NANYANG TECHNOLOGICAL UNIVERSITY

COMPUTATIONAL MODELLING AND DATA ANALYSIS ON THE VIRULENCE OF INFLUENZA VIRUSES

ZHOU XINRUI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

2019

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.

24 March 2020	
Date	Zhou Xinrui

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.

24 March 2020

Date

Kwoh Chee Keong

Authorship Attribution Statement

This thesis contains material from seven papers published in the following peer-reviewed journals and papers accepted at conferences in which I am listed as an author.

Chapter 3 is published as:

(a) Fransiskus Xaverius Ivan, Xinrui Zhou, Akhila Deshpande, Rui Yin, Jie Zheng, and Chee Keong 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*, pages 21–26. ACM, 2017

- FXI conceived the idea, directed the project. XZ and AD conducted the experiments, interpreted the results, and edited the manuscript. RY and JZ vetted through the manuscript. CK provided overall supervision to the research.

(b) Haifen Chen, Xinrui Zhou, Jie Zheng, and Chee-Keong Kwoh. Rules of co-occurring mutations characterize the antigenic evolution of human influenza A/H3N2, A/H1N1 and B viruses. *BMC Medical Genomics*, 9(3):69, 2016

- HC conceived and directed the project. HC and XZ performed experiments, interpreted results, and wrote the manuscript. JZ and CK revised the paper, provided overall supervision, direction and leadership to the research.

Chapter 4 is published as:

(a) @2018 IEEE. Reprinted, with permission, from Xinrui Zhou, Rui Yin, Jie Zheng, and Chee-Keong Kwoh. An encoding scheme capturing generic priors and properties of amino acids improves protein classification. *IEEE Access*, 2018

- XZ conceived the project, performed experiments, and wrote the manuscript. RY and JZ helped interpret the results and revise the manuscript. CK provided overall leadership to the research and vetted through the manuscript.

(b) Xinrui Zhou, Rui Yin, Chee-Keong Kwoh, and Jie Zheng. A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza a viruses. *BMC Genomics*, 19(10):936, 2018

- XZ conceived the project, performed experiments, and wrote the manuscript. RY helped interpret the results. CK and JZ revised the manuscript and provided supervision during the overall research.

Chapter 5 is published as:

(a) Xinrui Zhou, Jie Zheng, Fransiskus Xaverius Ivan, Rui Yin, Shoba Ranganathan, Vincent TK Chow, and Chee-Keong Kwoh. Computational analysis of the receptor binding specificity of novel influenza A/H7N9 viruses. *BMC genomics*, 19(2):88, 2018

- XZ conceived the project, performed experiments, and wrote the manuscript. JZ revised the manuscript and helped interpret the results. FXI, RY, SR vetted through the manuscript. CK provided supervision during the overall research.

Chapter 6 is published as:

(a) Fransiskus Xaverius Ivan[†], Xinrui Zhou[†], Lau Suk Hiang[†], Shamima Banu Binte SM Rashid, Douglas Tay, Jasmine Shi Min Teo, Hong Kai Lee, Mark I-Cheng Chen, Chee Keong Kwoh[*], Vincent Tak Kwong Chow[*]. Molecular insights into evolution, mutations and receptor-binding specificity of influenza A and B viruses from outpatients and inpatients in Singapore. *International Journal of Infectious Disease*, 90 (2020): 84-96 .

- [†] Authors with equal contribution;
- [*] Joint corresponding authors;
- CK and VC conceived and supervised the overall project. FXI, XZ and LSH equally contributed to the work. LSH, DT, JSMT, HKL and MIC cultured influenza viral samples and conducted genome sequencing. FXI, XZ and SR conducted computational analyses on the influenza viral sequences. All the authors vetted through the manuscript.

24 March 2020

Date

Zhou Xinrui

Abstract

Influenza virus, a rapidly evolving contagious virus causing seasonal flu, has been circulating globally for centuries. The first recorded influenza pandemic was in 1918, claiming at least 50 million lives. Until 1933, influenza viruses were isolated from human for the first time, proving that influenza is caused by a virus rather than a bacterium. Seasonal influenza has caused substantial social and economic burden worldwide, affecting school attendance, work absenteeism, industrial productivity, etc. It mainly infects the respiratory system and can cause pneumonia, severe complications, or deaths, especially among people at high risks. Flu vaccines have been designed as primary prevention to help defense viral infection in advance. However, the rapid and continuous evolution of influenza raises the challenge to prepare vaccine candidates matching the antigenicity of dominant circulating influenza viruses for the next season. Therefore, it is in urgent demand to characterize and predict the antigenicity of influenza in advance.

Given the feasibility of high throughput sequencing techniques and enriched protein structure database, lots of computational models have been proposed to characterize antigenic properties of influenza. There have been many studies working on evolutionary models for tracing back the genomic variations and predicting the antigenic variants of influenza. However, current models for predicting the antigenicity are only applicable to one pre-defined subtype of influenza virus. The universal models for multiple-subtypes are still lacking. When it comes to virulence, the ability of the virus to cause disease among humans, it is a more complex problem involving the interaction with the immune system. From the medical perspective, the virulence level of influenza viruses is measured with the severity of an infection, the capability of drug resistance and transmission among hosts. There are still no consistent measurements for quantifying the virulence level of an influenza viral strain. The objective of this dissertation is to construct computational models for profiling the virulence of influenza viruses.

In this dissertation, the virulence level is quantified from the virus perspective only, including the sequence analyses on the genomic variation, and structural analyses on the receptor binding. The proposed sequence models for genomic variation include a phylogenetic-tree based method for pairwise co-mutations of influenza intra-proteins, and a sequential rule mining based approach for co-occurring mutations at multiple sites, even on different proteins. For profiling the receptor binding specificity, a structure-based model was proposed to characterize the binding modes between the influenza viral membrane protein (HA) and the human receptors. Both sequence models and structural models are integrated into a pipeline to quickly profile the virulence of influenza viral strains. Results of this proposed pipeline on our newly

sampled influenza viral strains among outpatients and inpatients in Singapore highlighted viral subtypes and strains that are more infectious or pathogenic, which are consistent with the local observations.

Keywords: Influenza, computational models, antigenicity, receptor binding specificity, sequence-based analyses, structure-based analyses, molecular docking, molecular dynamics simulation, encoding scheme.

Acknowledgment

First and foremost, I want to express my appreciation to Prof Kwoh Chee-Keong and Zheng Jie, for their extensive supervising and supporting throughout my research. Also to Prof Vincent T.K. Chow and Lau Suk Hiang from NUS have provided kind help for sequencing the seasonal influenza viruses from outpatients and inpatients in Singapore. Prof Tan Meng How and Ke Yiping Kelly from the thesis assistant committee have given valuable advice in every meeting during my Ph.D. study. All the comments, suggestions, and support from them are deeply appreciated.

Also, I'd like to thank the Man of La Mancha, who taught me to dream the impossible dream, to fight the unbeatable foe, to run where the bravest dare not to go, and to reach the unreachable star. Thank Mr.Lin-Manuel Miranda, who created the incredible musical Hamilton. The lyrics, rhythms, and rhymes have been the best companion with me. Alexander Hamilton was always writing like running out of time and had never been satisfied. The story and the song "Non-Stop" have seriously inspired me to get to work. He wrote 51 essays, and I should be able to write a nice one. Hope I can watch a live performance of Hamilton in Broadway after passing my oral defense.

Last but not least, this dissertation is dedicated to my beloved, especially my grandma. I hope she would be satisfied and be proud of me in heaven. I'm grateful for all the care from my parents and friends who raised me up when I was down, who found and embraced me when I got lost. Without their support, I can never be who I am and go for what I want.

Author's Publications

1. Fransiskus Xaverius Ivan[†], Xinrui Zhou[†], Lau Suk Hiang[†], Shamima Banu Binte SM Rashid, Douglas Tay, Jasmine Shi Min Teo, Hong Kai Lee, Mark I-Cheng Chen, Chee Keong Kwoh[*], Vincent Tak Kwong Chow[*]. Molecular insights into evolution, mutations and receptor-binding specificity of influenza A and B viruses from outpatients and inpatients in Singapore. *International Journal of Infectious Disease*, 90 (2020): 84-96

2. Yin, Rui, Xinrui Zhou, Shamima Rashid, and Chee Keong Kwoh. "HopPER: an adaptive model for probability estimation of influenza reassortment through host prediction." *BMC Medical Genomics* 13, no. 1 (2020): 9.

3. Xinrui Zhou, Rui Yin, Jie Zheng, and Chee-Keong Kwoh. An encoding scheme capturing generic priors and properties of amino acids improves protein classification. *IEEE Access*, 2018

4. Xinrui Zhou, Rui Yin, Chee-Keong Kwoh, and Jie Zheng. A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza a viruses. *BMC Genomics*, 19(10):936, 2018

5. Xinrui Zhou, Jie Zheng, Fransiskus Xaverius Ivan, Rui Yin, Shoba Ranganathan, Vincent TK Chow, and Chee-Keong Kwoh. Computational analysis of the receptor binding specificity of novel influenza A/H7N9 viruses. *BMC genomics*, 19(2):88, 2018

6. Rui Yin, Junqiu Tan, Deshpande Akhila, Xinrui Zhou, and Chee Keong 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*, page 5. ACM, 2018

7. Rui Yin, Xinrui Zhou, Jie Zheng, and Chee Keong Kwoh. Computational identification of physicochemical signatures for host tropism of influenza a virus. *Journal of Bioinformatics and Computational Biology*, pages 1840023–1840023, 2018

8. Rui Yin, Viet Hung Tran, Xinrui Zhou, Jie Zheng, and Chee-Keong Kwoh. Predicting antigenic variants of H1N1 influenza virus based on epidemics and pandemics using a stacking model. *PLoS ONE*, 2018

9. Fransiskus Xaverius Ivan, Xinrui Zhou, Akhila Deshpande, Rui Yin, Jie Zheng, and Chee Keong 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*, pages 21–26. ACM, 2017

10. Rui Yin, Xinrui Zhou, Fransiskus Xaverius Ivan, Jie Zheng, Vincent TK Chow, and Chee Keong 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*, pages 30–36. ACM, 2017

11. Haifen Chen, Xinrui Zhou, Jie Zheng, and Chee-Keong Kwoh. Rules of co-occurring mutations characterize the antigenic evolution of human influenza A/H3N2, A/H1N1 and B viruses. *BMC Medical Genomics*, 9(3):69, 2016

Contents

Abstract v

Acknowledgement vii

Author's Publications viii

Lists of Figures xv

Lists of Tables xvii

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 The biology of influenza . . . 2
  - 1.2.1 Viral structure and genomic composition . . . 2
  - 1.2.2 The replication of influenza viruses . . . 5
- 1.3 Epidemics and pandemics of influenza in history . . . 6
- 1.4 Vaccination and drug therapies for influenza . . . 9
- 1.5 Organization of the dissertation . . . 11

2 Literature review 14
- 2.1 Role of influenza viral genes in virulence . . . 15
  - 2.1.1 Glycoproteins . . . 15
  - 2.1.2 Influenza viral polymerase . . . 20
  - 2.1.3 Non-structural proteins . . . 22
- 2.2 The antigenicity of influenza viruses . . . 25
- 2.3 Computational modeling on influenza . . . 29
  - 2.3.1 Sequence-based computational analyses on influenza . . . 29
  - 2.3.2 Structure-based computational analyses on influenza . . . 35
- 2.4 Summary . . . 37

3 Detecting co-occurring mutations and virulence signatures of influenza viruses 39
- 3.1 Known co-mutations in influenza viruses . . . 39
- 3.2 A phylogenetic tree-based method for detecting pairwise co-mutations . . . 41

3.2.1 Methods . . . 42
3.2.2 Results . . . 43
3.3 Association rules and sequential rules mining of mutations at multiple sites . . . 45
3.3.1 Methods . . . 45
3.3.2 Results . . . 49
3.4 Summary . . . 58

4 Predicting the antigenicity of influenza viruses from the HA sequences 60
4.1 Methods for predicting the antigenicity of influenza viruses . . . 60
4.1.1 Genotypic Analyses . . . 61
4.1.2 Phenotypic analyses . . . 62
4.1.3 Hybrid models integrating genotypic and phenotypic analyses . . . 63
4.2 CFreeEnS: a **C**ontext-**Free** **E**ncoding **S**cheme of protein sequences . . . 63
4.2.1 A typical pipeline of computational modeling . . . 63
4.2.2 CFreeEnS for protein sequences and protein sequence pairs . . . 64
4.2.3 Model evaluation . . . 67
4.3 Predicting the antigenicity of diverse influenza A viruses . . . 68
4.3.1 Data . . . 68
4.3.2 Results . . . 69
4.4 Summary . . . 75

5 Structure-based analysis to quantify viral binding preference with host cells 77
5.1 Structural basis of viral binding with host cells . . . 77
5.2 Methods to identify genetic markers for cross-species transmission or virulence determinants . . . 79
5.2.1 Animal models . . . 79
5.2.2 Computational models . . . 81
5.3 The receptor binding specificity of a novel influenza A/H7N9 virus . . . 83
5.3.1 Functional markers of the novel influenza A/H7N9 strain . . . 84
5.3.2 Molecular docking predicting the optimal conformations with host receptor analogs . . . 86
5.3.3 Molecular dynamics simulation revealing residues that contributing the the enhanced binding with host cell receptors . . . 88
5.4 Summary . . . 93

6 Profiling the evolution, mutations and receptor binding specificity of viral strains from outpatients and inpatients in Singapore 96
6.1 Introduction . . . 96
6.1.1 The burden of influenza infections in Singapore . . . 96
6.1.2 The necessity for regionally focused influenza study . . . 97
6.2 Influenza viral samples and genome sequencing . . . 98

6.2.1 RNA extraction and reverse transcription . . . 98
6.2.2 NGS and Sanger for genome sequencing . . . 99
6.3 Molecular evolution study of sampled sequences . . . 100
6.3.1 BLAST analysis . . . 100
6.3.2 Phylogenetic tree analyses . . . 101
6.3.3 Phenotypically or epidemiologically interesting mutations and genomic signatures discriminating outpatient and inpatient samples . . . 102
6.4 Receptor binding specificity of sampled viruses . . . 105
6.4.1 Molecular docking analyses of HA-receptor binding . . . 105
6.4.2 A/H1N1 receptor binding modes . . . 108
6.5 Summary . . . 113

7 Discussion and conclusion **114**
7.1 General discussion . . . 114
7.2 Future directions . . . 115

Appendices **142**

Appendix A CFreeEnS improves protein classification **142**
A.1 Datasets for protein classification . . . 142
A.1.1 iAMP . . . 143
A.1.2 TumorHPD . . . 143
A.1.3 HemoPI . . . 143
A.1.4 PVPred . . . 144
A.2 Results of protein classification . . . 144

Appendix B Supplements for MD simulation **146**
B.1 Total binding energy . . . 146
B.2 Average binding energy . . . 146

List of Figures

1.1 A typical schematic structure of an influenza A virus particle. (Horimoto and Kawaoka, 2005) . . . 3
1.2 The lifecycle of influenza viruses. . . . 6
1.3 The spread of major influenza epidemics and pandemics in recent centuries. . . . 8
1.4 The reassortment events resulting in the 2009 influenza A/H1N1 virus. . . . 9
1.5 Illustration of the scope and organization of this dissertation. . . . 12

2.1 A schematic of an unfolded polypeptide chain of HA0 of H1 virus . . . 16
2.2 The folded HA protein structure and functional domains. . . . 17
2.3 HA structure of H1 and H3 with receptor-binding sites and antigenic sites highlighted. . . . 18
2.4 NA structure of N2 with active sites highlighted. . . . 20
2.5 The cartoon representation of influenza A virus polymerase in different orientations (PDB ID: 4WSB). . . . 21
2.6 Structures of the RNA-binding domain and the effector domain of NS1. . . . 23
2.7 An example of hemagglutination assay (Eisfeld *et al.*, 2014). . . . 26
2.8 Mechanism of the hemagglutination inhibition assay. . . . 27
2.9 An example of hemagglutination inhibition assay. . . . 27
2.10 A typical phylogenetic tree for HA protein of influenza A/H3N2. . . . 30
2.11 An example of antigenic map for human influenza A/H3N2 viral isolates from 1968 to 2003 (Smith *et al.*, 2004). . . . 32
2.12 Antigenic evolution of influenza A/H1N1pdm. . . . 33
2.13 Illustration of molecular docking (Salmaso and Moro, 2018). . . . 37

3.1 An overall architecture for inferring co-occurring mutations from reconstructed phylogenetic trees. . . . 41
3.2 An Example of the phylogenetic tree of the HA sequences of influenza A/H3N2. 43
3.3 An overall architecture for uncovering the co-occurring mutation patterns from influenza protein sequences. . . . 46
3.4 Problem formulation: from protein sequences to transactions (*Seq2Trans*). . . . 47
3.5 The phylogenetic tree and co-occurring mutations of influenza A/H1N1. . . . 49
3.6 The phylogenetic tree and co-occurring mutations of influenza A/H3N2. . . . 50
3.7 The phylogenetic tree and co-occurring mutations of influenza B viruses. . . . 51

3.8 Mapping co-mutation sites detected onto HA1 protein of influenza B/Yamagata and B/Victoria (PDB ID: 4NRJ) (Ni *et al.*, 2014). 52
3.9 Co-occurring mutations on the PB2 of influenza A/pH1N1 viruses. 54
3.10 Co-occurring mutation patterns on the HA protein of influenza A/epH1N1, A/H3N2 and B viruses. 57
3.11 Comparison of co-occurring mutation patterns on the HA protein of influenza B/Victoria and B/Yamagata lineage. 57

4.1 A typical pipeline of supervised machine learning models in bioinformatics. . . 64
4.2 The diagram of CFreeEnS for *m* protein sequences or protein sequence pairs. . 66
4.3 Evaluation of all substitution matrices on datasets of single subtype. 70
4.4 Comparing F-score of models on datasets with singe subtype influenza viruse. . 71
4.5 Accuracy scores of transfer learning using three encoding schemes: MutCounts, RegionBand and CFreeEnS. 75
4.6 Learning curve of the random forest regressor trained on different datasets. . . 76

5.1 An illustration of an influenza viral particle binding with the host cells. 78
5.2 Structure of sialic acids recognized by the HA protein of influenza viruses. . . . 79
5.3 The workflow for structural analyses on HA-receptor bindings. 83
5.4 The HA and NA phylogenetic trees of influenza A/H7N9. 84
5.5 Binding affinity of host receptor analogs with the H7N9 HA proteins. 87
5.6 Superimpose the best and worst conformations of each HA-receptor binding predicted from molecular docking. 89
5.7 Typical steps for conducting molecular dynamics simulation. 89
5.8 Analysis of RMSD and total vacuum MM energy during the whole MD simulation. 90
5.9 Visualization of HA-receptor interactions in the optimally docked complexes of SH13-LSTa, SH13-LSTc, TW17-LSTa and TW17-LSTc. 92
5.10 Average energy contribution of residues that involved in receptor-ligand interactions in the optimally docked complexes. 92
5.11 Comparing residues contribution to HA-receptor binding energy. 94

6.1 Subtypes of collected samples. 100
6.2 The HA and NA phylogenetic trees of influenza A/H1N1, A/H3N2 and B viruses. 102
6.3 The power of sites in discriminating inpatient and outpatient strains. 106
6.4 Binding affinity of host cell receptors with the HA1 of the representative influenza A/H1N1, A/H3N2 and B viruses. 107
6.5 H1 HA with avian receptor analogs. 111
6.6 H1 HA with human receptor analogs. 112

A.1 Predicting accuracy of CFreeEnS on protein classification compared with traditional methods and *m-NGSG* 145

B.1 Comparison of total binding energy between two rounds 50 ns MD simulation for SH13-LSTa, SH13-LSTc, TW17-LSTa and TW17-LSTc. 147

List of Tables

1.1 The genome set of influenza A viruses and functions of encoded proteins. . . . 4
1.2 The mortality and virus composition of major influenza pandemics in recent centuries. . . . 7

2.1 Elements for the influenza risk assessment tool. . . . 14
2.2 A summary of antigenic sites and the primary receptor binding sites of H1, H3 and H7. . . . 18
2.3 Reported virulence determinants on NS1. . . . 24
2.4 A typical 2×2 HI table with four titers for viral strains V_i and V_j. . . . 28
2.5 Summary of nucleotide substitution models. . . . 31

3.1 Co-mutations of influenza viral proteins reported in literature. . . . 40
3.2 Dominant mutations detected from the longest path of the phylogenetic trees. . . 44
3.3 Top 10 confident site-pairs with varying distance thresholds d^*. . . . 44
3.4 An overview of extracted rules on the HA1 protein sequences of influenza A/H1N1, A/H3N2 and B viruses using association rule mining. . . . 51
3.5 An overview of the complete protein sequences used for analyses. . . . 53
3.6 Co-occurring mutations detected in each protein of influenza A/pH1N1, A/epH1N1, A/H3N2 and B viruses. . . . 53
3.7 Comparing the detected mutations on HA protein of influenza A/H3N2 with pure association rule mining method (Chen *et al.*, 2016) and a phylogenetic tree-based method (Ivan *et al.*, 2017) . . . 56
3.8 Frequent co-occurring mutations across the HA and NA protein of influenza A/pH1N1 and A/H3N2 . . . 58

4.1 Datasets for training and testing the predicting model. . . . 69
4.2 Performance comparison among five strategies on four single subtype datasets. 72
4.3 Performance comparison among five strategies on the combined dataset. . . . 73
4.4 Evaluation metrics for the performances of three encoding schemes on transfer learning. . . . 74

5.1 Advantages and disadvantages of different animal models. . . . 80
5.2 Mutations of the influenza TW17 strain compared with the reference SH13 strain. 86
5.3 T-test for HA-receptor docking experiments (N= 500) . . . 88

5.4 Average total binding energy (kJ/mol) of the HA-LSTa/LSTc complexes. 91
5.5 Interacting residues and the number of hydrogen bonds between host receptor analogs and the HA proteins of influenza A/H7N9. 91

6.1 Infections and deaths in Singapore during the past influenza pandemics. 97
6.2 An overview of the collected influenza positive samples from outpatient and inpatient subjects during 2012-2015. 99
6.3 The WHO recommended vaccine strains from 2010 to 2020. 101
6.4 Mutations of the positive influenza samples with high interest level. 104
6.5 Representative strains and the predicted optimal matching template for the HA proteins 280 of influenza A/H1N1, A/H3N2 and B viruses. 106
6.6 The group mean differences of the HA1 proteins of the representative influenza A/H1N1, A/H3N2 and B binding with 3'SLN and 6'SLN. 108
6.7 Solvent Accessible Surface Area of HA-receptor complexes. 109
6.8 Number of putative hydrogen bonds between the HA protein and host cell receptors. 110

A.1 An overview of datasets for protein classification 143
A.2 The classification results of CFreeEnS applied to iAMP, TumorHPD, HemoPI and PVPred datasets. 145

B.1 Average total binding energy (kJ/mol) of the HA-LSTa/LSTc complexes. 147

Chapter 1

Introduction

1.1 Background

Influenza virus, which is widely known as seasonal flu, has been identified as a highly contagious disease circulating globally for centuries since 1933 (Smith *et al.*, 1933). Seasonal influenza has caused substantial social and economic burden worldwide, affecting school attendance, work absenteeism, industrial productivity, etc. It mainly infects the respiratory system and can cause pneumonia, severe complications, or deaths, especially among people at high risks (Thompson *et al.*, 2009; Nair *et al.*, 2011). As documented by the World Health Organization (WHO), seasonal influenza is estimated to cause 3-5 million server infections and claim up to 650,000 lives (World Health Organization (WHO), 2018b). Apart from the seasonal flu, influenza viruses have caused four major pandemics in history, claiming thousands even millions of lives (Girard *et al.*, 2010; Khanna *et al.*, 2008; Stein, 2009).

