprecha, P. Vichiwattana, S. Korkong, S. Klinfueng, S. Vongpunsawad, T. Thongmee, A. Theamboonlers, and Y. Poovorawan, "Assessing antigenic drift of seasonal influenza a (h3n2) and a (h1n1) pdm09 viruses," *PloS one*, vol. 10, no. 10, p. e0139958, 2015.

[144] T. E. Ginting, K. Shinya, Y. Kyan, A. Makino, N. Matsumoto, S. Kaneda, and Y. Kawaoka, "Amino acid changes in hemagglutinin contribute to the replication of oseltamivir-resistant h1n1 influenza viruses," *Journal of virology*, vol. 86, no. 1, pp. 121–127, 2012.

[145] N. M. Bouvier and P. Palese, "The biology of influenza viruses," *Vaccine*, vol. 26, pp. D49–D53, 2008.

[146] R. G. Webster, "Influenza: an emerging disease." *Emerging infectious diseases*, vol. 4, no. 3, p. 436, 1998.

[147] N. Marshall, L. Priyamvada, Z. Ende, J. Steel, and A. C. Lowen, "Influenza virus reassortment occurs with high frequency in the absence of segment mismatch," *PLoS pathogens*, vol. 9, no. 6, p. e1003421, 2013.

[148] D. Vijaykrishna, R. Mukerji, and G. J. Smith, "Rna virus reassortment: an evolutionary mechanism for host jumps and immune evasion," *PLoS pathogens*, vol. 11, no. 7, p. e1004902, 2015.

[149] E. De Clercq, "Antiviral agents active against influenza A viruses," *Nature reviews drug discovery*, vol. 5, no. 12, p. 1015, 2006.

[150] S. Chang, J. Zhang, X. Liao, X. Zhu, D. Wang, J. Zhu, T. Feng, B. Zhu, G. F. Gao, J. Wang *et al.*, "Influenza virus database (ivdb): an integrated information resource and analysis platform for influenza virus research," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D376–D380, 2006.

[151] C. A. Macken, R. J. Webby, and W. J. Bruno, "Genotype turnover by reassortment of replication complex genes from avian influenza A virus," *Journal of general virology*, vol. 87, no. 10, pp. 2803–2815, 2006.

[152] S. L. Salzberg, C. Kingsford, G. Cattoli, D. J. Spiro, D. A. Janies, M. M. Aly, I. H. Brown, E. Couacy-Hymann, G. M. De Mia, D. H. Dung *et al.*, "Genome analysis linking recent european and african influenza (h5n1) viruses," *Emerging infectious diseases*, vol. 13, no. 5, p. 713, 2007.

[153] M. Villa and M. Lässig, "Fitness cost of reassortment in human influenza," *PLoS pathogens*, vol. 13, no. 11, p. e1006685, 2017.

[154] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[155] Y. Shu and J. McCauley, "Gisaid: Global initiative on sharing all influenza data–from vision to reality," *Eurosurveillance*, vol. 22, no. 13, 2017.

[156] S. Kawashima and M. Kanehisa, "Aaindex: amino acid index database," *Nucleic acids research*, vol. 28, no. 1, pp. 374–374, 2000.

[157] I. Dubchak, I. Muchnik, S. R. Holbrook, and S.-H. Kim, "Prediction of protein folding class using global description of amino acid sequence," *Proceedings of the National Academy of Sciences*, vol. 92, no. 19, pp. 8700–8704, 1995.

[158] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S.-H. Kim, "Recognition of a protein fold in the context of the SCOP classification," *Proteins: structure, function, and bioinformatics*, vol. 35, no. 4, pp. 401–407, 1999.

[159] K. Tomii and M. Kanehisa, "Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins," *Protein Engineering, Design and Selection*, vol. 9, no. 1, pp. 27–36, 1996.

[160] C. Eng, J. C. Tong, and T. W. Tan, "Predicting zoonotic risk of influenza a viruses from host tropism protein signature using random forest," *International journal of molecular sciences*, vol. 18, no. 6, p. 1135, 2017.