Flu vaccines have been designed as primary prevention to help defense viral infection in advance since the 1940s (Plotkin, 2014). However, influenza viruses are undergoing continuous evolution and possess the ability of genetic reassortment, which empowers the virus to escape the detection and defense of the human immune system (Bragstad *et al.*, 2008; Joseph *et al.*, 2017). The emerging novel influenza, which causes infections among humans and animals, has always been a great concern to cause pandemics when they evolve to gain enhanced ability to spread among human. The rapid evolution of influenza necessitates timely updates of flu vaccines (Gerdil, 2003).

Since 1952, the WHO has been monitoring the evolution and circulation of influenza viruses among humans through the Global Influenza Surveillance and Response System (GISRS). The surveillance is conducted under the help of health authorities and collaborations from worldwide, which has covered the influenza infections among animals more recently (World Health Organization (WHO), 2018a). The GISRS is primarily responsible for the early detection and antigenic characterization of influenza viruses, functioning as a global mechanism of surveillance, preparedness, and response for zoonotic, seasonal and pandemic influenza.

The WHO assesses and recommends the components of human influenza vaccines twice a

year before the coming winters of the Northern and Southern hemisphere. Current systems of influenza vaccination by WHO relies on the antigenicity of clinical isolates and cross-reactive immunity in human populations, combined with genetic and epidemiological data (World Health Organization *et al.*, 2018a). Such experiments are time-consuming and costly. Besides, to spare sufficient time for the manufacturing of vaccines, the authorities need to select effective vaccine strains more than six months before the onset of influenza seasons (Gerdil, 2003). Thus, the health authority needs to identify the circulating viral strains and predict the dominant strains in advance. The effectiveness of a vaccine is heavily dependent on the match of antigenicity between the circulating viral strains and those used for vaccination. However, influenza viruses are undergoing rapid mutation rate. Cumulative mutations may lead to new antigenic clusters with distinct antigenicity from current circulating strains. The new antigenic clusters may have not been reported or even emerged at the time of decision making for influenza vaccines, leading to low vaccine effectiveness. Therefore, it is also crucial to predict the mutations and antigenic variants, as well as quantify the risk of new influenza viral strains for a timely response.

Given the feasibility of high throughput sequencing techniques and enriched protein structure database, computational models for characterizing antigenic properties of influenza have been developed (Ghedin *et al.*, 2005; Berman *et al.*, 2000). There have been many computational models for tracing back the genomic variations and predicting the antigenic variants of influenza (Bragstad *et al.*, 2008; Arinaminpathy and Grenfell, 2011; Chen *et al.*, 2012). However, current models for predicting the antigenicity of influenza are only applicable to a few subtypes. The universal models for multiple-subtypes are still lacking. When it comes to virulence, the ability of the virus to cause disease among humans, it is a more complex problem involving the interaction with the immune system (Morens *et al.*, 2004). There are still no consistent measurements for quantifying the virulence level of an influenza viral strain. In this dissertation, the virulence level is quantified from the virus perspective only, including the sequence analyses on the genomic variation, and structural analyses on the receptor binding.

1.2 The biology of influenza

1.2.1 Viral structure and genomic composition

The influenza viruses, belonging to the family Orthomyxoviridae, can be categorized into three types according to their immunological groups, designated as type A, B, C, and D. The influenza A viruses have been isolated from several mammalian species and more than 105 species of the avian (Wahlgren, 2011). The type A influenza virus has the highest mutation rate among all types of influenza, which is the most prevalent and notorious type discovered in human and animals. Humans are the primary reservoir for type B and C influenza viruses. Only sporadic influenza B viruses have been isolated from seals (Osterhaus *et al.*, 2000; Bodewes *et al.*, 2013) and influenza C viruses from pigs, respectively (Yuanji *et al.*, 1983; Kimura *et al.*,

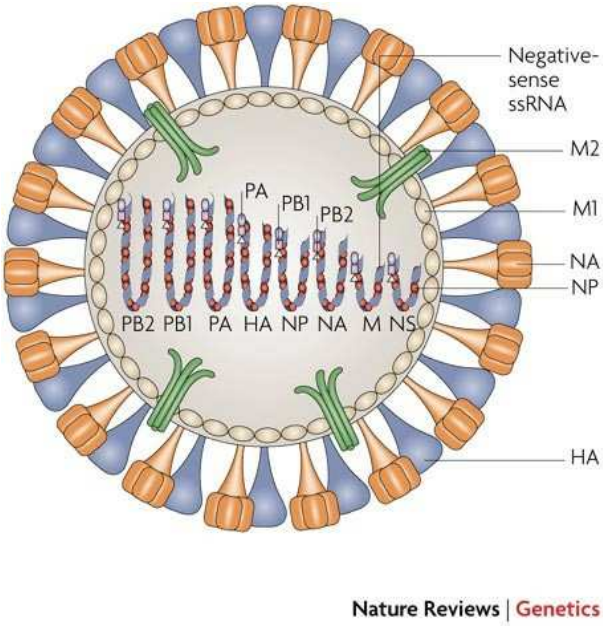


Figure 1.1: A typical schematic structure of an influenza A virus particle. (Horimoto and Kawaoka, 2005)

1997). Influenza D virus has been isolated from swine, cattle, and horses since 2011 (Ferguson *et al.*, 2016; Su *et al.*, 2017). The full range of hosts and the zoonotic risk of the novel influenza D virus are to be determined but currently not reported to cause sickness in humans (Nedland *et al.*, 2018; World Health Organization (WHO), 2018b). Seasonal epidemics are mainly caused by the type A and B of influenza, while the type C virus is less prevalent, which usually causes only upper respiratory infection in children with mild symptoms. Therefore, the influenza A and B viruses will be the focus throughout.

The four types of influenza viruses share similar viral structure and composition. Figure 1.1 shows a typical viral particle of influenza A virus. The influenza viral particle is roughly spherical about 80-120 nanometers in diameter (Horimoto and Kawaoka, 2005). The viral particle consists of a lipid bilayer envelope and a wrapped central core. The envelope contains two types of glycoproteins, known as hemagglutinin (HA) and neuraminidase (NA), respectively. The central core contains negative-sense single-stranded RNAs, which encode several viral proteins. Type A and B influenza have eight single-strand RNA segments, with two glycoproteins on the viral surface. However, the type C and D viruses have only seven segments, encoding one glycoprotein named hemagglutinin-esterase fusion (HEF) on the viral surface. The segmented structure of genomes is responsible for the rapidly generating diversity, which allows and facilitates the reassortment of influenza viruses when co-infecting a single host (Vijaykrishna *et al.*, 2010).

The total length of eight RNA segments of the influenza A virus is around 13 kb. Each segment encodes a major viral protein, including PB2, PB1, PA, HA, NP, NA, M2, and NS1. Later, the M2 and NS2 (NEP), two splicing variants of the 7th and the 8th segment were detected (Bouvier and Palese, 2008). Recently, more novel viral proteins expressed by splicing

Table 1.1: The genome set of influenza A viruses and functions of encoded proteins. *Segments are listed from the longest to the shortest by convention. The ten main proteins are colored with gray. (Bouvier and Palese, 2008; Hayashi* et al.*, 2015; Dubois* et al.*, 2014; Yamayoshi* et al.*, 2016; Hu* et al.*, 2015)*

Segment	Protein	Protein functions
1 PB2	PB2	Polymerase subunit; mRNA cap recognition; interacting with MAVS; regulating the host antiviral innate immune pathways
	PB2-S1	Detected in virus-infected cells and in cells transfected with a protein expression plasmid encoding PB2
2 PB1	PB1	Polymerase subunit, RNA elongation, endonuclease activity
	PB1-F2	Pro-apoptotic activity; responsible for pathogenicity; associated with pathogenicity, inducing apoptosis and a reduction in the mitochondrial inner membrane potential; aggravation of inflammation; secondary bacterial infection
	PB1-N40	N-terminal truncated polypeptide; interacting with the polymerase; viral replication
3 PA	PA	Polymerase subunit; protease activity
	PA-N155	N-terminally truncated form of PA; associated with viral replication and pathogenecity
	PA-N182	N-terminally truncated form of PA; associated with viral replication and pathogenecity
	PA-X	Viral Growth and suppression of the host antiviral and immune responses
4 HA	HA	Membrane glycoprotein; major antigen; receptor binding and fusion activities
5 NP	NP	RNA binding protein; nuclear import regulation; viral replication
6 NA	NA	Membrane glycoprotein; sialidase activity; viral release
7 M	M1	Matrix protein; vRNP interaction; RNA nuclear export regulation; viral budding
	M2	Proton ion channel; virus uncoating and assembly
	M42	An alternative proton channel in place of M2
8 NS	NS1	Interferon antagonist protein; regulation of host gene expression, host inflammatory responses; viral replication and pathogenicity
	NS2/NEP	Nuclear export of RNA; bound to M1 protein
	NS3	Associated with the adaptation of avian influenza A virus to new mammalian hosts

(e.g. PB2-S1, M42, and NS3 (Yamayoshi *et al.*, 2016; Wise *et al.*, 2012; Selman *et al.*, 2012)), alternative open reading frame (e.g. PB1-F2, PB1-N40, PA-N155, and PA-N182 (Chen *et al.*, 2001; Wise *et al.*, 2009; Muramoto *et al.*, 2013)), or ribosomal frameshifts (e.g. PA-X (Jagger *et al.*, 2012)) have been identified. Table 1.1 lists the genome segments (from the longest to the shortest by convention), viral proteins, and their functions.

The HA, NA, and the M2 ion channel are the three major components in the viral lipid envelope. The HA is involved in the process of viral binding and transporting viral genomes into the target cell, while NA, is responsible for cleaving the mature viral particles from the infected cells to release progeny viral particles (Gamblin and Skehel, 2010). In a typical viral particle, the HA is four to five times more than the NA. (Subbarao and Joseph, 2007). The excess of HA over NA indicates a weak interaction of HA binding with host cell receptors, in need of several connections for a stable interaction (Sauter *et al.*, 1989). M42 functions as an alternative proton channel in place of M2. The M1 is a matrix protein, one of the main components of influenza viral capsids and the most abundant structural protein beneath the influenza viral envelope. All the RNA segments reside in the central core beneath the M1 layer in the form of ribonucleoprotein (RNP) complexes.

There are three segments (PB2, PB1, and PA) encoding nine known viral polymerase complexes. For instance, the polymerase basic 2 (PB2), polymerase basic 1 (PB1), and polymerase acid (PA) form a trimeric viral RNA-dependent RNA polymerase. This polymerase functions in the transcription and replication process of the viral RNA (Engelhardt and Fodor, 2006). The PB2 protein is reported to be important in regulating the host antiviral innate immune pathways

(Graef *et al.*, 2010). The functions of the newly detected PB2-S1 are unknown yet. The PB1 regulates the endonuclease activity and the RNA elongation. PB1-F2 has multiple functions, including regulating the pro-apoptotic activity (Le Goffic *et al.*, 2011), inducing apoptosis (Zamarin *et al.*, 2005) and secondary bacterial infection (McAuley *et al.*, 2007a). PB1-N40 lacks the transcriptase function compared to PB1 but can interact with polymerase and regulate the viral replication (Wise *et al.*, 2009). PA is primarily responsible for the promoter binding, cap binding, endonuclease activity, and protease activity (Yuan *et al.*, 2009; Wang *et al.*, 2018). PA-N155 and PA-N182 are probably associated with viral replication and pathogenicity (Wise *et al.*, 2009; Muramoto *et al.*, 2013). PA-X can regulate viral growth, suppress the host immune responses, and contribute to the virulence (Jagger *et al.*, 2012; Desmet *et al.*, 2013; Hu *et al.*, 2015). The nucleoprotein (NP) is an RNA binding protein which mainly functions in viral replication, responsible for regulating the nuclear import and export of RNA. The nonstructural proteins (NS1, NS2, and NS3) are expressed in the infected cells, which are essential in regulating host gene expression and host inflammatory responses (Bouvier and Palese, 2008). The newly identified NS3 could be associated with the host adaptation from avian to mammalian (Selman *et al.*, 2012).

The influenza viruses can be further categorized based on antibody responses. Determined by the two primary antigens HA and NA, the influenza A viruses are subtyped as H*N*. To date, there are 18 HA subtypes, and 11 NA subtypes have been reported, named as H1-H18 and N1-N11. The wild aquatic birds have been found to harbour most subtypes of influenza A viruses, except H17N10 and H18N11, which have been detected only in bats. (Tong *et al.*, 2013). Influenza B viruses are clustered into two co-circulating lineages since 2001, B/Victoria/2/1987-like and B/Yamagata/16/1988-like, referred to as B/Yamagata and B/Victoria lineage, with distinct antigenic HA glycoproteins (Biere *et al.*, 2010).

1.2.2 The replication of influenza viruses

The invasion and replication of influenza viruses include several stages: entering, membrane-uncoating, component-reproducing, particle-assembling, and progeny-releasing. A schematic description of the lifecycle of influenza A viruses is shown as Figure 1.2 (Samji, 2009; Wilschut *et al.*, 2005).

First, the HA protein, which is on the influenza membrane, binds to the sialic acid on glycoproteins or glycolipid receptors of the host cells. The specificity of HA-receptor binding is dependent on the nature of the glycosidic linkage between the terminal sialic acid and the penultimate galactose residue on the receptor. The virus particles are then endocytosed by the host cell plasma membrane, forming an endocytic vesicle. During the endocytosing process, the HA protein undergoes a low PH environment, forcing a conformational change exposing a hydrophobic fusion peptide. The low PH environment also triggers uncoating the viral and endosomal membranes and releasing viral RNPs into the cytoplasm.

Subsequently, the vRNPs are transporting to the host nucleus. Primary transcription will be

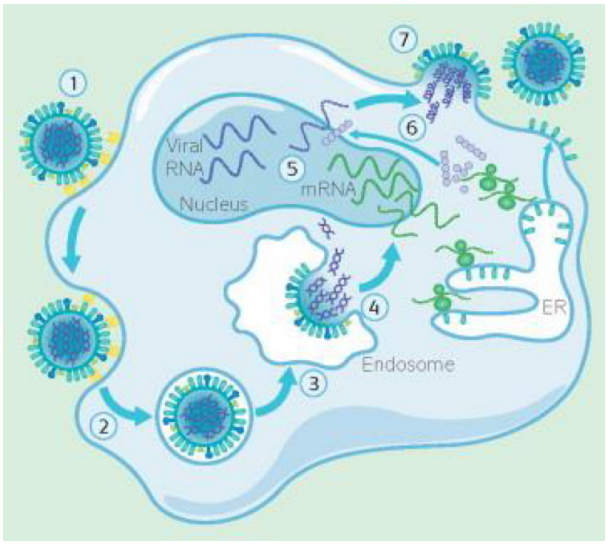


Figure 1.2: The lifecycle of influenza viruses. (Samji, 2009) *(1) Viral binding to host cell receptor. (2) Membrane fusion and viral endocytosing by the host cell plasma membrane. (3) Delivery of the virus to the endosomal cell compartment. (4) Uncoating viral and endosomal membrane uncoating and transporting vRNPs into the host nucleus. (5) Synthesis of mRNA, cRNA, and vRNPs. (6) Synthesis and post-process of viral membrane proteins. (7) Assembly and releasing of progeny viruses.*

done in the host nucleus to synthesize necessary proteins for replication, including the messenger RNA (mRNA) and complementary RNA (cRNA). In this process, the vRNPs are used as templates and the transcriptase (including PA, PB1, and PB2) carried by the vRNPs, are performing as catalyzers. These synthesized cRNAs are further utilized as templates for synthesizing negative-sense viral RNA segments and amplifying mRNA. NP, M1, NS2, and polymerases will also be imported into the nucleus for the final replication, and the assemble of vRNPs.

Then, the newly synthesized vRNPs will be exported into the cytoplasm of the host and then guided to the plasma membrane. Meanwhile, the HA and NA proteins are undergoing glycosylation, polymerization, and acylation in the cytoplasm. HA, NA and M1 will be directed to the plasma membrane as well. All components will be then assembled into progeny particles at the plasma membrane, and then the virus buds. The final stage is to cleave those progenies from host cells by the NA enzyme, thus allowing the release of viral progenies.

1.3 Epidemics and pandemics of influenza in history

Because of lacking pre-existing immunity, a novel influenza virus strain capable of spreading among humans is likely to cause a pandemic. During the pandemics, the viruses spread quickly worldwide in several waves. In recent centuries, the influenza A viruses have caused four pandemics, which have resulted from antigenic drift and shift (Girard *et al.*, 2010; Kilbourne, 2006). The origin, mortality and virus compositions of the four pandemics are summarized in Table 1.2, including the most devastating Spanish flu in 1918 caused by influenza A/H1N1 and the pandemics in 1957, 1968 and 2009 caused by influenza A/H2N2, A/H3N2, and A/H1N1, respectively. Figure 1.3 depicts the spread of major epidemics and pandemics in recent cen-

Table 1.2: The mortality and virus composition of major influenza pandemics in recent centuries.

Onset time	Pandemic	Virus composition	Mortality worldwide	Origin of virus genes
1918	Spanish flu	H1N1	50 million	Unclear origin, probably France; with mammalian and avian genetic information
1957	Asian flu	H2N2	2 million	Reassortant of human influenza A/H1N1 and avian influenza virus with surface antigens H2N2, acquiring the avian PB1
1968	Hong Kong flu	H3N2	1 million	Reassortant of human influenza A/H2N2 and avian influenza virus with surface antigen H3 acquiring the avian PB1
2009	Swine flu	H1N1	151,700 – 575,400	Reassortment of at least three parents (human, avian and swine), probably a 'quadruple' reassortant

turies (Krammer *et al.*, 2018).

The 1918 Spanish flu was a nightmare in history. There was no effective treatment or vaccine to influenza viruses at that time. It was the first and most devastating pandemic caused by influenza A/H1N1, estimated to have claimed 50 million lives throughout the world (Taubenberger and Morens, 2006). The origin countries and viral strains are not clear until now. However, the viruses in that pandemic had both mammalian and avian genetic elements. The virus responsible for the 1918 Spanish flu was reconstructed by reverse genetics, confirming its high morbidity and high mortality in humans (Tumpey *et al.*, 2005). It was claimed that a group of three genes were responsible for weakening a victim's bronchial tubes and lungs and clearing the way for bacterial pneumonia. Besides, the reconstructed strain can infect various species, such as mice, macaques, ferrets, and embryonated chicken eggs with a high fatality rate.