[161] R. Yin, X. Zhou, J. Zheng, and C. K. Kwoh, "Computational identification of physicochemical signatures for host tropism of influenza A virus." *Journal of bioinformatics and computational biology*, p. 1840023, 2018.

[162] T. K. Ho, "Random decision forests," in *Document analysis and recognition, 1995., proceedings of the third international conference on*, vol. 1. IEEE, 1995, pp. 278–282.

[163] M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *Icml*, vol. 97. Nashville, USA, 1997, pp. 179–186.

[164] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures

for models assessment over imbalanced datasets," *J Inf Eng Appl*, vol. 3, no. 10, 2013.

[165] P. E. Black, "Ratcliff/obershelp pattern recognition," *Dictionary of algorithms and data structures*, vol. 17, 2004.

[166] H. Boström, "Calibrating random forests," in *2008 Seventh International Conference on Machine Learning and Applications*. IEEE, 2008, pp. 121–126.

[167] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 625–632.

[168] H. Boström, "Estimating class probabilities in random forests," in *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on 13-15 Dec. 2007*. IEEE, 2007, pp. 211–216.

[169] C. Li, "Probability estimation in random forests," 2013.

[170] M. A. Olson and A. J. Wyner, "Making sense of random forest probabilities: a kernel perspective," *arXiv preprint arXiv:1812.05792*, 2018.

[171] J. D. Malley, J. Kruppa, A. Dasgupta, K. G. Malley, and A. Ziegler, "Probability machines," *Methods of information in medicine*, vol. 51, no. 01, pp. 74–81, 2012.

[172] V. Bourret, J. Lyall, S. D. Frost, A. Teillaud, C. A. Smith, S. Leclaire, J. Fu, S. Gandon, J.-L. Guérin, and L. S. Tiley, "Adaptation of avian influenza virus to a swine host," *Virus evolution*, vol. 3, no. 1, p. vex007, 2017.

[173] M. R. Castrucci, I. Donatelli, L. Sidoli, G. Barigazzi, Y. Kawaoka, and R. G. Webster, "Genetic reassortment between avian and human influenza A viruses in italian pigs," *Virology*, vol. 193, no. 1, pp. 503–506, 1993.

[174] R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka, "Evolution and ecology of influenza A viruses." *Microbiological reviews*, vol. 56, no. 1, pp. 152–179, 1992.

[175] T. Chambers, V. S. Hinshaw, Y. Kawaoka, B. Easterday, and R. Webster, "Influenza viral infection of swine in the united states 1988–1989,"

Archives of virology, vol. 116, no. 1-4, pp. 261–265, 1991.

[176] C. Olsen, S. Carey, L. Hinshaw, and A. Karasin, "Virologic and serologic surveillance for human, swine and avian influenza virus infections among pigs in the north-central united states," *Archives of virology*, vol. 145, no. 7, pp. 1399–1419, 2000.

[177] K. Li, Y. Guan, J. Wang, G. Smith, K. Xu, L. Duan, A. Rahardjo, P. Puthavathana, C. Buranathai, T. Nguyen *et al.*, "Genesis of a highly pathogenic and potentially pandemic h5n1 influenza virus in eastern asia," *Nature*, vol. 430, no. 6996, p. 209, 2004.

[178] C. E. Mills, J. M. Robins, and M. Lipsitch, "Transmissibility of 1918 pandemic influenza," *Nature*, vol. 432, no. 7019, p. 904, 2004.

[179] M. I. Nelson, C. Viboud, L. Simonsen, R. T. Bennett, S. B. Griesemer, K. S. George, J. Taylor, D. J. Spiro, N. A. Sengamalay, E. Ghedin *et al.*, "Multiple reassortment events in the evolutionary history of h1n1 influenza a virus since 1918," *PLoS Pathogens*, vol. 4, no. 2, p. e1000012, 2008.

[180] I. M. Berry, M. C. Melendrez, T. Li, A. W. Hawksworth, G. T. Brice, P. J. Blair, E. S. Halsey, M. Williams, S. Fernandez, I.-K. Yoon *et al.*, "Frequency of influenza h3n2 intra-subtype reassortment: attributes and implications of reassortant spread," *BMC biology*, vol. 14, no. 1, p. 117, 2016.

[181] D. C. Lye, B. S. Ang, and Y.-S. Leo, "Review of human infections with avian influenza h5n1 and proposed local clinical management guideline," *Annals-academy of medicine Singapore*, vol. 36, no. 4, p. 285, 2007.

[182] M. Gilbert, X. Xiao, D. U. Pfeiffer, M. Epprecht, S. Boles, C. Czarnecki, P. Chaitaweesub, W. Kalpravidh, P. Q. Minh, M. J. Otte *et al.*, "Mapping h5n1 highly pathogenic avian influenza risk in southeast asia," *Proceedings of the national academy of sciences*, vol. 105, no. 12, pp. 4769–4774, 2008.

[183] W. H. Organization *et al.*, "Recommended composition of influenza virus vaccines for use in the 2015-2016 northern hemisphere influenza season," *Weekly epidemiological record= Relevé épidémiologique hebdomadaire*, vol. 90, no. 11, pp. 97–108, 2015.

[184] P. P.-H. Cheung, I. B. Rogozin, K.-T. Choy, H. Y. Ng, J. S. M. Peiris, and H.-L. Yen, "Comparative mutational analyses of influenza a viruses," *Rna*, vol. 21, no. 1, pp. 36–47, 2015.

[185] Y. Poovorawan, S. Pyungporn, S. Prachayangprecha, and J. Makkoch, "Global alert to avian influenza virus infection: from h5n1 to h7n9," *Pathogens and global health*, vol. 107, no. 5, pp. 217–223, 2013.

[186] S. Su, Y. Bi, G. Wong, G. C. Gray, G. F. Gao, and S. Li, "Epidemiology, evolution, and recent outbreaks of avian influenza virus in china," *Journal of virology*, vol. 89, no. 17, pp. 8671–8676, 2015.

[187] M.-J. Ma, C. Liu, M.-N. Wu, T. Zhao, G.-L. Wang, Y. Yang, H.-J. Gu, P.-W. Cui, Y.-Y. Pang, Y.-Y. Tan *et al.*, "Influenza a (h7n9) virus antibody responses in survivors 1 year after infection, china, 2017," *Emerging infectious diseases*, vol. 24, no. 4, p. 663, 2018.

[188] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[189] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[190] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.

[191] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[192] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.

[193] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level

concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[194] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[195] K. Ganchev, J. Gillenwater, B. Taskar *et al.*, "Posterior regularization for structured latent variable models," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2001–2049, 2010.

[196] L.-a. Pirofski and A. Casadevall, "Q&a: What is a pathogen? a question that begs the point," *BMC biology*, vol. 10, no. 1, p. 6, 2012.

[197] F. X. Ivan and C.-K. Kwoh, "Rule-based meta-analysis reveals the major role of pb2 in influencing influenza a virus virulence in mice," *bioRxiv*, p. 556647, 2019.

[198] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen *et al.*, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 40, no. D1, pp. D13–D25, 2012.

[199] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[200] R. Yin, J. Tan, D. Akhila, X. Zhou, and C. K. Kwoh, "Inference of sequence homology by blast visualization of influenza genome set," in *Proceedings of the 9th International Conference on Computational Systems-Biology and Bioinformatics*, 2018, pp. 1–6.

[201] K. Katoh and D. M. Standley, "Mafft multiple sequence alignment software version 7: improvements in performance and usability," *Molecular biology and evolution*, vol. 30, no. 4, pp. 772–780, 2013.

[202] T. Hall, I. Biosciences, and C. Carlsbad, "Bioedit: an important software for molecular biology," *GERF Bull Biosci*, vol. 2, no. 1, pp. 60–61, 2011.

[203] Z. Hu, X. Ma, Z. Liu, E. Hovy, and E. Xing, "Harnessing deep neural networks with logic rules," *arXiv preprint arXiv:1603.06318*, 2016.

[204] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, "Risk prediction on

electronic health records with prior medical knowledge," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1910–1919.