Since 1918, there have been another three pandemics. In 1957, the Asian flu caused by influenza A/H2N2 killed approximately two million people. The viral strain was a reassortment of A/H1N1 with the avian virus, acquiring avian H2 HA, N2 NA, and PB1 segments. In 1968, the Hong Kong flu by influenza A/H3N2 claimed around one million lives. The influenza A/H3N2 was reassorted from H2N2 and avian virus, acquiring avian H3 HA and PB1. In 1977, influenza A H1N1 was detected to be circulating among humans. However, no pandemic was raised, which may be partially due to the antigenic similarity between the circulating strain and that preceded in the 1957 influenza A/H2N2 pandemic, causing the pre-existing immunity among the population (Krammer *et al.*, 2018).

The most recent pandemic in 2009 originated from China. It was caused by influenza A/H1N1, which was a triple reassortant, or probably a quadruple reassortant, resulting in an antigenically very distinct strain to the preceded seasonal influenza A/H1N1 epidemic (Coburn *et al.*, 2009; Girard *et al.*, 2010). Figure 1.4 shows the reassortant events that created the influenza A/H1N1 strain causing the 2009 pandemic. The sixth and seventh segments NA and M (light green) are similar to the Eurasian swine strain, while the other segments are similar to the North American swine influenza A/H1N2 and the triple-reassortant swine influenza

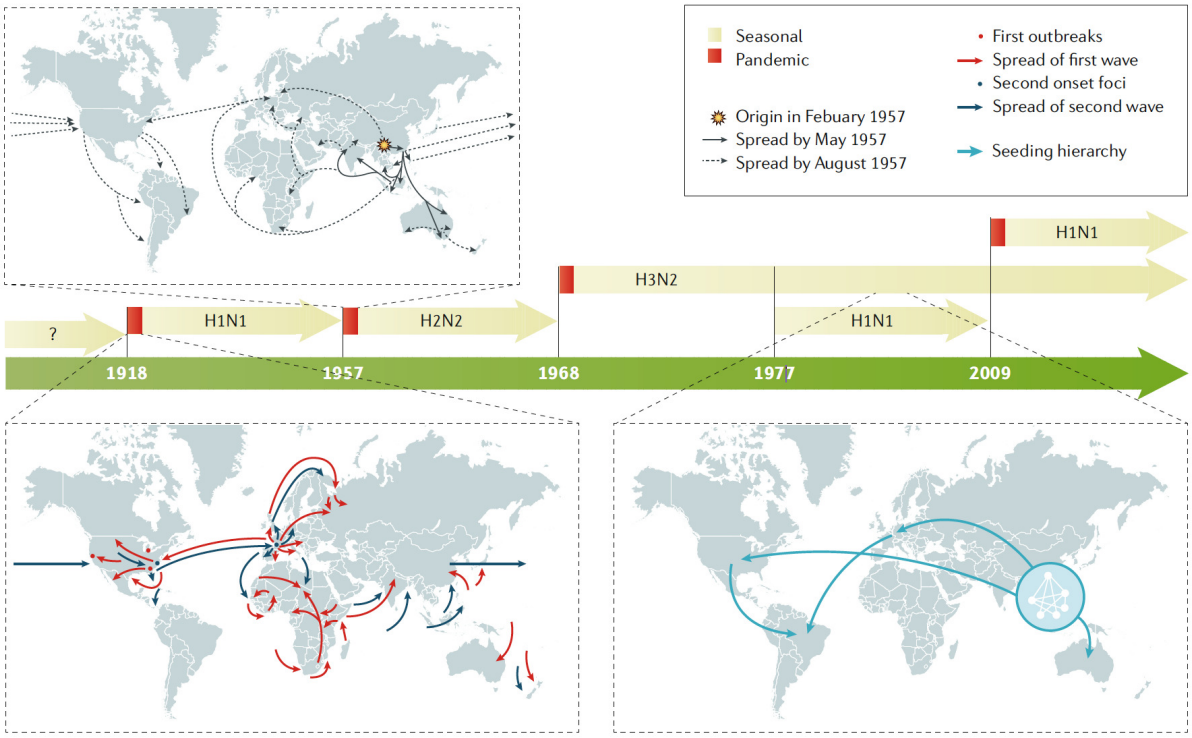


Figure 1.3: The spread of major influenza epidemics and pandemics in recent centuries. *The arrows for the pandemics represent the first and second waves of transmission, while the arrows for seasonal influenza A/H3N2 represent the seeding hierarchy over five years. Seasonal influenza B viruses that are co-circulating in humans are not shown. (Krammer* et al., *2018)*

A/H3N2 viruses (Trifonov *et al.*, 2009b). Dawood *et al.* estimated 151,700-575,400 deaths in the first year of the 2009 pandemic due to influenza infections (Dawood *et al.*, 2012). Currently, two lineages of B viruses (B/Yamagata and B/Victoria), the influenza A/H1N1 strain causing the 2009 pandemic, and A/H3N2 are co-circulating as seasonal influenza viral strains among human.

Some strains that are routinely circulating in animals can also infect humans, for instances, the avian influenza A/H5N1, A/H7N9, and A/H9N2 and the swine influenza A/H1N1 and A/H3N2. Those viral strains infecting humans are generally reassortment of human and avian viruses. Infections of human by a complete avian virus were first detected in 1997. The influenza A/H5N1, a complete avian strain, was found transmitted to the human, causing fatal disease. Then, the avian A/H7N9 and H9N2 were also reported as direct transmissible to human, without the lethal disease (Liu *et al.*, 2013). Generally, human infections from such zoonotic influenza require either exposure in contaminated environments or intermediate contact with infected animals. Besides, the transmission among humans is not as efficient as human influenza viruses. However, it is always a concern that when such a virus evolved with enhanced ability to spread among human by adaptation or acquisition of specific genes from human influenza viruses, a new epidemic or even a pandemic may onset. Lessons should be learned from the pandemics to strengthen the surveillance of the emergence and spread of influenza virus strains. Also, quick responses to novel influenza strains are in demand, including

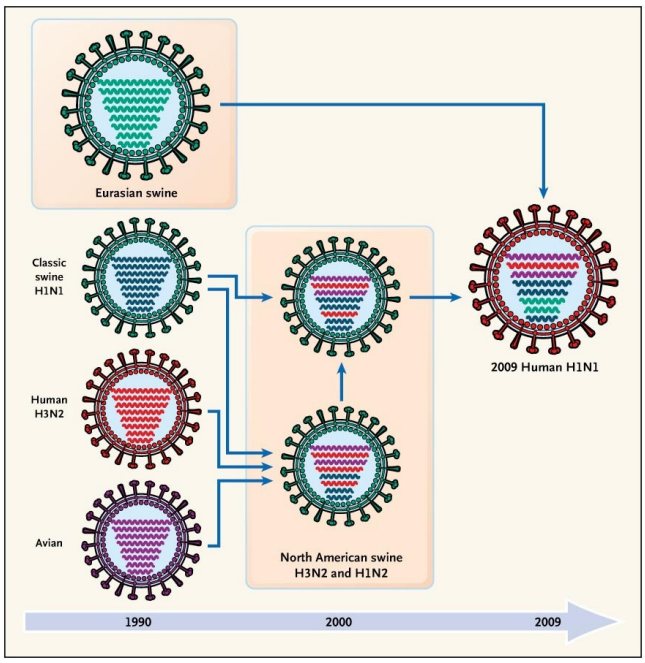


Figure 1.4: The reassortment events resulting in the 2009 influenza A/H1N1 virus. *The RNA segments shown within each virus are ranked from the longest to the shortest, namely the PB2, PB1, PA, HA, NP, NA, M and NS from top to bottom. (Trifonov* et al., *2009b)*

drug therapies and vaccination to the high-risk populations.

1.4 Vaccination and drug therapies for influenza

The vaccination program is the primary way for the public to prevent and control the spread of infectious diseases. Influenza vaccines were first applied to the human population in 1945 (Grabenstein *et al.*, 2006). Two types of influenza vaccines are developed; namely, the injected inactivated influenza vaccine, abbreviated as IIV, and the intranasally administered lively attenuated influenza vaccine, abbreviated as LAIV. The IIV is made from inactivated influenza viruses, while the LAIV is made from live attenuated viruses, both of which make the immune system recognize the antigens, produce antibodies in response, but do not cause influenza infections. The injected IIV takes up larger portion of influenza vaccination (Osterholm *et al.*, 2012).

Trivalent vaccines of both IIV and LAIV protect against three different influenza viral strains from influenza A/H1N1, A/H3N2 and one of the circulating influenza B lineage. Recently, quadrivalent vaccines, which include both influenza B lineage viruses, are available (World Health Organization *et al.*, 2018b). The compositions of both IIV and LAIV, reflecting the most dominant and recent circulating influenza viral strains, are assessed and updated twice annually by WHO.

The effectiveness of vaccination varies for seasonal influenza and can be significantly reduced in some seasons when the predominant strains are antigenically distant from the vaccine

strains. The insufficient protection of vaccines may contribute to severe influenza outbreaks such as observed in 1947 (Kilbourne *et al.*, 2002), 1997-1998 (de Jong *et al.*, 2000) and 2014-2015 influenza seasons (Rondy *et al.*, 2018). In contrast, when most circulating strains are matched with the compositions of influenza vaccines, the risk of illness by influenza can be reduced by 40%-60% (CDC *et al.*, 2018). For example, influenza vaccination was estimated to have prevented 5.3 million from sickness, 2.6 million medical visits, and 85,000 hospitalizations caused by influenza infections during 2016-2017. In 2017-2018, the vaccination was more effective, preventing around 7.1 million illness, 3.7 million medical visits, 109,000 hospitalizations, and 8,000 deaths caused by influenza infections (CDC *et al.*, 2018). Studies on the vaccine effectiveness (VE) from 2004-2014 suggested better protection against influenza A/H1N1 and influenza B than influenza A/H3N2 viruses (Belongia *et al.*, 2016). One reason to explain the lower VE against influenza A/H3N2 is that the influenza A/H3N2 has more mutations resulting in the antigenic change than the influenza A/H1N1 and B viruses. Thus, the influenza A/H3N2 viruses are more likely to be antigenically different from the recommended influenza A/H3N2 composition in influenza vaccines, reducing the VE against influenza A/H3N2 (CDC *et al.*, 2018).

Antiviral drugs are used as complements of vaccines for treatment. At present, two categories of drug therapies for influenza viruses have been approved for alleviating symptoms of human infection, including M2 ion channel inhibitors and NA inhibitors. Amantadine and rimantadine are M2 ion channel inhibitors against only influenza A viruses, preventing the viral invading to host cells. Zanamivir, oseltamivir, and peramivir are NA inhibitors against influenza A and B viruses, preventing the release of viral progenies (Centers for Disease Control and Prevention (CDC), 2018).

However, influenza viruses can develop drug resistance to those compounds. For example, the influenza A/H1N1 viruses responsible for the pandemic in 2009 showed resistance to both amantadine and rimantadine, compromising the drug therapies. Temporarily, antiviral resistance to NA inhibitors among circulating influenza viruses is low. Therefore, the M2 ion channel inhibitors are now not recommended for antiviral treatment. It should be noted that the drug recommendation will be changed if the circulating strains developed distinct drug binding preferences.

Mutations at active sites may intervene in the process of drug binding and thus compromise drug efficacy. It is significant to monitor those active sites and develop in-depth knowledge of potential resistance mutations before they are clinically observed. With the help of diverse computational models, it is possible to probe the binding interactions between a drug and a target protein (Trifonov *et al.*, 2009b). By simulating the drug binding process and calculating energies involved in this process, drug resistance can be measured.

1.5 Organization of the dissertation

Currently, profiling the pathogenicity of influenza viruses largely depends on animal models, and the risk assessment guidelines from WHO. The pandemic risk of an influenza viral strain is influenced by many factors, including the virus, the infected host, the population, and other pathogens. In this dissertation, the risk is only assessed from the perspective of the virus, especially the antigenicity and receptor binding specificity of influenza viruses. Computational models proposed in this dissertation mainly use the viral sequences, with less help from experimental assays and animal testing.

As to current computational models for predicting the mutations and antigenicity of influenza viruses, most works focus on mutations and pairwise co-mutations on the HA protein. Hemagglutination inhibition assay data and subtype specific features are often required for the prediction. In this dissertation, mutation patterns across influenza proteins are investigated. A universal sequence model for predicting the antigenicity of multiple influenza subtypes is proposed, which can help facilitate the quantification of antigenicity of highly pathogenic influenza subtypes (e.g. H5, H7, and H9).

Figure 1.5 illustrates the scope and organization of this dissertation. This dissertation intends to construct computational models for profiling the pathogenicity of influenza viruses.

Viral sequences are the primary data for computational analysis. Nucleotide sequences and protein sequences of influenza viruses are retrieved from public databases, such as National Center for Biotechnology Information (NCBI) (NCBI, 2014), Influenza Research Database (Squires *et al.*, 2012) and Global Initiative on Sharing All Influenza Data (GISAID) (Shu and McCauley, 2017). Besides, to obtain a more comprehensive description of the viral capability of host infection and drug resistance, structural analyses such as molecular docking and molecular dynamics simulation will also be incorporated. The dissertation is organized into chapters covering different aspects of the research on the virulence of influenza viruses.

Chapter 1 has briefly introduced the background of research, including the fundamental biology of influenza, the social burden caused by influenza, and current treatments for influenza. The main threat of influenza is resulted from the rapid evolution, leading to mismatch with recommended influenza vaccines. Therefore, in the following chapters, the mutations of influenza and how the mutations affect the phenotype, especially virulence-related aspects of influenza will be the main focus.

Chapter 2 provides a comprehensive literature review on current knowledge and computational models on the virulence determinants of influenza. The literature review covers well-established knowledge on the structural basis of influenza virus infection to host cells, especially the host receptor binding specificity and antigenicity of influenza. Besides, two types of state-of-the-art computational analyses on influenza are presented, i.e., the sequence-based modeling and the structure-based modeling on influenza.

Chapter 3 describes a phylogenetic-tree based method for uncovering pairwise co-mutations of intra-proteins, and a sequential rule mining based method for co-occurring mutations at

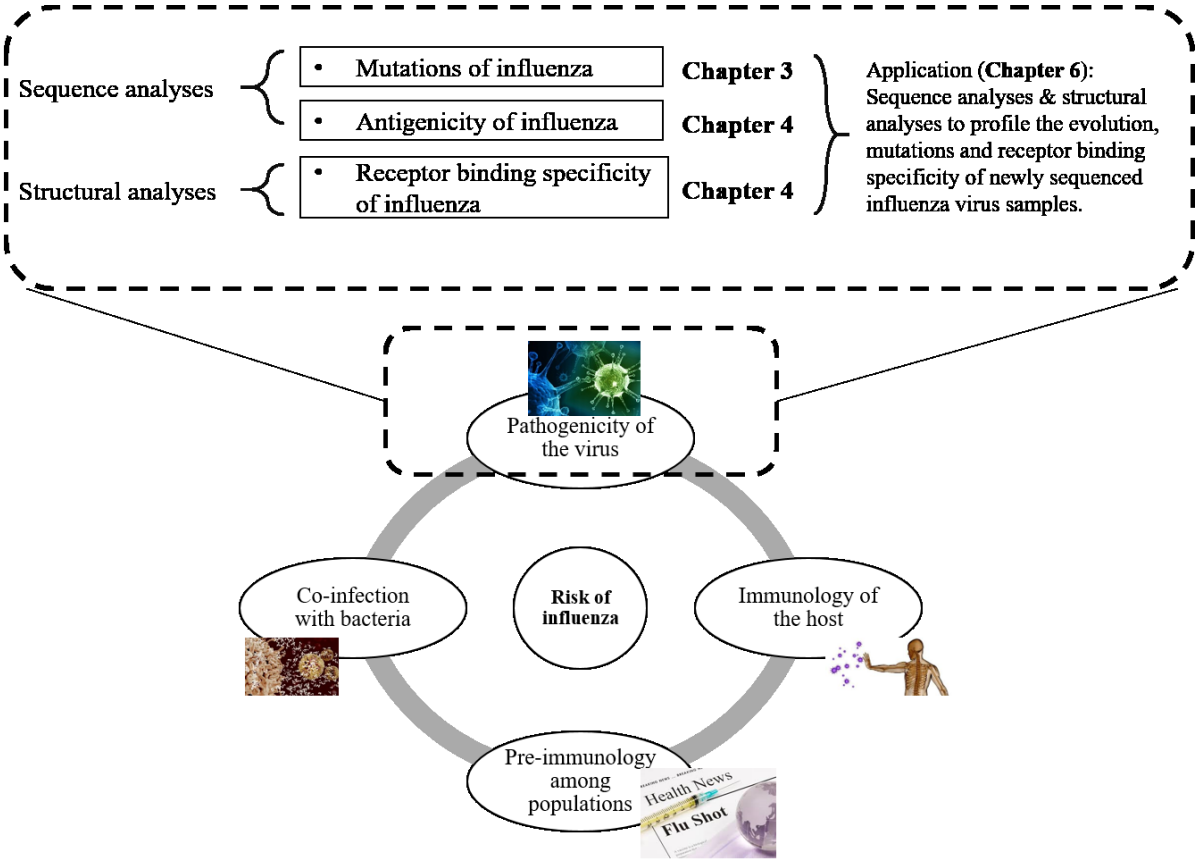


Figure 1.5: Illustration of the scope and organization of this dissertation.

multiple sites, even on different proteins. A list of case studies on influenza viral proteins is illustrated. The proposed methods have successfully detected dominant mutations responsible for the antigenic evolution of HA, and mutations clustering at functional domains of PB2.

Chapter 4 focuses on sequence-based computational models for predicting the antigenicity of influenza viruses. A context-free encoding scheme (CFreeEnS) for protein sequences has been proposed. Using CFreeEnS, the antigenicity of multiple subtypes of influenza viruses has been predicted accurately.

Chapter 5 focuses on profiling the receptor binding specificity of influenza viruses using structure-based methods. Taking a newly emerging influenza A/H7N9 as an example, the impact of HA mutations on the viral binding preference with host cells has been analyzed computationally in a systematical way.

Chapter 6 presents an application of the mentioned methods on newly sampled influenza viral strains. With the help of our collaborators, influenza viruses were sampled among outpatients and outpatients in Singapore during 2012-2015. The viral sequences have been investigated systematically to highlight distinctions between strains isolated from outpatients and inpatients.

Finally, Chapter 7 concludes the dissertation by wrapping up the methods and results introduced in previous chapters, mapping out possible directions for future research.

Chapter 2

Literature review

Because of the rapid evolution of influenza A viruses, it is always a concern that a newly emerging influenza viral strain may spread among humans, and even raise a worldwide pandemic. The WHO uses the Influenza Risk Assessment Tool (IRAT) (Trock *et al.*, 2012; Cox *et al.*, 2014) to evaluate the pandemic risk of a new influenza viral strain before it circulates among human. The evaluation considers factors from the virus, the population, and the ecology, as listed in Table 2.1. From the viral perspective, the virulence is determinant on the genomic variation, the ability for receptor binding, transmission among hosts, and defending from host immunity, including antiviral treatments. From the perspective of population and ecology, all the factors, such as the existing population immunity, the effectiveness of recommended vaccines, the global geographic distribution, the levels of population heterogeneity and human mobility, may contribute to rapid diffusion of pandemic influenza (Merler and Ajelli, 2009).

This dissertation mainly focuses on the virus, investigating viral sequences and empirical data collected from functional assays. This chapter presents a comprehensive literature review on current knowledge about the virulence determinants of influenza, mainly covering the structure basis (Section 2.1) and the state-of-the-art computational models for analyzing the receptor binding specificity and the antigenicity of influenza viruses. Sequence-based computational models and structure-based models are introduced in Section 2.3.2 respectively.

Table 2.1: Elements for the influenza risk assessment tool.

Virus	1. Genomic variation
	2. Receptor binding
	3. Transmissibility in laboratory animal models
	4. Antiviral and treatment options
Population	5. Existing population immunity
	6. Disease severity and pathogenesis
	7. Antigenic relationship and vaccine to candidates
Ecology	8. Global geographic distribution
	9. Infections in animals, human risk of infection
	10. Human infections and transmission

2.1 Role of influenza viral genes in virulence

Empirical experiments have reported several particular genetic mutations contribute to the enhancement of activity in viral life (as illustrated in Figure 1.2), including the viral binding to host cell receptors, fusion of the viral membrane for entry into host cells, genome transcription and translation, the assemble and release of virus progeny, and the escape from immune responses. Such aspects and mutations affecting the processes are determinant for the virulence of influenza viruses. This section reviews the functions of virulence determinant proteins.

2.1.1 Glycoproteins

The glycoproteins HA and NA have complementary functions, responsible for the virus entering into and releasing from host cells respectively. The functional balance of HA and NA are important for efficient circulation among hosts, which could be a critical indicator for the potential of the viral strain causing a pandemic (Xu *et al.*, 2012a). This section presents the functional domains and known virulence determinants on HA and NA respectively.

HA

HA is a membrane fusion glycoprotein on an influenza viral particle. The HA bears both receptor-binding sites for binding host cells and antigenic sites for neutralizing antibodies (Skehel and Wiley, 2000). Wiley *et al.* resolved the HA structure for the first time (Wiley *et al.*, 1981). The HA protein is homotrimeric, composed of a globular head domain and a stalk extending from the viral membrane. The receptor-binding sites and the antigenic sites reside the globular head. Figure 2.1 is a schematic representation of an unfolded polypeptide chain of HA0 of H1, with functional regions depicted in different colors, while Figure 2.2 shows the folded 3D structure of HA.

The HA0 is a precursor, consisting of two polypeptides named HA1 and HA2. After an enzymatic proteolytic cleavage, the amino terminus of HA2 is exposed, which contains the fusion peptide and allows for the viral fusion with the infected cell. Whether the HA0 precursor contains a multi-basic cleavage site is reported as a major virulence determinant in influenza viruses (Tscherne and García-Sastre, 2011). A conserved arginine (Arg/R) is found prevalent at cleavage site of low pathogenic avian influenza viruses, abbreviated as LPAI. This conserved cleavage site can be mediated by host trypsin-like enzymes. In contrast, the high pathogenic avian influenza (HPAI) contains multiple cleavage sites on the HA0 precursor, which allow for the recognition of HA2 by other host enzymes (e.g. furin and PC6). Therefore, the multi-basic cleavage site can result in more efficient invading to host cells (Kawaoka and Webster, 1988). To date, the HA proteins of HPAI viruses with multiple cleavage sites in the HA0 precursor, are exclusively identified as H5 and H7 (Tscherne and García-Sastre, 2011).

The esterase of HA1 locates between the fusion domain (F') and the receptor-binding domain (almost in the middle of the unfolded HA1). In the folded HA protein structure, as shown

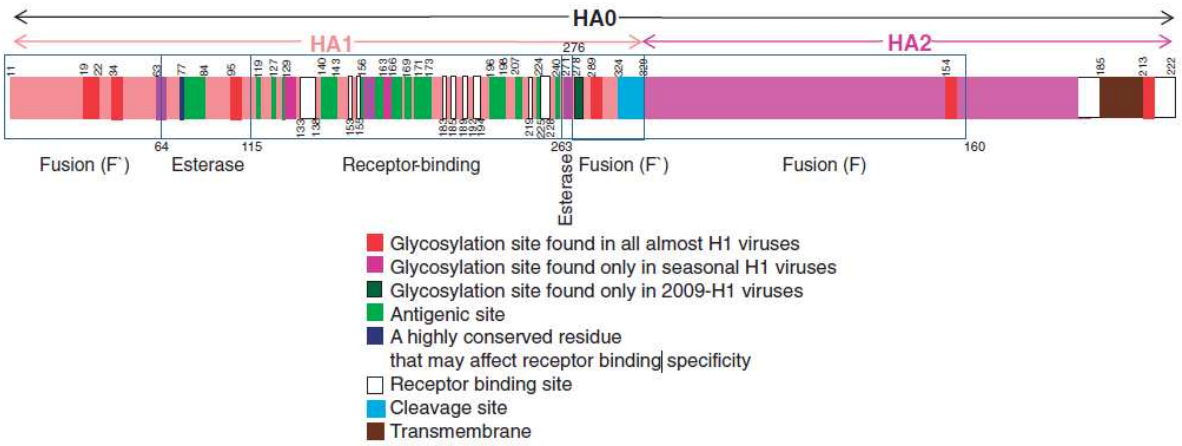


Figure 2.1: A schematic of an unfolded polypeptide chain of HA0 of H1 virus. (Skehel and Wiley, 2000) *The HA0 is composed of HA1 and HA2, mainly containing the fusion, esterase and receptor binding regions.*

in Figure 2.2, the receptor-binding domain is at the globular head, forming a shallow pocked where the four conserved residues Y98, W153, H183 and Y195 (named after the H3 numbering system (Burke and Smith, 2014)) serve as bases of the binding pocket. The binding pocked for sialic acid is formed by three structural domains nearby the four conserved residues, namely the 130-loop, 190-helix and 220-loop. The length and amino acid compositions in the receptor-binding domain vary between strains and can the determinants of the preferential receptors.

The HA plays a critical role in differentiating hosts. The avian influenza virus strains have preference for the sialic acid moieties terminated with the $\alpha_{2,3}$ linkage, while the human influenza viruses often prefer binding with the sialic acid moieties with an $\alpha_{2,6}$ linkage. The glycans with $\alpha_{2,3}$ and $\alpha_{2,6}$ linkage are often named as the "avian" and "human" receptors respectively. Both types of receptors can be found in human, but the $\alpha_{2,6}$-linked sialic acid moieties are mostly distributed in the upper respiratory tract. The $\alpha_{2,3}$-linked avian receptor are more preferentially expressed in the lower respiratory tract. The infection in the upper respiratory tract often results in mild symptoms, whereas the infection in the lower respiratory tract can cause severe illness (e.g. lung infection). Acquiring the ability to bind human receptors is critical for an avian influenza A viral strain to spread efficiently among humans and cause pandemics. Several residues in HA have been reported to control the receptor binding specificity of influenza viruses. For the HAs of H1, receptor binding for avian receptors are supported by residues E190 and G225 (H3 numbering), while the binding for human receptors are indicated by D190 and D225 (Glaser *et al.*, 2005). The HAs of H2 and H3 containing Q226 and G228 have binding preference for avian receptors, whereas the HAs containing L226 and S228 have binding preference for human receptors (Stevens *et al.*, 2004).

Influenza epitopes have been identified through exposure to neutralizing monoclonal antibodies (mAbs), which are produced by identical immune cells (Webster and Laver, 1980). Under selective pressure, influenza viruses typically can generate genetic variants to reduce antibody binding. By analyzing the viral sequences, the amino acid positions comprising an

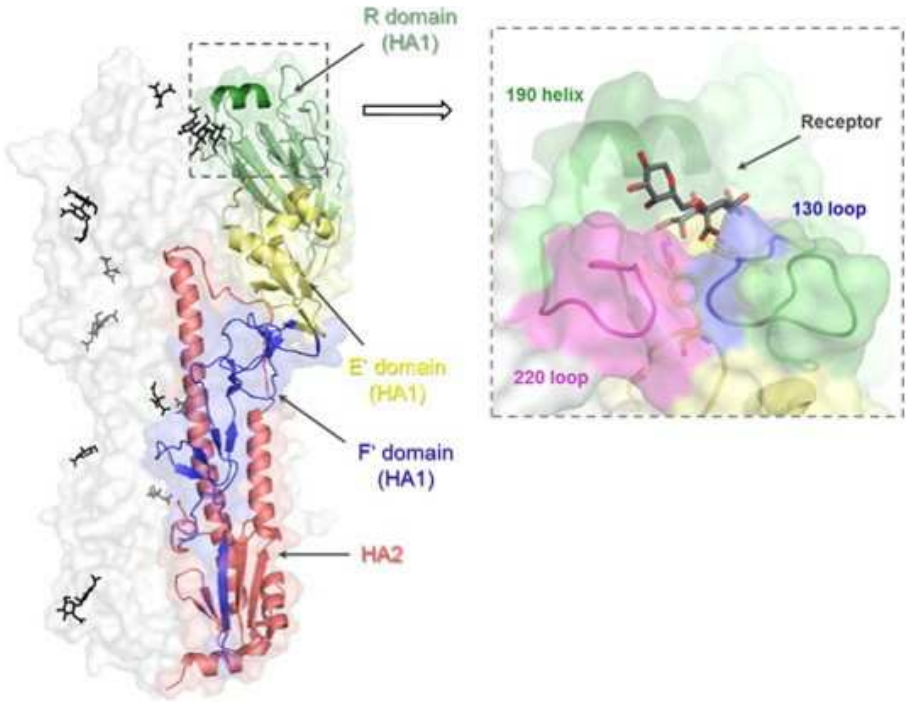
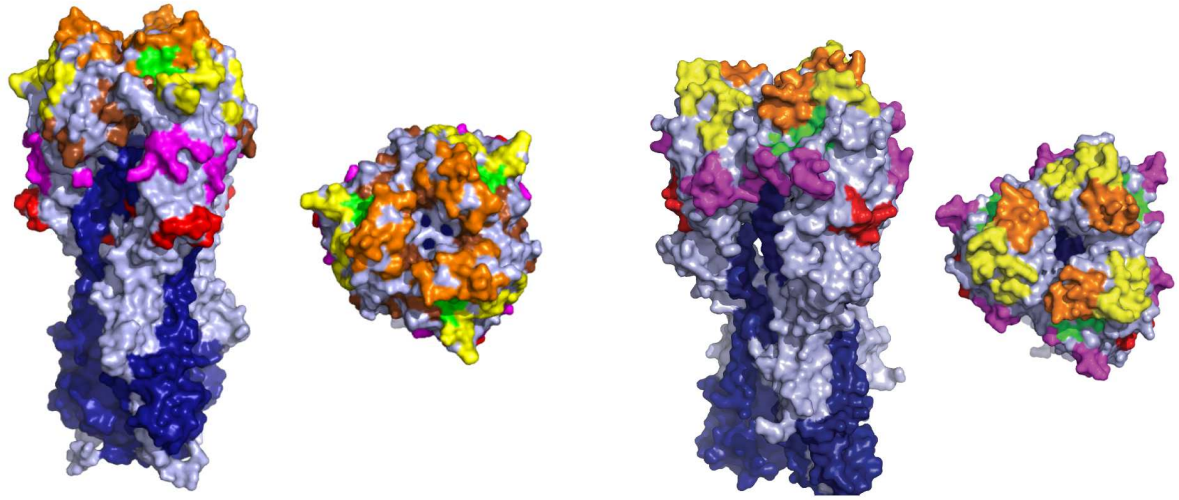


Figure 2.2: The folded HA protein structure and functional domains. (Mair *et al.*, 2014)

epitope can be identified. Substitutions on those sites are likely to enable the influenza variant to escape the mAb binding. Five antigenic sites, named as A-E, have been identified on the HA1 surface of H3 by conducting the mAb competition assays (Webster and Laver, 1980; Skehel *et al.*, 1984). Subsequently, the substitutions were mapped onto the HA structure to define five distinct clusters with antigenic sites A-E and surface-exposed residues, shown in Figure 2.3a. Similarly, five comparable antigenic sites for HA1 of H1 were identified at the globular head of HA1 (as shown in Figure 2.3b). Studies have revealed that all the antigenic sites are surface-exposed, and most of them are around the receptor-binding sites (Skehel and Wiley, 2000). Besides, the antigenic sites mainly reside on loop-like structures. Mutations on the loop-like structures are more likely to keep the stability and functions of the HA protein, and therefore cause less cost in viral fitness (Wilson and Cox, 1990). Table 2.2 listed the receptor binding sites and antigenic sites of H1, H3 and H7. The antigenic sites of H7 are predicted conserved epitopes with minimal variation using ABCpred, BepiPred and LBtope (Wang *et al.*, 2016).

Because the receptor binding sites and antigenic sites are too close and even with many overlaps, mutations can reshape the receptor-binding specificity, antigenicity and avidity of the influenza virus all together (Daniels *et al.*, 1984; Hensley *et al.*, 2009; Li *et al.*, 2013). For example, the HA mutation Q226L changes the viral binding preference from avian to human receptors. Also, the viruses with residues HA Q226 and L226 are distinct in antigenicity which can be distinguished through the mAbs assays (Daniels *et al.*, 1984). Similarly, the HA mutation N145K was reported to affect the antigenicity by altering the binding with the host cells and the antibody responses (Li *et al.*, 2013).



(a) HA Structure of A/Aichi/2/68(H3N2) (PDB ID: 1HGG (Sauter *et al.*, 1992)). Antigenic sites A-E are colored yellow, orange, red, brown and pink respectively.

(b) HA Structure of A/Puerto Rico/8/34(H1N1) (PDB ID: 1RU7 (Gamblin *et al.*, 2004)). Antigenic sites (Ca_1 and Ca_2), Cb, Sa, and Sb are colored pink, red, yellow, and orange, respectively.

Figure 2.3: HA structure of H1 and H3 with antigenic sites and receptor binding sites highlighted. *The HA1 and HA2 are colored light blue and dark blue respectively, while the receptor binding sites are colored green.*

Table 2.2: A summary of antigenic sites and the primary receptor binding sites of H1, H3 and H7. * *The antigenic sites of H7 are predicted results (Wang* et al.*, 2016); the receptor binding sites are the corresponding sites mapped from H3.*

Subtype	Functional domain		HA1 amino acid position
H1	Antigenic site (Brownlee and Fodor, 2001) (Kash *et al.*, 2010) (Caton *et al.*, 1982)	Ca_1	166, 167, 168, 169, 170, 203, 204, 205, 235, 236, 237
		Ca_2	137, 138, 139, 140, 141, 142, 221, 222
		Cb	69, 70, 71, 72, 73, 74
		Sa	124, 125, 153, 154, 155, 156, 157, 159, 160, 161, 162, 163, 164
		Sb	184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195
	Receptor binding site (Skehel and Wiley, 2000) (Soundararajan *et al.*, 2011)		91, 94, 142, 150, 151, 152, 127-135, 180-191, 216-225
H3	Antigenic site (Bush *et al.*, 1999) (Shih *et al.*, 2007)	A	122, 124, 126, 130, 131, 132, 133, 135, 137, 138, 140, 142, 143, 144, 145, 146, 150, 152, 168
		B	128, 129, 155, 156, 157, 158, 159, 160, 163, 164, 165, 186, 187, 188, 189, 190, 192, 193, 194, 196, 197, 198, 199
		C	44, 45, 46, 47, 48, 49, 50, 51, 53, 54, 273, 275, 276, 278, 279, 280, 294, 297, 299, 300, 304, 305, 307, 308, 309, 310, 311, 312
		D	96, 102, 103, 117, 121, 167, 170, 171, 172, 173, 174, 175, 176, 177, 179, 182, 201, 203, 207, 208, 209, 212, 213, 214, 215, 216, 217, 218, 219, 222, 223, 225, 226, 227, 228, 229, 230, 233, 238, 240, 242, 244, 246, 247, 248
		E	57, 59, 62, 63, 67, 75, 78, 80, 81, 82, 83, 86, 87, 88, 91, 92, 94, 109, 260, 261, 262, 265
	Receptor binding site (Weis *et al.*, 1988) (Skehel and Wiley, 2000)		98, 145, 153, 154, 155, 183, 195, 131-138, 183-195, 219-228
H7*	Antigenic site (Wang *et al.*, 2016)		37-52, 131-142, 215-234
	Receptor binding site		106, 152, 160, 161, 162, 192, 204, 139-146, 192-204, 228-237

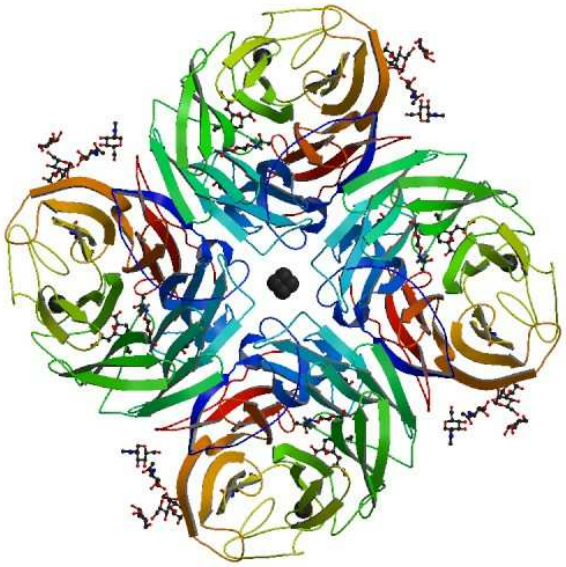
In addition, the substitutions introducing or destroying glycosylation sites are important virulence determinants. Introducing an N-glycosylation site near the antigenic sites may contribute to viral escape from antibody neutralization and zoonotic transmission (Skehel *et al.*, 1984; Gambaryan *et al.*, 1998; Kim *et al.*, 2018). For instance, the human influenza H3 have acquired four novel glycosylation sites during 1968-1999, which may have contributed to the influenza A/H3N2 pandemic in 1968 (Skehel and Wiley, 2000). Recent study on the 1918 and 2009 influenza A/H1N1 viruses highlighted the importance of substitutions in glycosylation sites, which possibly caused immune escape while maintained the receptor binding specificity (Kim *et al.*, 2018).

NA

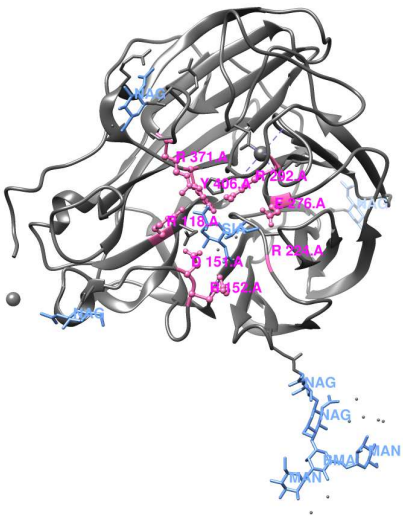
The primary function of NA is to cleave the viral progeny from host cells in virus replication process. The functional balance between HA and NA is critical to the efficiency of viral replication. For example, studies have indicated that the N2 adapted to cleave both the avian and human receptors (with $\alpha_{2,3}$ and $\alpha_{2,6}$-linked sialic acid moieties respectively) to meet the receptor binding specificity of H3. The balanced function between H3 and N2 may partially contribute to the influenza A/H3N2 pandemic in 1968 (Baum and Paulson, 1991). Besides, the NA structure shows a large and highly conserved active site, making it a good target for antiviral drugs. Efforts on the structure-based design of antiviral drugs have been made since 1970s (Air, 2012). Currently, antiviral drugs with NA inhibitors are recommended for the treatment against influenza A and B viruses, including zanamivir, oseltamivir and peramivir (Centers for Disease Control and Prevention (CDC), 2018). The NA inhibitors can bind the active sites of NA on the surface of infected cells, and therefore prevent the NA from cleaving viral progeny from infected cells.

Figure 2.4a shows the structure of NA, which is a tetramer with four identical polypeptides containing about 470 amino acids for each (Zhu *et al.*, 2012b). The folded structure of NA consists of four domains, namely the N-terminal cytoplasmic, the following transmembrane, the head containing enzyme active sites and calcium binding domain, and stalk connecting the transmembrane with the head (Shtyrya *et al.*, 2009). Figure 2.4b shows NA in complex with sialic acid. For the N2, residues in direct contact with sialic acid and account for the enzyme catalyst are Arg/R118, Asp/D151, Arg/R152, Arg/R224, Glu/E276, Arg/R292, Arg/R371 and Tyr/Y406. Other structural amino acid residues are Glu/E119, Arg/R156, Trp/W178, Ser/S179, Asp/D (or Asn/N in N7 and N9) 198, Ile/I222, Glu/E227, Glu/E277, Asp/D293, and Glu/E425.

The active site of NA is conservative in structural across subtypes of influenza and cross types of the enzyme. The conservation of NA active site points to the importance of all components, and makes it the main target for antiviral drugs. However, the antiviral drugs can also increase selective pressure for the influenza viruses, where the survived strains may have gained mutations resistant to the drug. For example, the NA H274Y substitution is a drug-resistant signature, but this mutation also compromises the viral fitness. Fucntioning with two other substitutions V234M and R222Q, surface-expressed NA were enhanced, making the virus



(a) The 3D structure of NA (PDB ID: 4GZQ) with four identical polypeptides (Zhu *et al.*, 2012b).



(b) The neuraminidase N2 from influenza A/Tanzania/205/2010 H3N2 in complex with sialic acid (PDB ID: 4GZQ (Zhu *et al.*, 2012b)). Active sites are colored pink while sialic acid and glycans are colored blue.

Figure 2.4: NA structure of N2 with active sites highlighted.

resistant to oseltamivir predominant during the 2017/18 influenza season (Bloom *et al.*, 2010). Monitoring the NA signatures of viral resistance is important for upcoming influenza seasons.

2.1.2 Influenza viral polymerase

The RNA polymerase of influenza A and B viruses consists of three subunits, including PB2, PB1, and PA. The viral polymerase complex mainly functions in trascription of viral genes and replication of the vRNA, making it an important contributor to the viral pathogenicity. The capping activity of influenza viral polymerase is dependent on cap-donors from host capped RNAs (Plotch *et al.*, 1981). Typically, the PB2 captures the 5'-end of nascent mRNAs in host cells using the cap-binding domain. Subsequently, the PA uses the endonuclease domain to cleave the capped RNA fragments as primers initiating transcription (Dias *et al.*, 2009).