[205] F. Pérez-Cruz, "Kullback-leibler divergence estimation of continuous distributions," in *2008 IEEE international symposium on information theory*. IEEE, 2008, pp. 1666–1670.

[206] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[207] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[208] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[209] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[210] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[211] W. H. Organization *et al.*, "Fact sheet no. 211. influenza (seasonal). april, 2009," 2010.

[212] J. Stevens, A. L. Corper, C. F. Basler, J. K. Taubenberger, P. Palese, and I. A. Wilson, "Structure of the uncleaved human h1 hemagglutinin from the extinct 1918 influenza virus," *Science*, vol. 303, no. 5665, pp. 1866–1870, 2004.

[213] A. Harding and N. Heaton, "Efforts to improve the seasonal influenza vaccine," *Vaccines*, vol. 6, no. 2, p. 19, 2018.

[214] G. A. Sautto, G. A. Kirchenbaum, and T. M. Ross, "Towards a universal

influenza vaccine: different approaches for one goal," *Virology journal*, vol. 15, no. 1, p. 17, 2018.

[215] A. Palache, W. Beyer, G. Rimmelzwaan, A. Boon, A. Osterhaus *et al.*, "Haemagglutination-inhibiting antibody to influenza virus." *Developments in biologicals*, vol. 115, pp. 63–73, 2003.

[216] J.-W. Huang, W.-F. Lin, and J.-M. Yang, "Antigenic sites of h1n1 influenza virus hemagglutinin revealed by natural isolates and inhibition assays," *Vaccine*, vol. 30, no. 44, pp. 6327–6337, 2012.

[217] W. T. Harvey, D. J. Benton, V. Gregory, J. P. Hall, R. S. Daniels, T. Bedford, D. T. Haydon, A. J. Hay, J. W. McCauley, and R. Reeve, "Identification of low-and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza a (h1n1) viruses," *PLoS pathogens*, vol. 12, no. 4, p. e1005526, 2016.

[218] R. Yin, V. H. Tran, X. Zhou, J. Zheng, and C. K. Kwoh, "Predicting antigenic variants of h1n1 influenza virus based on epidemics and pandemics using a stacking model," *PloS one*, vol. 13, no. 12, p. e0207777, 2018.

[219] S. W. Taju, T.-T.-D. Nguyen, N.-Q.-K. Le, R. M. I. Kusuma, and Y.-Y. Ou, "Deepefflux: a 2d convolutional neural network model for identifying families of efflux proteins in transporters," *Bioinformatics*, vol. 34, no. 18, pp. 3111–3117, 2018.

[220] Y. S. Vang and X. Xie, "Hla class i binding prediction via convolutional neural networks," *Bioinformatics*, vol. 33, no. 17, pp. 2658–2665, 2017.

[221] Z. Li and Y. Yu, "Protein secondary structure prediction using cascaded convolutional and recurrent neural networks," *arXiv preprint arXiv:1604.07176*, 2016.

[222] T. Sun, B. Zhou, L. Lai, and J. Pei, "Sequence-based prediction of protein protein interaction using a deep-learning algorithm," *BMC bioinformatics*, vol. 18, no. 1, p. 277, 2017.

[223] W. Ndifon, J. Dushoff, and S. A. Levin, "On the use of hemagglutination-inhibition for influenza surveillance: surveillance data are predictive of

influenza vaccine effectiveness," *Vaccine*, vol. 27, no. 18, pp. 2447–2452, 2009.

[224] K.-C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal chemistry*, vol. 11, no. 3, pp. 218–234, 2015.

[225] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[226] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188–1196.

[227] E. Asgari and M. R. Mofrad, "Continuous distributed representation of biological sequences for deep proteomics and genomics," *PloS one*, vol. 10, no. 11, p. e0141287, 2015.

[228] T. Bepler and B. Berger, "Learning protein sequence embeddings using information from structure," *arXiv preprint arXiv:1902.08661*, 2019.

[229] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2285–2294.

[230] T. K. Lee and T. Nguyen, "Protein family classification with neural networks," 2016.