Figure 2.5 shows the structure of a RNA polymerase heterotrimer with subunits PB2, PB1 and PA. The PB1, locating at the center of the structure, contains six motifs (A-F) and a primary loop (shown in Figure 2.5a). The PB1 has a canonical right handed fold arrangement, with additional N- and C-terminal extensions facilitating the interaction with PA and PB2 (Figure 2.5b). PA consists of two main domains (C- and N-terminal domain) connected by a long linker, wrapping the external side of PB1 finger (Figure 2.5c). PB2 is composed of an N-terminus interacting with PB1, the mid-domain, the cap-binding domain, the cap-627 linker, and the C-terminal nuclear localization signal (Figure 2.5d). The N-terminus of PB2 and the central cavity in PB1 form a polymerase active site.

The cap-627 linker is named after a host range and virulence determinant of PB2. Specifi-

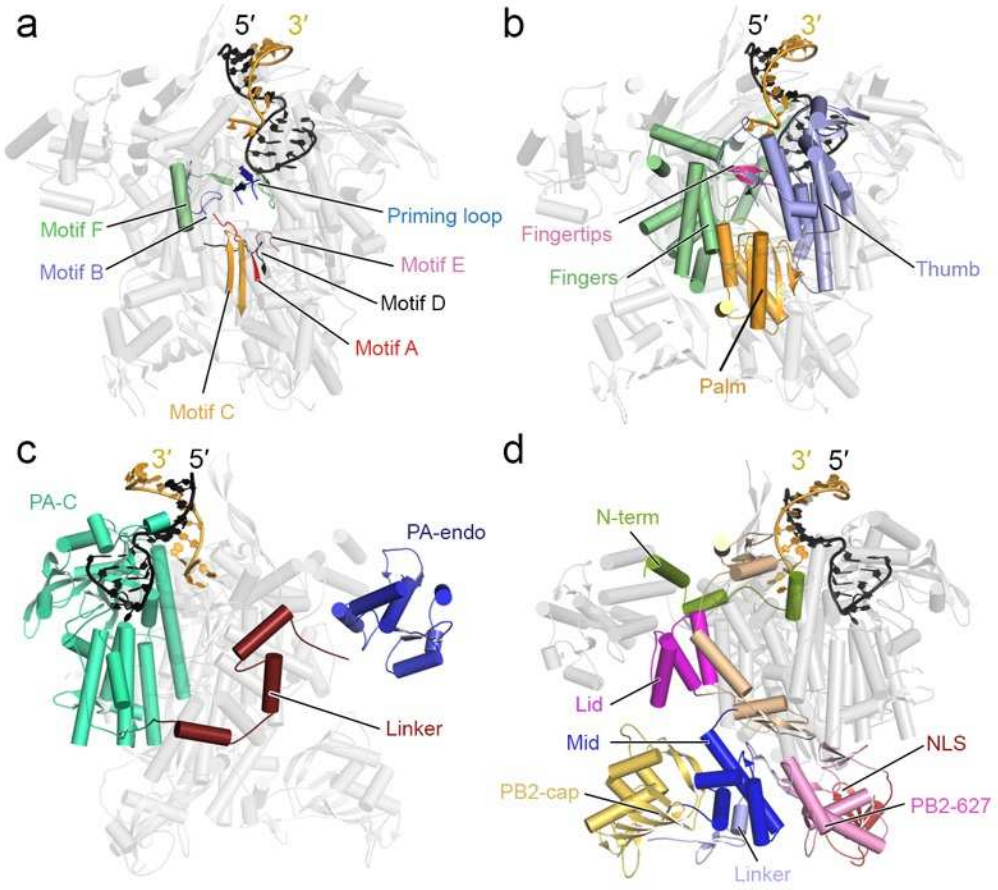


Figure 2.5: The structure of influenza A virus polymerase in different orientations (PDB ID: 4WSB) (Reich *et al.*, 2014; Te Velthuis and Fodor, 2016,?). *(a). PB1 motifs and the priming loop; (b). the right handed arrangement; (c). PA; (d). PB2 domains.*

cally, the PB2-E627 is typically isolated from the avian influenza strains, while the PB2-K627 is generally found in human influenza strains (Subbarao *et al.*, 1993). Besides, researchers observed that the PB2-E627K substitution in avian influenza A/H5N1, causing lethal infections in mice after a single passage but becoming non-pathogenic for mammals (Hatta *et al.*, 2001, 2007; Mase *et al.*, 2006). Researchers have investigated how the substituion E627K enhances the pathogenicity in mice and adaptation to mammals. Experiments suggested that the PB2 627 controls the temperature sensitivity of the vRNA replication (Massin *et al.*, 2001). It is well-established knowledge that the human influenza viruses prefer replicating at 33°C, the same temperature as in the human upper respiratory tract. In contrast, the avian influenza viruses prefer a higher temperature of about 41°C. Experiments showed that the PB2-E627K substitution enhanced the replication of avian influenza viruses in mammalian cells at 33°C *in vitro*. In addition, viruses with PB2-K627E showed reduced transmission in the guinea pig, which could also be explained by the temperature sensitivity resulting in weakened replication in the upper respiratory tract (Steel *et al.*, 2009).

Another virulence determinant in PB2 is the site 701. The PB2 substitution D701N facilitates the viruses to replicate in mammal cells, most likely by enhancing the PB2 binding with importin-α_1 (Gabriel *et al.*, 2008). Besides, the PB2 N701 is reported to be compensate for PB2 E627 which reduces the transmissibility among mammals (Steel *et al.*, 2009). The 590 and 591 on PB2 are two potential virulence determinants. Both S590 and R591 are found in the swine influenza A/H1N1 strains with the absence of PB2 K627 and N701 (Mehle and Doudna, 2009a).

2.1.3 Non-structural proteins

In virology, a non-structural protein is encoded by a virus but expressed in the host cells instead of being part of the viral particle. Specifically for the influenza viral proteins, NEP (formerly known as NS2) and NS1 are encoded by the NS segment; PB1-F2 and PB1-N40 are encoded by the PB1 segment; PA-X, PA-N155, and PA-N182 are encoded by the PA segment. All of them are expressed in the infected cells by splicing, alternative open reading frames, or ribosomal frameshifts. None of them is the composition of the influenza viral particle. These proteins, with localization and functionality probably regulated by post-translational modifications, are speculated to be multi-functional (Klemm *et al.*, 2018).

NS1

NS1 is probably the most active non-structural protein, responsible for many interactions between the infected cells and the viral particles. Notably, it facilitates the virus to antagonize the antiviral response of human immune system (Egorov *et al.*, 1998). The NS1 is composed of an RNA-binding domain (RBD) and an effector domain (ED) joined by an flexible linkage. Typically, the folded 3D structures of NS1 are determined independently for each domain. Figure 2.6 shows crystallographic structures of the RBD, a helix-helix dimer of the ED and a

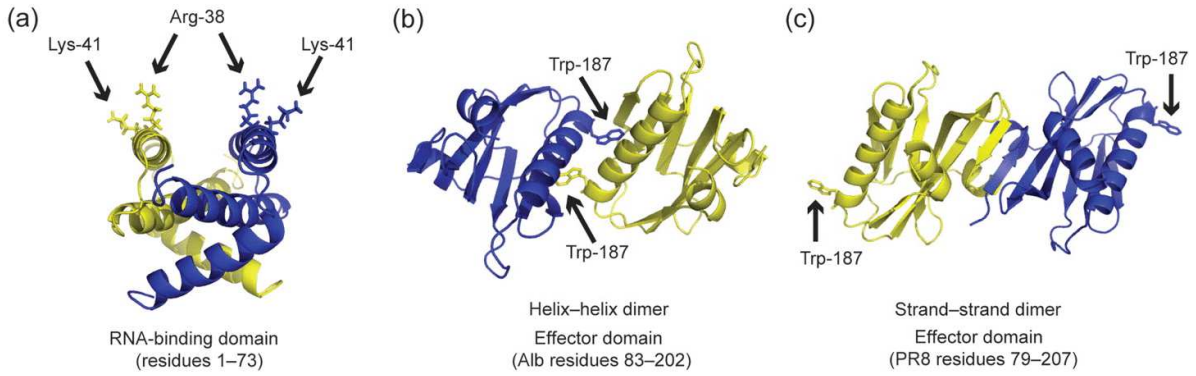


Figure 2.6: Structures of the RNA-binding domain and the effector domain of NS1 (Marc, 2014). *(a). The dimeric RNA-binding domain of NS1 (influenza A/Udorn/72(H3N2) PDB ID:1NS1 (Chien* et al.*, 1997)) (b). The helix-helix dimer of the effector domain (influenza A/Duck/Albany/76(H12N5) PDB ID: 2GX9 (Bornholdt and Prasad, 2006)) (c). The strand-strand dimer of the effector domain (influenza A/Puerto/Rico/8/34 (H1N1) PDB ID: 3D6R (Hale* et al.*, 2008b))*

strand-strand dimer of ED (Hale *et al.*, 2008a).

The NS1 mainly functions to forge a multipronged attack against the immune response by antagonizing IFN-α/β-mediated antiviral responses (García-Sastre, 2001).Besides, the NS1 bears many other functions associated with the replication and pathogenicity of influenza, such as the temporal regulation of the synthesis of vRNA, the splicing and translation of viral mRNA. The functions of NS1 largely depends on its involvement in protein-RNA and protein-protein interactions.

Several virulence determinants on NS1 have been reported. Table 2.3 summarizes reported virulence determinants on NS1. For example, the NS1 substitution P42S contributes to increased virulence in mice and decreased level of IFN-α/β *in vitro* (Jiao *et al.*, 2008). Similarly, the NS1 E92 is a signature for HPAI H5N1, while D92 typically appears in LPAI strains (Heui Seo *et al.*, 2002). Besides, the viral strains with NS1 E92 can replicate in the presence of IFN *in vitro*, which may partially explain the pathogenicity in pigs. The NS1 substitutions L103F and I106M have been reported to facilitate the viral replication *in vitro* probably because the mutant NS1 suppressed the expression of IFN-α/β mRNAs (Twu *et al.*, 2007). Besides, the three substitutions R108K, E125D and G189D have been reported to restore the CPSF30 binding activity of the swine pandemic influenza A/H1N1 virus, and attenuate its virulence in ferrets and mice (Hale *et al.*, 2010).

Apart from signature positions, the C-terminus portion of NS1 has been found associated with the pathogenicity of influenza (Obenauer *et al.*, 2006). Both the influenza A/H1N1 in 1918 pandemic and the HPAI A/H5N1 contain four C-terminus amino acids, showing increased virulence in mouse models (Jackson *et al.*, 2008). Structural analysis shows that the C-terminus portion forms a PDZ ligand domain. Although the mechanism or casual relationship between the C-terminus portion and the increased virulence is unclear, the C-terminus portion of NS1 has been taken as a virulence determinant of influenza viruses.

Table 2.3: Reported virulence determinants on NS1.

Virulence determinants	Description	Reference
P42S	Increased virulence in mice	(Jiao *et al.*, 2008)
D92E	High pathogenicity in avian and pig	(Heui Seo *et al.*, 2002)
L103F	Increased viral replication	(Twu *et al.*, 2007)
I106M	Increased viral replication	(Twu *et al.*, 2007)
R108K	Attenuated virulence in ferrets and mice	(Hale *et al.*, 2010)
E125D	Attenuated virulence in ferrets and mice	(Hale *et al.*, 2010)
G189D	Attenuated virulence in ferrets and mice	(Hale *et al.*, 2010)
The C-terminus portion	Forming PDZ ligand domain motif; Increased virulence in mice	(Obenauer *et al.*, 2006) (Jackson *et al.*, 2008)

PB1-F2

PB1-F2 is encoded by the PB1 segment, transcribed from an alternative open reading frame after a ribosomal frame shifting (Chen *et al.*, 2001). PB1-F2 is not expressed in all influenza viruses, but almost in all avian strains. PB1-F2 is expressed with varying length, especially tending to be truncated when introduced to mammalian cells (McAuley *et al.*, 2010). The prevalence of truncated PB1-F2 suggests that the expression of full-length PB1-F2 is not essential for the viral fitness (Klemm *et al.*, 2018).

PB1-F2 is indispensable in regulating the viral infection, promoting the secondary bacterial infection, and modulating host immune responses (McAuley *et al.*, 2007b; Dudek *et al.*, 2011).

PB1-F2 shows the ability to bind the PB1 polymerase and enhance the polymerase activity *in vitro* (Mazur *et al.*, 2008). Besides, it can inhibit natural killer cells and cause exacerbated inflammation (Vidy *et al.*, 2016). However, the molecular basis of how PB1-F2 functions in the viral infection or antagonizes the human immune system needs further investigation.

Researchers found that PB1-F2 could be responsible for the increased pathogenicity of the pandemic strains in 1918, 1957, and 1968, also the recent high pathogenicity avian influenza A/H5N1 strains (McAuley *et al.*, 2007b, 2010). Interestingly, truncated PB1-F2 with only the first 11 amino acid of the N-terminus was expressed in the 2009 pandemic influenza A/H1N1 strains, which turned to be non-functional, and even signature of low pathogenicity (Trifonov *et al.*, 2009a). The function of PB1-F2 may be host and strain specific, greatly influenced by the context of viral genome (Tscherne and García-Sastre, 2011).

Besides the truncation of PB1-F2, researchers have revealed additional cytoplasmic or nuclear localization of PB1-F2 controlled by amino acid 68-71 (Cheng *et al.*, 2017). Several substitutions have been reported important to the pathogenicity of viral strains. The PB1-F2 N66S mutation was reported to cause more severe infections, lung titers and cytokine production in mice (Conenello *et al.*, 2007). The increase virulence is speculated to be caused by the interaction between PB1-F2 and the mitochondrial antiviral signaling protein (MAVS), which blocks the IFN response (Dudek *et al.*, 2011; Varga *et al.*, 2011).

In addition, post-translational modifications are speculated to be important for PB1-F2. For

example, the T27 and S35 are phosphorylation sites for PB1-F2, which function in the viral infection process. Substitutions at these sites can modulate the proapoptotic in monocytes (Mitzner *et al.*, 2009). Ubiquitylation sites K73, K78 and K85 are highly conserved. Mutant avian influenza strains with substitutions K73R, K79R and K85R on PB1-F2 show enhanced polymerase activity, modulated IFN antagonism and increased virulence (Košík *et al.*, 2015).

2.2 The antigenicity of influenza viruses

In the immune system, antigen molecules are often specifically targeted by and bind with antigen receptors such as antibodies. The capability of an antigen in binding with the receptors is called antigenicity. Antigenic variation is the major way for antigens to escape the immune system. Influenza viruses are undergoing constant immune selective pressure. The antigenic variation of influenza viruses can be divided into antigenic drift and antigenic shift, characterized by both HA and NA proteins on viral surface (Webster *et al.*, 1982). In numerical terms, antigenic drift is to describe minor changes (one or two substitutions) in viral nucleotide sequences, while antigenic shift is for abrupt and major changes by intermixing two or more strains. The antigenic shift is mainly resulted from reassortment, when different influenza viruses co-infects a single host, they may take segments from each other (Treanor, 2004). Such event will result in a sudden drastic change in viral genome and produce a new viral strain. In this way, virtually all antigenic determinants of the HA or NA antigen altered. The subtype of influenza, denoted as HxNx, roughly represents the antigenicity of influenza viruses. The antigenic dissimilarity between two viral strains of the same subtype also needs to be measured.

Hemagglutination assay

Hemagglutination inhibition (HI) assay, micro-neutralization (MN) assay and enzyme-linked immunosorbent assay (ELISA) are the most widely used binding assays to characterize the antigenicity, measuring the serum cross-reactive antibodies to antigens. It is worth noting that the results are sometimes reported with inconsistency (Grund *et al.*, 2011). The Centers for Disease Control and Prevention (CDC) use HI assays to characterize the antigenicity of influenza viruses. Every year, about 2,000 influenza viruses are antigenically characterized by the CDC to determine the similarities of current circulating strains to those that were included in the influenza vaccines, providing insights on the efficacy of the flu vaccines (CDC *et al.*, 2015).

HA is a glycoprotein on the surface of influenza viruses, which binds to the sialic acid receptors when invading the host cell. The HA protein can also bind the red blood cells (RBCs, or erythrocytes), causing the erythrocytes clumping together (Amano and Cheng, 2005). This process and property is called hemagglutination. When agglutinated by a viral suspension, the erythrocytes will be prevented from settling out of suspension. This serves as the basis of tittering influenza viruses in a sample (Killian, 2008). Generally, the HI assay is conducted in a

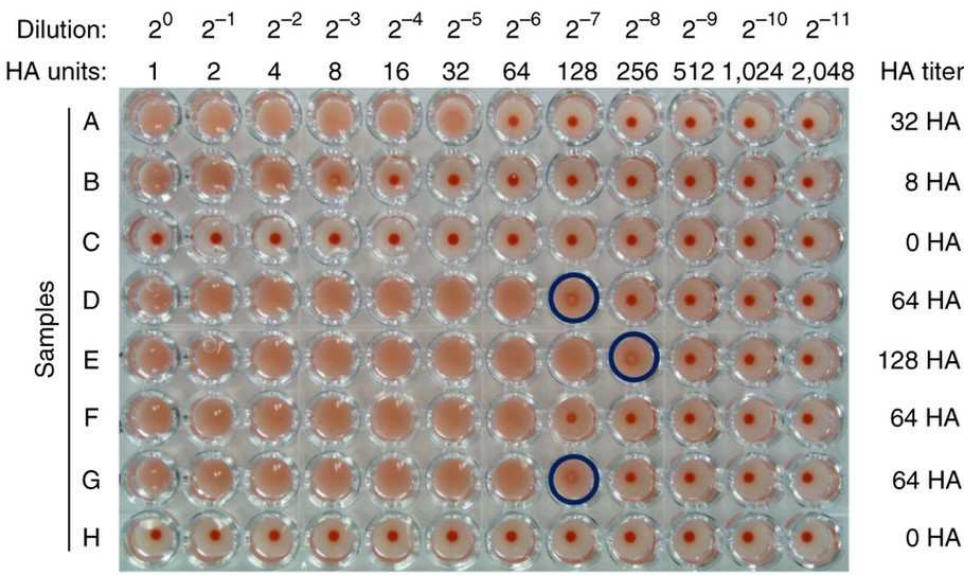


Figure 2.7: An example of hemagglutination assay (Eisfeld *et al.*, 2014).

96-well plate (8×12) with two-fold serial dilutions of an influenza virus, as shown in Figure 2.7 (Eisfeld *et al.*, 2014). A consistent number of erythrocytes are then added to wells of the plastic tray. After hemagglutination, the erythrocytes attached by the influenza viral particles will form a lattice coating the well. Otherwise, those not bound will sink to the bottom and form a button. From the dilution and HA titres, the amount of virus can be estimated. To be more specific, in the example presented in Figure 2.7, the samples containing influenza viruses (rows A-H) with two-fold serial dilutions ($2^0 - 2^{-11}$) were subjected to hemagglutination assay, mixed with turkey erythrocytes in a 96-well plate. After 30-minute incubation at room temperature, the wells were photographed. The HA titres for each sample were recorded to the right of the figure. Wells indicating partial agglutination have been highlighted with dark blue circles. In this example, the sample A formed a button at the dilution up to 2-6, indicating that the HA titer is 32; the sample C contained no detectable virus; sample D has an HA titer of 64.

Hemagglutination inhibition assay and antigenic distance

The hemagglutination inhibition assay takes advantage of the antibodies binding with influenza viruses, which will prevent the viral binding to RBCs. Figure 2.8 describes the interactions between the viral strain, antibody and RBCs (Palese, 2006). If the serum contains no antibodies reacting the viral strain or the antibodies diluted sufficiently, hemagglutination will be observed (Figure 2.8-B). Otherwise, with the presence of antibodies, hemagglutination will be inhibited, forming a button at the bottom of the plate (Figure 2.8-C). Finally, an HI titre provides information about the corresponding affinity of an antibody to the viral strain. For example, in Figure 2.9, hemagglutination is observed at the 2560 dilution. The 1280 dilution of antibody can still recognize the antigens on the surface of the viral sample and thus block hemagglutination, forming a button in the centre. Thus, the HI titre of this viral sample is 1280. The HI assay has been serving as a golden standard to characterize the antigenicity of influenza viruses

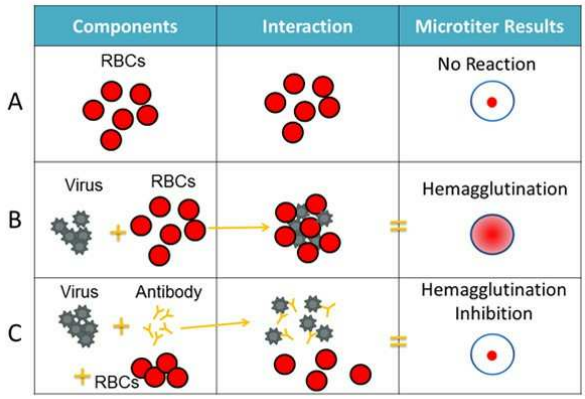


Figure 2.8: Mechanism of the hemagglutination inhibition assay (Palese, 2006). *A). The red blood cells sink to the bottom without reaction. B). Hemagglutination is observed when the viruses bind to red blood cells. C). Hemagglutination is inhibited when the antibody binds to the virus. Hence, the red blood cells sink to the bottom.*

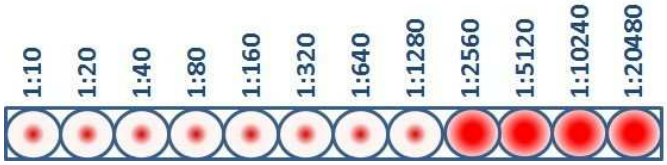


Figure 2.9: An example of hemagglutination inhibition assay.

and provide insights into the efficacy of the selected vaccine (Hobson *et al.*, 1972). But several limitations are born with this methodology. First, the grown of influenza viruses to produce vaccines must be in eggs, during which genetic changes can occur (also called egg-adapted changes). Second, current vaccine production takes 6-8 months, during which the circulating viruses may have evolved to be distinct from the selected vaccine strain. Thus, the HI tests can be sensitive to detecting antigenic differences and measuring the efficacy of the selected vaccine strain, but lack the ability to select effective vaccine strains. The recommendation of vaccine selection requires predicting antigenic variants of the next season.

Antigenic distance, conceptually the reciprocal of antigenic relationship, is a quantitative description of the antigenic difference between viral strains (Rédei, 2008). It can be derived from purely viral sequences or results of binding assays. The titer result of an HI assay can be taken as a measurement of distance between the antiserum and viral strain. Thus, pairwise antigenic differences between influenza viral strains can also be roughly estimated.

It should be emphasized that the HI data are mostly been interpreted by observing the wells. Thus, the precision is not high. Currently, there are no reliable methods to quantify such binding assay results. The interpretation of HI data is only applicable to judging large antigenic differences. In terms of influenza, such antigenic difference is often large enough to update the flu vaccine. Typical HI data for two strains and their corresponding antisera is shown as Table 2.4. The titer H_ij represents the maximum dilution for the antibody A_j inhibiting hemagglutination of V_i. Numbers on the diagonal of the table (H_{ii} highlighted as

Table 2.4: A typical 2×2 HI table with four titers for viral strains V_i and V_j.

	('Antiserum') A_i	('Antiserum') A_j
('Virus' or 'Antigen') V_i	H_{ii}	H_{ij}
('Virus' or 'Antigen') V_j	H_{ji}	H_{jj}

bold) represent homologous titers.

There are many ways to define the antigenic distance between viral strain V_i and V_j can be defined. Two widely used definitions are: $d_i = log_2(\frac{H_{ii}}{H_{ji}})$ (Smith *et al.*, 1999) and $d_2 = \sqrt{\frac{H_{ii}H_{jj}}{H_{ji}H_{ij}}}$ (Lee and Chen, 2004). Note that d_i is asymmetrical. In this context, the V_i is vaccine strain and V_j is the dominant circulating strain.

This 2×2 HI table can be extended to $n \times m$, where n viral strains and m antisera are analyzed. In practice, the health authorities provide HI tables with at least eight antisera to evaluate the antigenic distances between candidate vaccines and dominant circulating viral strains. The antigenic distances should be calculated pairwisely by picking up the corresponding four entries. Average distance (A-distance) is defined to measure the average difference of antigen-antisera interaction effect of two antigens, while the largest distance (L-distance) measures the maximum distance. AD_{ij} and LD_{ij}, denoted for the A-distance and L-distance between virus V_j and V_j, are defined as Equation (2.1) and (2.2).

$$AD_{ij} = \frac{1}{n}\sum_{t=1}^{n}|H_{it} - H_{jt}| \tag{2.1}$$

$$LD_{ij} = max_{t=1,2,...n}\{|H_{it} - H_{jt}|\} \tag{2.2}$$

In addition, antigens and antisera are usually clustered by flu seasons. When selecting vaccines, we should focus on the HI assays during the period containing antigens V_i or V_j. To take the temporal information into account, Cai *et al.* proposed a Mutual antigenic distance (M-distance), considering only the interaction of antigen V_i and V_j to antisera A_t in the same period as the two antigens (Cai *et al.*, 2012). Suppose the n antigens and m antisera are partitioned into k clusters, denoted as C_x (x is the cluster index, $x = 1, 2, ..., k$). The M-distance between viruses $V_i \in C_x$ and $V_j \in C_y$, denoted as MD_{ij} is calculated as Equation (2.3), where the $|\{A_t \in C_x \cup A_t \in C_y\}|$ calculates the number of antisera in the union of cluster x and cluster y.

$$MD_{ij} = \frac{\sum_{A_t \in C_x \cup A_t \in C_y}|H_{it} - H_{jt}|}{|\{A_t \in C_x \cup A_t \in C_y\}|} \tag{2.3}$$

Generally, the M-distance (MD) is the most robust among all the mentioned definitions. while A-distance (AD) is sensitive to noise. MD performs well in simulation studies for vaccine selection. But the results need to be validated on real data.

As the high-throughput sequencing techniques are becoming quicker and cheaper, it is more convenient to characterize antigenic distance by HA sequences. Constructing robust computational models to predict the antigenicity of influenza viruses is in demand.

2.3 Computational modeling on influenza

Benefited from the development of high throughput sequencing techniques and accumulated experimental data, computational models have been an attractive complement to experimental studies, safer and cheaper than animal models. However, the Achille's heel is the lack of established validating framework for the numerous computational methods, resulting in the concerns about accuracy of the predictions. The experimental results of animal models provide valuable information on the relationship between genotype and phenotype. By integrating the laboratory genotype-phenotype assessment and large-scale viral sequences, it is possible to delineate the biochemical traits of the viral protein and predict phenotype from sequences. In this section, both sequence-based computational analyses and structure-based computational models are reviewed briefly.

2.3.1 Sequence-based computational analyses on influenza

Phylogenetic analysis

A phylogenetic tree is the most traditional yet useful model for reconstructing adaptive evolutionary paths from genetic sequence data, inferring the origins of viral strains, highlighting decisive epidemiological or immunological events (Bedford *et al.*, 2011; Barrero *et al.*, 2011). For example, phylogenetic analyses on the genome set of influenza A/H3N2 viruses circulating in Japan suggested the reassortment event of six internal gene segments (Lindstrom *et al.*, 1998). Barrero *et al.* profiled the evolution of pandemic influenza A/H1N1 viruses in Buenos Aires, and monitored the drug resistance indicated by the NA fragment by constructing phylogenetic tress Barrero *et al.* (2011).

The BEAST (Bayesian Evolutionary Analysis Sampling Trees) program is widely used to estimate phylogenetic trees and divergence times (Drummond and Rambaut, 2007). Typically, the BEAST applies MCMC model to weight each tree proportional to the tree space, uses the maximum credibility to evaluate the posterior trees, and identifies the maximum clade credibility (MCC) tree.

Figure 2.10 presents a typical phylogenetic tree for HA of influenza A/H3N2 is constructed using BEAST with the MCMC model (Bedford *et al.*, 2011). Bedford *et al.* used 229 representative HA1 sequences of human influenza A/H3N2 viral strains from 1968 to 2013. Each virus sample is represented as a terminal node of the phylogenetic tree. The edge between parent and child nodes indicate a hypothetical ancestor-descendant relationship between the two samples. Evolutionary path is composed of branches between the ancestor and the descendant nodes, or referred to as tree topology between nodes. The root of the tree is inferred as the ancestor of all samples. Alternatively, in an unrooted tree, each node is positioned relative to one another without indicating the direction of evolution.

Typically the phylogenetic tree is presented in the horizontal dimension, while the vertical dimension is just for visualization. The length in horizontal stands for time or number of sub-

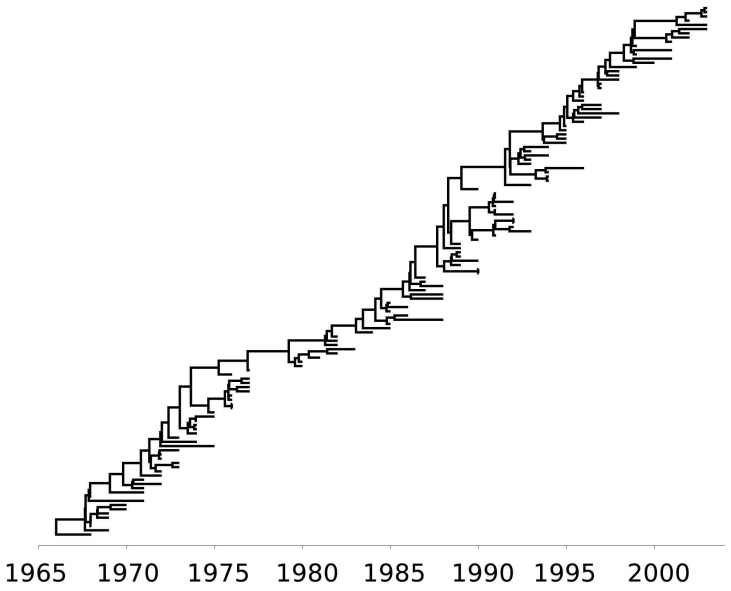


Figure 2.10: A typical phylogenetic tree for HA protein of influenza A/H3N2 (Drummond and Rambaut, 2007). *Viruses sampled from 1968 to 2002 and the tree is constructed by BEAST using the Markov chain Monte Carlo (MCMC) model.*

stitutions, while the length in vertical stands for nothing. In Figure 2.10, the node which has extended branches leading to more descendant nodes is positioned at the top. Substitution models are used for computing the genetic distances between samples, which provide a statistical description of the evolution supporting the construction of a phylogenetic tree.

Table 2.5 summarizes the assumptions of different nucleotide substitution models. The simplest nucleotide substitution model JC69 was proposed by Jukes *et al.*, hypothesizing equal frequencies as of 0.25 for the four nucleotide bases A, G, C, and T, respectively. The JC69 model also hypothesizes that the rates of one nucleotide substitutes any other are the same Jukes *et al.* (1969). The K80 model keeps the assumption of equal nucleotide frequencies of JC69, but distinguishes the nucleotide replacements between transitions, namely the replacements A$\leftrightarrow$G or C$\leftrightarrow$T, and transversions, namely the replacements between A/G $\leftrightarrow$ C/T (Kimura, 1980). The F81 model extended JC69 by allowing for different base frequencies of the four nucleotides (Felsenstein, 1981). Subsequently, Hasegawa *et al.* proposed HKY85 by integrating the K80 and F81, so that both the base frequencies and substitutions are allowed to vary (Hasegawa *et al.*, 1985). The general time reversible (GTR) model is a maximum-likelihood based approach, where the nucleotide substitutions are assumed to follow a homogeneous Markov chain (Tavaré, 1986; Yang, 1994). The molecular clock model assumes the substitutions occur in lineages at a constant rate (Zuckerkandl and Pauling, 1965), where the computed genetic distances to the root also represent the time (as shown in Figure 2.10). The time-resolved phylogenetic trees allow for the estimation of the time when an important divergent event happened.

Researchers have applied phylogenetic tree analyses to reveal the evolutionary patterns of influenza. For example, Bedford *et al.* looked into the evolution of human influenza A/H3N2 viral strains between 1968 and 2002, showing taht they have been undergoing continuous se-

Table 2.5: Summary of nucleotide substitution models.

Model	Assumptions and description	Reference
JC69	Equal nucleotide frequencies; identical substitution rates	Jukes *et al.* (1969)
K80	Equal nucleotide frequencies; distinctions between transtions and transversions.	Kimura (1980)
F81	An extension of JC69, allowing base frequencies different from 0.25.	Felsenstein (1981)
HKY85	Combination of K80 and F81	Hasegawa *et al.* (1985)
GTR (General Time Reversible model)	A stationary Markov process assuming constant equilibrium character state frequencies and the instantaneous transition probabilities through time.	Tavaré (1986) Yang (1994)
Molecular Clock	The model hypothesizes a particular substitution rate in lineages of a tree, and therefore genetic distance to the root also represents the time	Zuckerkandl and Pauling (1965)

lective pressure. The side branches typically persist 1-5 years before extinction, while a single trunk lineage has been predominant throughout the time (Bedford *et al.*, 2011). In contrast, the influenza B viruses diverge into two lineages co-circulating worldwide, named as influenza B/Yamagata and B/Victoria respectively (Biere *et al.*, 2010). Mining the nucleotide substitution patterns is important to reconstructing a reliable phylogenetic tree from genetic sequences, and thereby affecting the understanding of molecular sequence evolution. The main problem in phylogenetic tree analyses is evaluating the reliability and interpreting the results. Generally, the phylogenetic tree analyses are cooperated with other approaches for supporting evidence.

Phenotypic analyses

Phynotypic data of influenza is generally collected from experimental assays, such as immunological data from hemagglutinination inhibition (HI) assays and micro-neuraminidase (MN) assays. Researchers have endeavored to reveal the antigenic evolution patterns and predict the antigenicity of influenza. For example, Plotkin *et al.* investigated the spatio-temporal evolution of influenza viruses by grouping strains into antigenic clusters (Plotkin *et al.*, 2002). Smith *et al.* pioneered to project HI data of influenza viruses into a two-dimensional map (antigenic map) using conventional multidimensional scaling, where the distances between viral isolates represent the HI measurements with least error (Smith *et al.*, 2004). The antigenic map is free-oriented, which provides a spatial layout representing relative positions of antigens and antisera. This computational approach to analyse binding assay data is called antigenic cartography.

Antigenic cartography does not only function as a visualization tool, but also allows for estimating the antigenic similarity between a pair of influenza viruses which are not directly tested under HI assays. Therefore, the antigenic cartography facilitates profiling the antigenicity of viral population with limited viral samples. In addition, the antigenic cartography can help reveal substitutions that are responsible for the transition of antigenic clusters, shedding light on the further investigations by experiments or structural analyses. More specifically, when the viruses present clusters, the amino acids that can differentiate between clusters are candidate substitutions for the change of antigenicity.

Figure 2.11 is the first antigenic map. The viral strains are human influenza A/H3N2 iso-

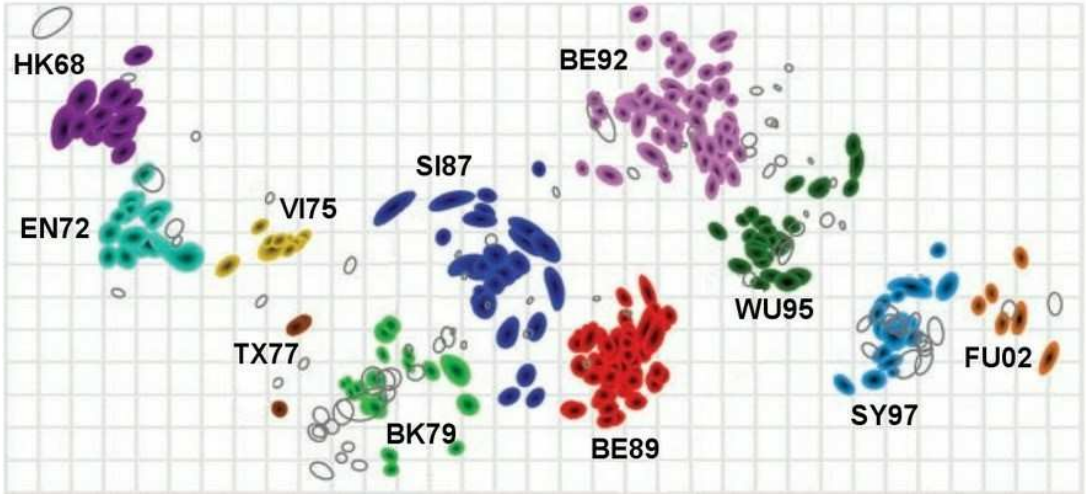


Figure 2.11: An example of antigenic map for human influenza A/H3N2 viral isolates from 1968 to 2003 (Smith *et al.*, 2004).

lated from 1968 to 2003. The distances between colored antigen strains and uncolored antisera strains represent the HI measurements with the least error. Strains with the largest antigenic distance are positioned in horizontal and colored by antigenic clusters. The antigenic clusters are determined with k-means clustering and represented byr the first vaccine strain within the cluster. For instance, the purple HK68 cluster at the left and the yellow FU02 cluster at the right. The transitions of antigenic clusters are approximately from left to right, corresponding to the evolution of antigenicity through time.

Besides, Smith *et al.* compared the genetic map and antigenic map of influenza A/H3N2. Results indicated that the antigenic evolution was punctuated compared to the consecutive molecular evolution. Following this original analysis on influenza A/H3N2, researchers have extended the cartography to other subtypes and even other viruses. For example, Koel *et al.* investigated amino acids responsible for the cluster transitions of influenza A/H3N2 and A/H5N1 viruses. All of the detected substitutions locate close to the receptor binding domain (Koel *et al.*, 2014). Barr *et al.* characterized the antigenicity of seasonal human influenza A/H1N1, A/H3N2 and B viruses, providing clues for recommending vaccine candidates for the northern hemisphere 2009–2010 influenza season (Barr *et al.*, 2010).

In addition, Smith's method was improved by cooperating with many algorithms. Given the fact that the HI data is often incomplete and noisy, Cai *et al.* improved the approach by completing the matrix first, named temporal Matrix Completion-Multidimensional Scaling (MC-MDS) (Cai *et al.*, 2010). This approach was then applied to both influenza A/H3N2 and A/H1N1 2009-pandemic strains, showing effectiveness and efficiency in constructing antigenic map (Cai *et al.*, 2012). Besides, Du *et al.* integrated a Bayesian network with antigenic cartography to predict the antigenic relationship between influenza viral strains and identify antigenic clusters (Du *et al.*, 2012).

Furthermore, many hybrid models have been proposed to investigate the relationship between genotype and phenotype (i.e. antigenicity) by integrating genetic sequence data and

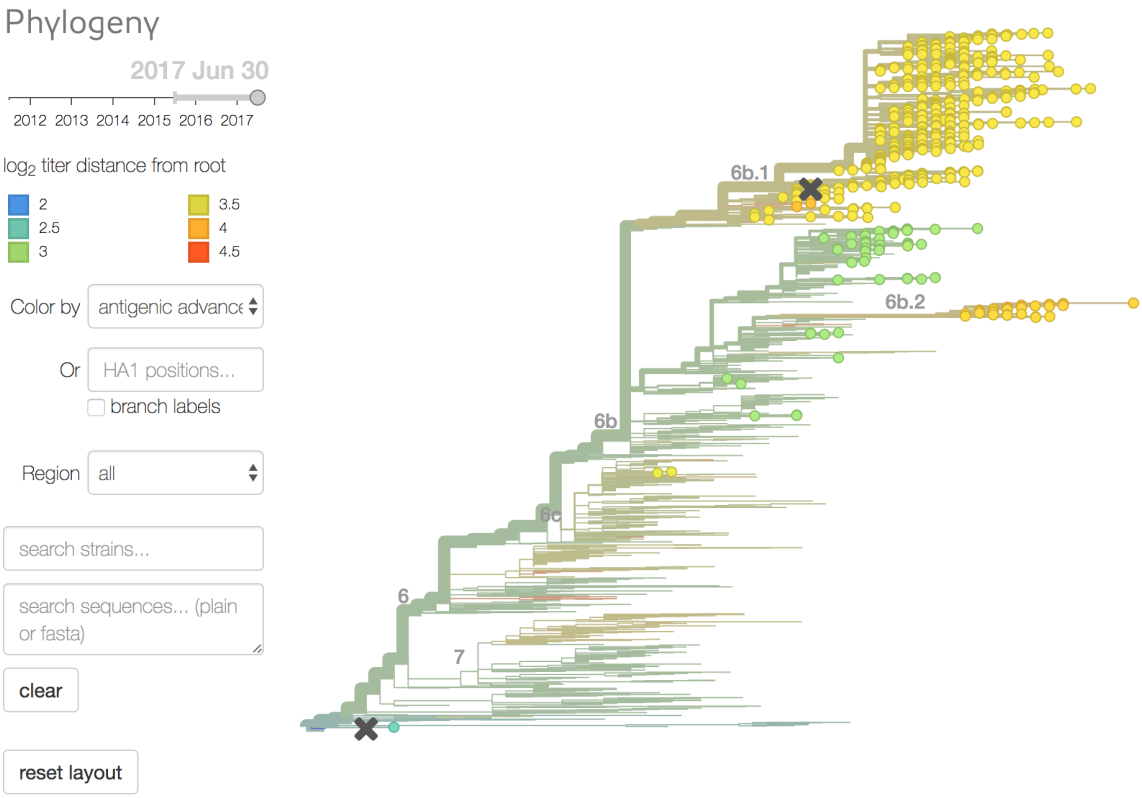


Figure 2.12: Antigenic evolution of influenza A/H1N1pdm (Neher and Bedford, 2015). *Branches are colored by antigenic log_2 titer distance from root.*

immunological assays. To be more specific, researchers endeavored to construct models to measure the antigenicity by sequence, which can accurately reflect the antigenic information measured by immunological data (typically HI data). For example, Lee and Chen tried to predict antigenic distances simply based on the number of substitutions in five epitopes (Lee and Chen, 2004). Liao *et al.* improved this approach by grouping amino acid according to their physicochemical properties and only substitutions across different classes would be counted (Liao *et al.*, 2008). The substitution counts were fed into a regression model to predict the antigenicity. Similarly, Huang *et al.* used Shannon entropy of residues as a predictor of antigenicity (Huang *et al.*, 2009a). Recently, Bedford *et al.* proposed a diffusion model to characterize both genetic and antigenic evolution of influenza viruses at the same time (Bedford *et al.*, 2014). Furthermore, Neher *et al.* integrated a phylogenetic-tree based model with a substitution model to predict HI titers from sequences, providing "nextflu" (an interactive real-time online tool) to track the evolutionary path of influenza viruses. Also, the "nextflu" enables the users to visualize antigenic distances on the phylogenetic tree (Neher and Bedford, 2015). Figure 2.12 is an example of phylogenetic tree showing antigenic evolution of influenza A/H1N1pdm, where the viral strains are colored by antigenic distances that are inferred from HI data. In addition, the online tool provides convenient way to color the viruses by the number of substitutions in different functional domains (e.g. epitopes, non-epitope regions, receptor binding sites), or by geographic region. With timely updates of new data, "nextflu" can serve as a real-time tracking

tool to monitor the evolution of influenza viruses.