[231] N. Q. K. Le and V.-N. Nguyen, "Snare-cnn: a 2d convolutional neural network architecture to identify snare proteins from high-throughput sequencing data," *PeerJ Computer Science*, vol. 5, p. e177, 2019.

[232] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, and D. R. Koes, "Protein–ligand scoring with convolutional neural networks," *Journal of chemical information and modeling*, vol. 57, no. 4, pp. 942–957, 2017.

[233] R. Gao, M. Wang, J. Zhou, Y. Fu, M. Liang, D. Guo, and J. Nie, "Prediction of enzyme function based on three parallel deep cnn and amino acid mutation," *International journal of molecular sciences*, vol. 20, no. 11, p. 2845, 2019.

[234] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[235] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[236] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[237] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[238] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[239] T. Tieleman and G. Hinton, "Rmsprop gradient optimization," *URL http://www. cs. toronto. edu/tijmen/csc321/slides/lecture_slides_lec6. pdf*, 2014.

[240] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[241] C. A. Russell, P. M. Kasson, R. O. Donis, S. Riley, J. Dunbar, A. Rambaut, J. Asher, S. Burke, C. T. Davis, R. J. Garten *et al.*, "Science forum: Improving pandemic influenza risk assessment," *Elife*, vol. 3, p. e03883, 2014.

[242] K. A. Herrick, F. Huettmann, and M. A. Lindgren, "A global model of avian influenza prediction in wild birds: the importance of northern regions," *Veterinary research*, vol. 44, no. 1, p. 42, 2013.

[243] X. Chen, S. Liu, M. U. Goraya, M. Maarouf, S. Huang, and J.-L. Chen, "Host immune response to influenza a virus infection," *Frontiers in immunology*, vol. 9, p. 320, 2018.

[244] H. Sooryanarain and S. Elankumaran, "Environmental role in influenza virus outbreaks," *Annu. Rev. Anim. Biosci.*, vol. 3, no. 1, pp. 347–373, 2015.

[245] G. Chowell and H. Nishiura, "Transmission dynamics and control of ebola virus disease (evd): a review," *BMC medicine*, vol. 12, no. 1, p. 196, 2014.

[246] B. Rockx and L.-F. Wang, "Zoonotic henipavirus transmission," *Journal of Clinical Virology*, vol. 58, no. 2, pp. 354–356, 2013.

[247] B. E. Martina, P. Koraka, and A. D. Osterhaus, "Dengue virus pathogenesis: an integrated view," *Clinical microbiology reviews*, vol. 22, no. 4, pp. 564–581, 2009.

[248] Y.-R. Guo, Q.-D. Cao, Z.-S. Hong, Y.-Y. Tan, S.-D. Chen, H.-J. Jin, K.-S. Tan, D.-Y. Wang, and Y. Yan, "The origin, transmission and clinical therapies on coronavirus disease 2019 (covid-19) outbreak–an update on the status," *Military Medical Research*, vol. 7, no. 1, pp. 1–10, 2020.

[249] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *The Lancet infectious diseases*, 2020.

[250] N. C. Peeri, N. Shrestha, M. S. Rahman, R. Zaki, Z. Tan, S. Bibi, M. Baghbanzadeh, N. Aghamohammadi, W. Zhang, and U. Haque, "The sars, mers and novel coronavirus (covid-19) epidemics, the newest and biggest global health threats: what lessons have we learned?" *International journal of epidemiology*, 2020.

[251] Z. Li, Y. Yi, X. Luo, N. Xiong, Y. Liu, S. Li, R. Sun, Y. Wang, B. Hu, W. Chen *et al.*, "Development and clinical application of a rapid igm-igg combined antibody test for sars-cov-2 infection diagnosis," *Journal of medical virology*, 2020.

[252] Y. Cao, L. Li, Z. Feng, S. Wan, P. Huang, X. Sun, F. Wen, X. Huang, G. Ning, and W. Wang, "Comparative genetic analysis of the novel coronavirus (2019-ncov/sars-cov-2) receptor ace2 in different populations," *Cell Discovery*, vol. 6, no. 1, pp. 1–4, 2020.