Also, the antigenic cartography can help highlight substitutions responsible for the cluster transition, shedding light on further investigation by experiments or structural analyses. It has subsequently been applied to many other pathogens. For instance, researchers have applied the cartographic method to characterize the foot-and-mouth disease virus (Ludi *et al.*, 2014), the enterovirus (Huang *et al.*, 2009b), and the lyssavirus (Horton *et al.*, 2010). Typically, two dimensional cartography is applied in analyzing the antigenicity of influenza viruses, while three dimensional cartography could be more effective for other viruses (Ludi *et al.*, 2014). (Sun *et al.*, 2013) developed tools for visualizing the influenza HI data with both 2D and 3D antigenic cartography (Sun *et al.*, 2013).

There have been tools collecting the findings and highlight those interpretable residues of an input influenza protein sequence. For example, the influenza A/H5N1 influenza genetic changes inventory compiled by CDC supports international surveillance on the highly pathogenic strain (`https://www.cdc.gov/flu/avianflu/h5n1-genetic-changes.htm`). Besides, the FluSurver enables the researchers to quickly access the residue annotations given an influenza protein sequence (`http://flusurver.bii.a-star.edu.sg/`). Large-scale sequences available makes it possible to apply machine learning and deep learning techniques to extract evolutionary information from past strains. The reconstruction of evolutionary pathways helps to understand the evolutionary mechanism that allows adaptive substitutions controlling host tropism and pathogenicity.

Besides, great efforts have been made in annotating signatures for cross-species transmission and increased virulence, which can facilitate early detection of potential pandemic strains. For example, Miotto *et al.* highlighted a catalog of PB2 residues as human-to-human transmission markers by comparing a set of human-transmissible and avian isolates (Miotto *et al.*, 2008). Furthermore, Eng *et al.* utilized the reported genomic markers in literature to profile influenza viral proteins. With the help of those genomic markers, simple SVM could successfully classify the avian and human influenza viral strains (Eng *et al.*, 2014). Similarly, Qiang and Kou employed energy-feature-vectors into a protein sequences and used an artificial neural network (ANN) model to discriminant the avian influenza A viruses from the human influenza viral strains (Qiang and Kou, 2010).

However, those sequence-based computational models require solid validation against experimental data, particularly for revealing the context dependency of these substitutions before the models coming into service for informing policy-making. One debate on machine learning algorithms, especially the inner workings of neural networks, is that they perform like a black-box in the sense that they can approximate any function without giving any insights into the structure of the function being approximated. In terms of the application in biology, the estimated functions for prediction may not have substantial biological meaning.

Sequence models on co-mutation of influenza

Traditional experimental assays are challenging to design and high-priced, including site-directed mutagenesis, deep mutational scanning (DMS) and functional assays, for measuring the effects of site mutations and viral fitness are high-priced. Given the enriched viral sequences, the alternative by constructing computational models to detect mutation patterns is a more favorable approach for screening, providing clues for site-directed mutagenesis to identify how the mutation hinders or contributes to viral evolution.

Approaches using genetic sequences mainly detect mutations by reconstructing the adaptive evolution trajectory of influenza viruses. Antigenically relevant mutations, especially mutations lead to cluster transition are the main focuses (Akand and Downard, 2018). Smith *et al.* mapped the results of hemagglutinin assays into a phylogenetic tree using a maximum-likelihood approach to analyze the antigenic evolution of the HA1 protein of influenza A/H3N2 (Smith *et al.*, 2004). This work was extended by Bedford *et al.*, who analyzed both the antigenic and genetic evolution were analyzed using a diffusion model (Bedford *et al.*, 2014).

Statistical analysis usually depends on the changes in the frequency of residues. Shih *et al.* drew a frequency diagram of residues on the HA1 of influenza A/H3N2, finding that residues sharing similar patterns tend to cluster in antigenic sites (Shih *et al.*, 2007). Du *et al.* proposed a Naïve Bayes classifier based network to predict antigenic clusters (Du *et al.*, 2012). The co-occurring mutations observed at antigenic sites of HA could cumulatively drive the antigenic drift evolution of influenza viruses. Similarly, Bhatt *et al.* estimated the viral evolution rate by explicitly investigating the changes in the frequency of mutations through time (Bhatt *et al.*, 2011).

Alternatively, machine learning approaches have also been applied to analyze co-occurring mutation patterns in the viral sequences. For example, Substitution matrices have been used in many algorithms to formulate measurements (such as mutual information, information gain, joint entropy) to analyze the correlations between pairwise mutations (Baker and Porollo, 2016; Yip *et al.*, 2007; Xia *et al.*, 2009).

2.3.2 Structure-based computational analyses on influenza

Structural analysis, benefited from the enriched protein structure database, may compensate for that by analysing the impact of substitutions on protein structure, simulating the protein-ligand binding process and thereby optimizing therapies (Koday *et al.*, 2016). Structure-based computational analyses on influenza mainly focus on the receptor binding by HA and the drug resistance by NA. Methods include the ab initial fragment molecular orbital method, molecular docking and molecular dynamics (MD) simulation.

Homology modeling

The availability of protein structures is critical to computational analyses of protein functions and to modeling the protein-ligand or protein-protein interactions. However, constrained by

complexity of obtaining materials for the crystallization of protein structures, there is often delay to obtain an experimentally solved protein structure compared to the protein sequence. Besides, limited number of experimentally solved protein structures have called for the necessity of predicting protein structures from sequences.

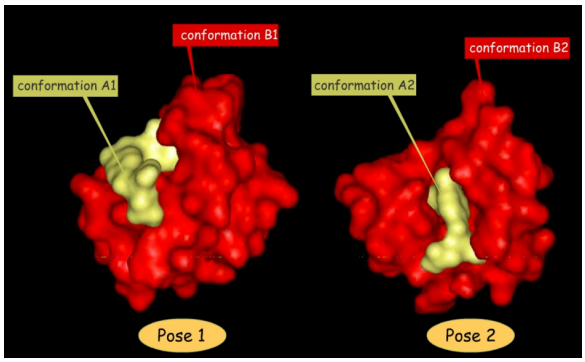
Homology modeling refers to the techniques to build a three-dimensional protein structure using available experimentally determined structures of homologous proteins as template (Krieger *et al.*, 2003). So it is known as template-based modeling. Homology modeling is based on the fact that proteins with similar amino acid sequences also share similar structures, especially for proteins belonging to the same family (Kaczanowski and Zielenkiewicz, 2010). Therefor, it is also called comparative modeling. Typically, the spatial arrangement information will be provided in the predicted protein structure, especially for important residues (Waterhouse *et al.*, 2018). The predicted structural information can shed light on new experiments, such as screening drugs and conducting site-directed mutagenesis.

Typical steps for homology modeling include the matching templates, aligning the protein sequences, generating and optimizing the backbone and the side chain, building *ab initio* loop, optimizing the model and measuring the quality of models. SWISS-MODEL provides online automated homology modeling services following the mentioned steps, along with visualization, evaluation, and interpretation of the modeling results (Arnold *et al.*, 2006; Guex *et al.*, 2009). The automated homology modeling service allows users with little expertise in computational biology to obtain reliable protein structural models for further analyses (Waterhouse *et al.*, 2018).

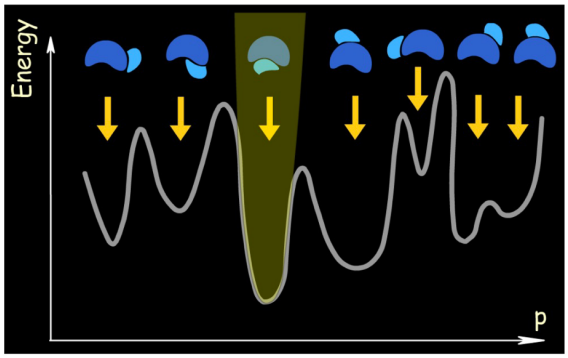
Molecular docking

Molecular docking has been an important tool for modeling protein-ligand interactions at the atomic level for a long time. The molecular docking facilitates the characterization of small molecules in the binding sites of target proteins, shedding light into the elucidation of fundamental biochemical process (Malathi and Ramaiah, 2018). Essentially, molecular docking aims to predict the ligand-receptor complex structure using computation methods.

A typical docking process includes predicting the ligand conformation (both position and orientation within binding sites) and the binding affinity for assessment. Therefore, the three-dimensional structure of the target protein is needed, which can be experimentally resolved (e.g. by X-ray crystallography or nuclear magnetic resonance) or computationally predicted (e.g. by homology modeling) (Salmaso, 2018). Sampling methods are applied to search the conformational space. Subsequently, the binding affinity scoring methods help to rank the generated conformations. Figure 2.13 illustrate the selection of protein-ligand docking pose and the binding affinity of the conformation ensemble model. Figure 2.13a shows examples of different poses of protein-ligand conformation. In a searching space, the binding energy of different conformation defined as Equation 2.4 is calculated (Figure 2.13b), which associates the native bound-conformation to the global minimum of the energy hypersurface (Salmaso and Moro, 2018). Generally, the conformation with the lowest global binding energy is selected as



(a) Examples of the conformation ensemble model.



(b) Illustration of the binding affinity in ensemble model.

Figure 2.13: Illustration of molecular docking (Salmaso and Moro, 2018).

the optimal docked conformation for further analyses.

$$\Delta G_{binding} = E_{complex} - E_{protein} - E_{ligand} \tag{2.4}$$

Molecular dynamics (MD) simulation

MD simulation is a technique for calculating the time dependent behavior of a molecular system, which has been widely used to understand relationship between macromolecular structure and molecular functions (Rapaport and Rapaport, 2004). The MD simulation can be used to facilitate X-ray crystallography or NMR experiments for determining the crystallized structure of target molecules. Besides, it produces plenty of information about the conformational changes of the biological molecules and the fluctuation of the system (Allen *et al.*, 2004). Therefore, it has been broadly applied to analyzing the structure and dynamics of target molecules and conformation ensembles.

Structure-based analyses on influenza mainly involves homology modeling, molecular docking and molecular dynamics simulation. For example, Su *et al.* combined molecular docking and molecular dynamics simulation to analyze conformation changes of drugs binding to NA of H7N9 under the mutation R289K, finding that this particular mutation may contribute to the enhanced drug resistance of influenza A/H7N9 (Su *et al.*, 2013). Similarly, Kannan and Kolandaivel; Pan *et al.* investigated the binding preference of A/H1N1 HA protein with different host cell receptors, giving insight on the enhanced virulence by the HA mutation D222G (Kannan and Kolandaivel, 2016; Pan *et al.*, 2012).

2.4 Summary

This chapter presents a comprehensive literature review on current knowledge about the virulence determinants of influenza, mainly covering the structure basis (Section 2.1), antigenicity (Section 2.2) and the state-of-the-art computational models for the receptor binding specificity,

antigenicity of influenza viruses. Sequence-based computational models and structure-based models are introduced in Section 2.3 respectively.

Empirical experiments have reported several particular genetic mutations contribute to the enhancement of activity in viral life, including binding with the host cells, fusion the membrane, entry into the infected cells, genome transcription and translation, the assemble and release of virus progeny, and the escape from immune responses. Such aspects and relevant mutations are virulence determinants of influenza viruses. The glycoproteins HA and NA have complementary functions, responsible for the virus entry into and release from host cells respectively. The functional balance of HA and NA are important for efficient transmission among hosts and may be a critical indicator for the pandemic potential of an influenza viral strain, also the main target for analyzing the antigenicity and receptor binding specificity of influenza viruses.

Computational analyses on influenza antigenicity and receptor binding specificity mainly include sequence-based models and structure-based models. Sequence-base models mainly focus on the evolutionary path and antigenic clusters, highlighting mutations probably leading to the change of antigenicity. Besides mutations on each protein, this dissertation also investigates mutation patterns across viral proteins. Instead of handcrafting features for each subtype using domain knowledge, a universal model for predicting the antigenicity of multiple influenza subtypes is proposed. However, those sequence-based computational models require solid validation against experimental data. Besides, the context dependency of the these substitutions need more investigation before the models coming into service for informing policy-making. Therefore, structure-based computational models are applied to measure how the residues substitution would affect the protein structure and viral functions.

Chapter 3

Detecting co-occurring mutations and virulence signatures of influenza viruses

The rapid evolution of influenza has limited the efficacy of flu vaccines. Various computational models have been proposed to uncover the mutational patterns and to predict mutations of rapidly evolving influenza viruses. A problem drawing attention is to identify sites that potentially contribute to the viral fitness or virulence of influenza.

This chapter aims to annotate virulence signatures and sites co-mutate with the signatures by analyzing the sequence pattern of influenza viruses. Section 3.1 summarizes known co-mutations in influenza viruses, which are mainly observations from functional assays, animal models and clinical data. For detecting the mutations contribute to viral fitness, a phylogenetic tree-based method is presented to infer pairwise co-mutations of intra-proteins in Section 3.2. Besides, residues cooperating or compensating each other for the same function tend to share a similar mutation pattern, which is not restricted to pairwise. Therefore, a sequential rule mining based method is proposed in Section 3.3 to explore co-occurring mutations at multiple sites, including mutations between different proteins.

Work of this chapter is mainly based on the publications Ivan *et al.* (2017); Chen *et al.* (2016); Zhou *et al.* (2019).

3.1 Known co-mutations in influenza viruses

The physical and genetic interactions among amino acids are the most elementary and decisive factors for the structure, function and evolution of proteins (Starr and Thornton, 2016). An amino acid mutation on a protein can neutralize or strengthen a mutation at another site, occurring together or in chronological order and jointly affecting the protein function. This phenomenon is generally named co-evolution or co-mutation (Codoñer and Fares, 2008). For example, Hopf *et al.* proposed an unsupervised statistical model for predicting the effects of mutations, reporting mutations with residue dependencies for about 7000 human proteins (Hopf *et al.*, 2017). The identification of co-occurring mutations and residue dependencies can help

Table 3.1: Co-mutations of influenza viral proteins reported in literature.

Protein	Mutations and effects	References
PB2	• E627K, D701N ⇒ high pathogenicity; • G590S, Q591R ⇒ high pathogenicity	Tscherne and García-Sastre (2011)
HA	• Antigenic sites ⇒ antigenic drift & antigenic escape; • Receptor binding sites ⇒ host cell receptor binding specificity	Shih *et al.* (2007); Du *et al.* (2008); Wu *et al.* (2017b)
NP	• R65K, D127E, E375G, R384G ⇒ nucleoprotein functionality & viral fitness & immune escape	Rimmelzwaan *et al.* (2004, 2005)
NA	• R222Q, V234M, H274Y ⇒ drug resistance	Abed *et al.* (2011); Durrant *et al.* (2015); Tscherne and García-Sastre (2011)
M2	• V27A, S31N ⇒ drug resistance	Durrant *et al.* (2015)
NS1	• R127K, V205I, N209D ⇒ increased virulence in mice	Pu *et al.* (2010)

uncover possible interactions among them and thereby improve our understanding of protein functions and the mutational dynamics. However, to discriminate the co-occurring mutations by chance from the co-mutations jointly affect a function remains a challenge.

Co-mutations also exist in influenza viral proteins. The co-occurring mutations on all viral proteins have not been systematically analyzed yet. Table 3.1 is a summary of co-mutations reported from experimental assays in literature.

Most researches focus on the surface glycoproteins HA and NA, which bear higher evolution rate and tolerance to mutations (Lyons and Lauring, 2018). Evidence has shown that co-occurring mutations of the HA protein can cumulatively facilitate the antigenic evolution of influenza (Shih *et al.*, 2007; Du *et al.*, 2008), enhance immune escape (Kryazhimskiy *et al.*, 2011), and alter the host cell receptor binding specificity of influenza viruses (Wu *et al.*, 2017b). Similarly, it has been observed that co-mutations on NA may affect its adaptive evolution and confer viral resistance to oseltamivir by altering the stability or enzymatic activity of protein (Durrant *et al.*, 2015; Lyons and Lauring, 2018).

There are sporadic examples of co-mutations reported for other influenza viral proteins. Rimmelzwaan *et al.* found that the NP with compensatory mutations demonstrates stronger flexibility to escape from cytotoxic T lymphocyte (Rimmelzwaan *et al.*, 2004, 2005). Mutations V27A and S31N on M2 associated with amantadine-resistance have been observed more frequently than predicted by chance, and strains with the dual mutations are more virulent to mice than with single mutation (Durrant *et al.*, 2015).

Co-mutations do not only exist within the same protein (intra-proteins), but also have been reported across different proteins (inter-proteins). It has been reported that the co-mutation of HA and NA protein may contribute to the formulation of an epidemic lineage (Chong and Ikematsu, 2018). The HA mutations can promote the replication of influenza viruses with the presence of mutation H274Y on NA (Ginting *et al.*, 2012). Pauly *et al.* identified mutations

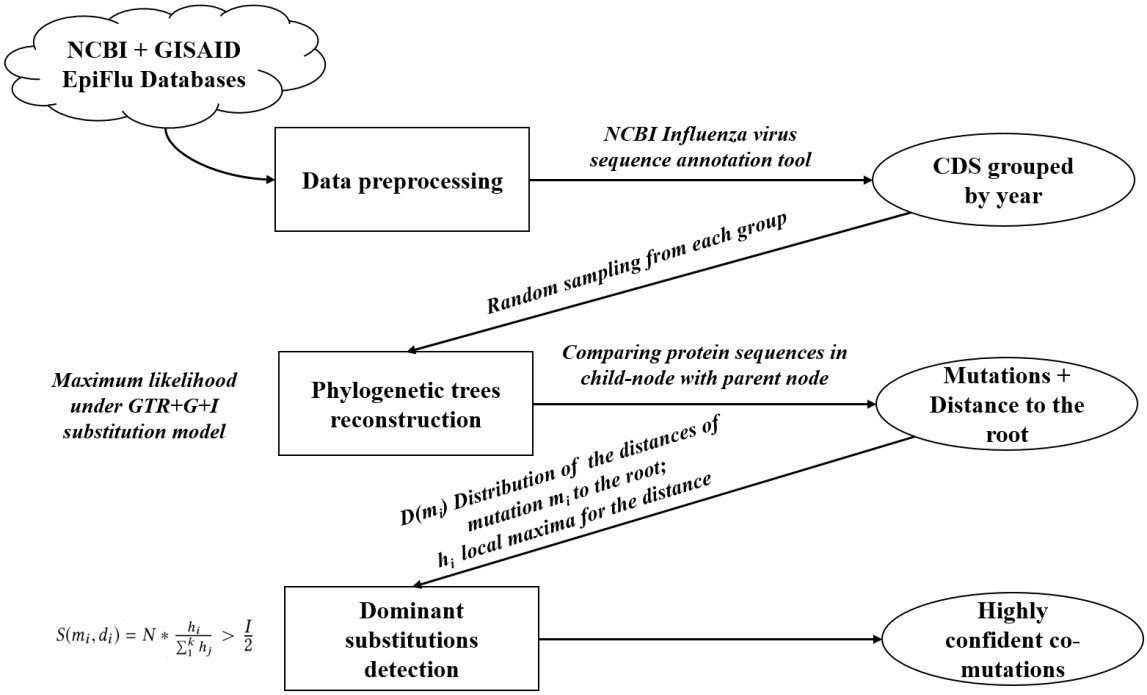


Figure 3.1: An overall architecture for inferring co-occurring mutations from reconstructed phylogenetic trees. *Phylogenetic trees of a influenza viral coding region are constructed using randomly sampled sequences from each year-group. Mutations and distance to the root are detected and smoothed by Gaussian kernel. Dominant mutations and highly confident co-mutations are further screened.*

T97I in PA and T123A in PB1 contribute to mediating mutagenic resistance and replication fidelity *in vitro*. Besides, co-mutation of multiple residues in the NS and PA can enhance the virulence of related influenza viruses in mice (Taft *et al.*, 2015; Choi *et al.*, 2017) while viruses containing mutations A36T in PA and H357N in PB2 have increased replication rate in human cells and enhanced virulence in mice (Zhu *et al.*, 2012a).

The mentioned co-mutations and the effects on protein functions are mainly measured by traditional experimental assays, which is costly. With enriched viral protein sequences, constructing computational models to detect co-occurring mutations is more promising.

3.2 A phylogenetic tree-based method for detecting pairwise co-mutations

This section presents a method based on reconstructing phylogenetic trees from coding DNA sequences(CDS) and protein sequences to infer co-mutations. This method takes the evolutionary structures in the sequences data into consideration and can capture most of the known HA sites contributing to major antigenic cluster transition.

3.2.1 Methods

Figure 3.1 shows the architecture for inferring co-mutations from reconstructed phylogenetic trees. First, coding regions and HA protein sequences of influenza A/H3N2 were extracted and grouped into years from NCBI and GISAID EpiFlu using the influenza viral sequence annotation tool (NCBI, 2014; Bao *et al.*, 2007; Shu and McCauley, 2017). Phylogenetic trees were then constructed using the randomly sampled sequences from each group. Mutations and their distances to the root were inferred and smoothed with a Gaussian kernel. Dominant mutations and site-pairs with a high confidence level would be selected afterwards.

The reconstruction of phylogenetic trees

All HA1 sequences of influenza A/H3N2 were obtained from NCBI and grouped by year bins, ensuring that the number of sequences in each bin was more than 20. Random HA sequences were sampled and aligned from each bin (Edgar, 2004). The phylogenetic tree was then reconstructed using the *GTR+G+I* substitution model with maximum likelihood (Schliep, 2011; Fitch, 1971). The process was repeated for k times to reconstruct sampled phylogenetic trees. Amino acid mutations were detected at each node (except for the root) by comparing with its parent's HA sequence. Each mutation was denoted as $m_i = A_a(p)A_b$, representing an amino acid A_a in the parent node to another amino acid A_b in the child node at a given site p in the sequence. The distances of each node to the root of the reconstructed phylogenetic tree were also recorded.

The detection of dominant mutations

Each mutation m_i in the phylogenetic trees were mapped to real numbers $d_1, d_2, ..., d_k$ representing the distances between the nodes where the mutations were observed and the root of phylogenetic tree $T_1, T_2, ..., T_k$. Afterwards, the distance distribution was smoothed with a Gaussian kernel density estimate (Sheather and Jones, 1991), followed by the detection of the peaks. Assuming $h_1, h_2, ..., h_k$ are the heights of the detected peaks for mutation m_i at distance to the root $d_1, d_2, ..., d_k$. A signal strength was defined as Equation 3.1 to measure the portion of observations that support the observed mutation m_i.

$$S(m_i, d_i) = N * \frac{h_i}{\sum_1^k h_j} \tag{3.1}$$

N represents the number of observations for the mutation m_i. If $S(m_i, d_i) > \frac{k}{2}$, the mutation m_i at distance d_i was taken as a dominant mutation.

The detection of confident site-pairs

For a dominant mutation m_p and the mutations locating at the descendant nodes within distance d^*, each pair of mutation (m_p, m_q) was observed N_q times. The distances of the ancestor

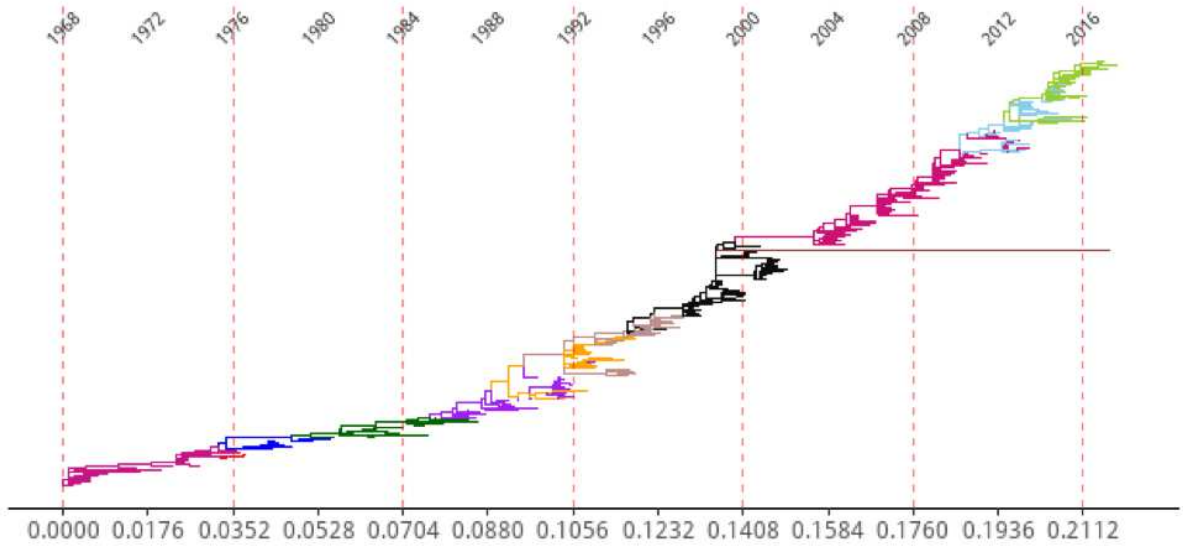


Figure 3.2: An Example of the phylogenetic tree of the HA sequences of influenza A/H3N2. *The distance to the root is mapped to time (year) at the top.*

mutation to the roots after smoothed by a Gaussian kernel were taken as the distances of the mutation pair to the roots. Similarly, $h_1, h_2, ..., h_k$ were the heights of the detected peaks for mutation pair (m_p, m_q) at distance $d_1, d_2, ..., d_k$. The support for mutation pair (m_p, m_q, d_i) was calculated as Equation 3.2. If $S(m_p, m_q, d_i) > \frac{k}{2}$, the site-pair was considered confident.

$$S(m_p, m_q, d_i) = N_q * \frac{h_i}{\sum_1^k h_j} \tag{3.2}$$

3.2.2 Results

To analyze the HA co-mutations of influenza A/H3N2, 1000 phylogenetic trees were generated using 510 random HA sequences from 36 year groups. Figure 3.2 shows a typical phylogenetic tree structure of the HA sequences of influenza A/H3N2, following a ladder-like structure. The distance to the root has been mapped to year at the top where the distance 0.0044 is approximately equivalent to 1 year.

Dominant mutations recovered from the longest path of the supported re-sampled trees are presented in Table 3.2. The number of dominant mutations are about uniformly distributed over the years of evolution. 80% of the dominant mutations locate at the epitope domains. Most mutations were detected at distance to root less than 0.14 or greater than 0.21, i.e. 1968-1999 and 2015 afterwards. Mutations responsible for antigenic cluster transitions as reported in Smith *et al.* (2004) are denoted bold and underlined. The early mutations are highly overlapped with the mutations responsible for the antigenic clusters transition. The lack of dominant mutations between 2000 and 2015 might be associated with the reduced binding affinity with human host cell receptors (Lin *et al.*, 2012).

The proposed method successfully identified a significant subset of dominant mutations responsible for the antigenic evolution of HA. Similar approaches may reliably uncover pairs

Table 3.2: Dominant mutations detected from the longest path of the phylogenetic trees.

	HK68			EN72		VI75	TX77	BA79					SI87	BE89		BE92	WU95	SY97			FU02		...					
Year	1968	1970	1972	1974	1976	1978	1980	1982	1984	1986	1988	1990	1992	1994	1996	1998	2000	2002	2004	2006	2008	2010	2012	2014	2016
Distance to root	0.0000	0.0088	0.0176	0.0264	0.0352	0.0440	0.0528	0.0616	0.0704	0.0792	0.0880	0.0968	0.1056	0.1144	0.1232	0.1320	0.1408	0.1496	0.1584	0.1672	0.1760	0.1848	0.1936	0.2024	0.2112
Epitope A			G144D T122N	T126N S145N	N137Y	G146S	N133S P143S	D144V	G124D	T131A	S133D	G135K	D124G N145K	K135T G124S D133N G142R	V144I Y137S	I144N		K145N				N145S			
Epitope B			T155Y N188D	S193N Q189K	G158E	K156E T160K	Q197R		V163A S159Y K189R Y155H		N193S E156K E190D	R189S S157L	R197Q	V196A	E158K K156Q	T192I S186G					K158N				
Epitope C		D275G		I278S N53D	K50R	N54S		K307R		K299R		T276N	S278N	G275D	N276K					G50E					
Epitope D		V242I R207K		I217V		V244L	D172G V217I	I213V N248T N173K			L226Q	I214T Q226L T214I	L226I I121T G172D	I226V T121N		D172E		V226I							
Epitope E	V78G			D63N T83K	M260I	I62K				F94Y E82K K83E	T262N			N262S	K62E R57Q			E83K							
Non-epitope	N31D D31N	E479G	L3F L331I	F3L		D2N I347V		R453K N2K V384L				K450R		D375N		R452K E386G			N375D			D489N			

Table 3.3: Top 10 confident site-pairs with varying distance thresholds d^*. *The darker the cell, the more frequent the site-pair appears in every ranking.*

Ranking	d*							
	0.0011	0.0022	0.0033	0.0044	0.0055	0.0066	0.0077	0.0088
1	142->196	484->142	3->145	3->145	3->145	189->226	189->145	189->145
2	212->312	406->142	144->220	124->145	83->145	83->145	83->145	83->145
3	121->196	171->142	213->137	144->220	124->145	3->145	124->145	155->145
4	144->213	213->137	276->226	347->220	347->220	189->145	189->226	156->226
5	128->347	248->193	219->145	197->145	106->220	124->145	156->226	225->142
6	142->347	142->347	217->145	106->220	276->226	156->226	144->220	124->145
7	144->3	3->331	278->246	219->145	144->220	276->226	197->145	189->226
8	62->3	188->331	83->189	172->145	156->226	347->220	172->145	144->220
9	156->3	193->50	347->220	83->145	189->226	106->220	3->145	278->145
10	212->45	225->50	144->226	142->45	189->145	219->145	225->142	144->142

of HA sites that have ancestor-predecessor relationship. The site-pairs within distance $d^* = 0.0011, 0.0022, 0.0033, ..., 0.0088$ were presented in Table 3.3, which corresponded to $\frac{1}{4}, \frac{2}{4}, \frac{3}{4}, ...,$ and $\frac{8}{4}$ years respectively. The top 10 confident site-pairs with varying distance threshold d^* were listed in Table 3.3. The co-mutations appearing over 4 times in each cases were highlighted in darken cells. Site 145 was observed to be affected by many other sites, including site 3, 83 and 124 in non-epitope, epitope E and A respectively. Co-mutations among site 145, 189 and 226 also appeared frequently. Site 145 and 189 located in the receptor binding sites in the HA protein, and were involved in the EN72→VI75 antigenic cluster transition. Site 226 was reported to be determinant for the binding specificity with host cells (Rogers *et al.*, 1983).

3.3 Association rules and sequential rules mining of mutations at multiple sites

Co-mutational relationship is not restricted to pairwise. In this section, an association rule based method was first applied to explore the co-mutations of multiple sites on the HA protein of influenza A/H1N1, A/H3N2 and B viruses. The results of multiple subtypes suggested that mutations on the HA protein could characterize the antigenic evolution of influenza viruses. Furthermore, this method was extended and improved by integrating time-constraints. The improved method was applied to all proteins of influenza A/H1N1, A/H3N2 and B viruses to detect intra-mutations. Some obtained co-mutations were supported by existing biological findings. Besides, the function and evolution of a viral protein can also be affected by the other proteins (Poole *et al.*, 2004). Therefore, the improved method was then applied to explore the inter-mutations, i.e. co-mutations across proteins. The results indicated that the mutation NA: N369K could function coordinately with HA: S185T and S451N to enhance the host cell binding and immune escaping. The finding was consistent existing evidence shown that the co-mutations of HA and NA might contribute to an epidemic lineage (Chong and Ikematsu, 2018).

3.3.1 Methods

To detect the co-occurring mutation patterns of influenza proteins, three modules have been constructed as illustrated in Figure 3.3. The first module *Seq2Trans* formulates the problem, converting protein sequences to a list of mutational transactions. The second module *Trans2Rules* integrates algorithms for detecting the association and sequential rules of mutations. Results from those algorithms are taken together to obtain a more concise subset of co-occurring mutation pattern.

Problem formulation: from protein sequences to mutational transactions

Figure 3.4 clarifies how to formulate the problem, focusing on converting the protein sequences to a transaction database *(transDB)* and a sequential database *(sequentialDB)* that can be fed into the downstream pattern recognition module. With protein sequences from consecutive year Y_m till Y_n, we selected samples from each year randomly with replacement. Repeat the process for K times to amplify the transactions for training. The pairwise comparison was applied to sequences s_{Y_i} and $s_{Y_{i+1}}$ from two adjacent years to get a binary encoding of mutations, where 0 and 1 represented "non-mutations" and "mutations" respectively. Finally, the index of all 1, i.e. the positions of mutations in a protein sequence, were outputted as a transaction record T_{i+1}, forming a transaction database *transDB*. For a dataset with sequences from year m to year n, we would get $(m-n) \times K$ transactions, with empty transactions included. In each transaction, mutation sites have been sorted in ascending order. Afterward, transactions from every year are concatenated with timestamps as a mutation sequence. Mutation sequences resulted from

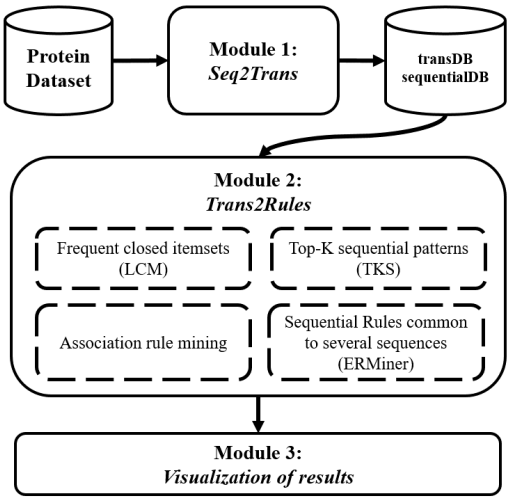


Figure 3.3: An overall architecture for uncovering the co-occurring mutation patterns from influenza protein sequences. *Module 1 Seq2Trans constructs a transaction database (transDB) and a sequential database (sequentialDB) from the inputting protein sequences. Module 2 Trans2Rules mines co-occurring rules and sequential rules from the transDB and sequentialDB. Module 3 visualizes the rules as a directed graph, treating each mutation site in results as a node.*

the K rounds of sampling consist of a sequential database (*sequentialDB*) for further sequential pattern analyses.

Similarly, to analyze the co-occurring patterns of inter-proteins P_α and P_β, we first concatenate the sequences from the same strain and then formulate the problem as stated to get the binary mutation patterns between the proteins. Patterns observed across proteins are exclusively kept.

Association and sequential rule mining

A transaction database $T = [T_1, T_2, ...T_l]$ obtained from the module *seq2trans* is defined as a list of items T_i representing mutation sites. To detect the co-occurring mutations in the same year, we first obtained frequent closed itemsets using LCM (Uno *et al.*, 2004) and then selected the top k association rules using TopKRules (Fournier-Viger *et al.*, 2012) higher than a confidence level. The confidence is defined as Equation (3.3).

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{3.3}$$

where the $|supp(X)|$ denotes the number of transactions where all the items (i.e. mutation sites) of X appears.

Furthermore, we constructed the sequential database $S = [s_1, s_2, ..., s_K]$, where the sequence s_i includes mutations along with timestamps from consecutive years Y_m till Y_n. A mutation sequence is defined as Equation (3.4), where $T^i_{m+1}(i = 1, 2, 3, ...K)$ represents a mutation list detected in year Y_{m+1} in the i^{th} round of experiment. We selected $K = 5000$ in our experiment

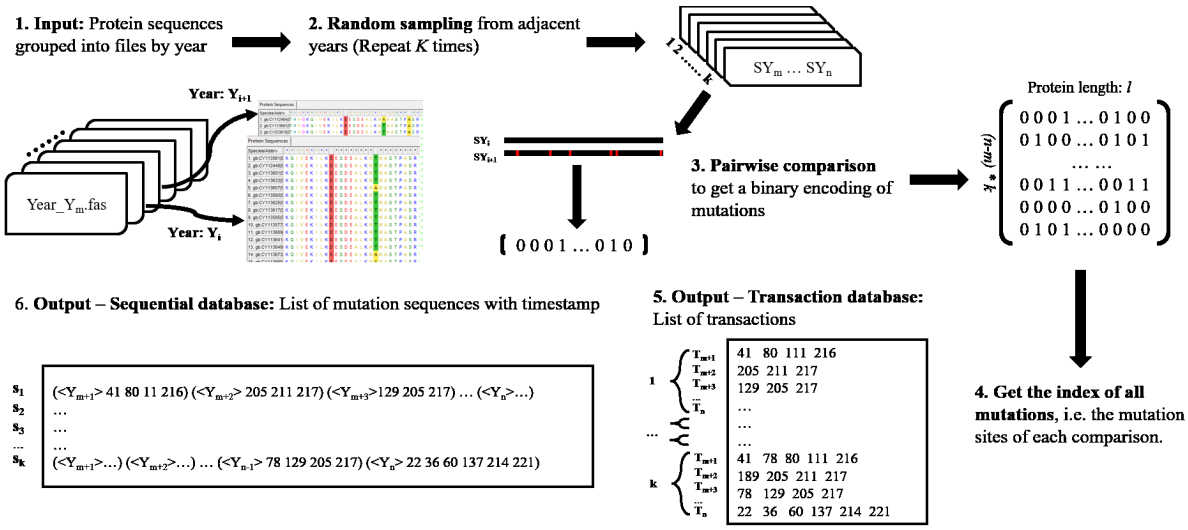


Figure 3.4: Problem formulation: from protein sequences to transactions (*Seq2Trans*). *This module outputs a transaction database transDB and a sequential database sequentialDB from protein sequences in year range Y_m and Y_n.*

and constructed a sequential database with 5000 mutation sequences.

$$s_i = (T^i_{m+1}, T^i_{m+2}, ..., T^i_n) \tag{3.4}$$

A sequential rule $X \Rightarrow Y$ represents the sequential relationship between mutations in set X and set Y, meaning that if mutations in X occur in a mutation sequence, the mutations in Y will occur afterwards. The reliability of a rule is measured with support and confidence, defined as Equation (3.5).

$$conf(X \Rightarrow Y) = \frac{|sids(X \Rightarrow Y)|}{|sids(X)|} \tag{3.5}$$

where $|sids(X)|$ denotes the number of sequences where all the items of X appears. A **T**op-k **N**on-redundant **S**equential rule mining algorithm (**TNS**) has been applied in our analyses, which is an approximate but efficient algorithm eliminating redundancy in results (Fournier-Viger and Tseng, 2013). Moreover, with the availability of timestamps of sequences, we applied an extension of sequential pattern mining with time constraints (Fournier-Viger *et al.*, 2008) to focus on mutations within Δt years. Finally, the intersection of association rules and top-k TNS were kept as the output mutation patterns.

Algorithm 1 describes how to get mutation rules from a batch of protein sequences S with timestamps ranging from y_1 to y_2. The inner loop (line 5-11) deals with a batch of randomly selected protein sequences from continuous years, outputting $y_2 - y_1$ transactions. Those transactions with timestamps form a sequential instance seq_i for mutational events. Repeat the process for K times to obtain enough records in *transDB* and *sequentialDB*, which can be fed into frequent items mining and sequential pattern mining algorithms respectively. In our study, $K = 5000$. LCM was used to mine co-occurring mutations *armRules* in *transDB* with the support $sup = 8000$ and confidence level $conf = 0.95$. Top-10 TNS was used to mine sequential

mutations *seqRules* in *sequentialDB* with the same confidence level. The intersection of *armRules* and *seqRules* was taken as the output.

Algorithm 1 Pseudocode for mining rules from historical protein sequences with timestamps

```
function SEQ2RULE(S, y1, y2, K):   ▷ S: protein sequences with timestamps ranging from y1 to y2; K: Sampling rounds
    define transDB = [ ]
    define sequentialDB = [ ]
    for i = 1 to K do:
        for y = y1 to y2 − 1 do:
            define seq_i = { }
            s1 = rand(S[y]); s2 = rand(S[y+1])
            muts = compare(s1, s2)
            trans[y+1] = getIndex(muts)
            transDB.append(trans[y+1])
            seq_i.update({y+1: trans[y+1]})
        sequentialDB.append(seq_i)
    armRules = LCM(transDB, sup, conf)
    seqRules = TNS(sequentialDB, conf, k)
    rules = intersection(armRules, seqRules)
    return rules
```

To visualize the rules with format $X \Rightarrow Y$, items in X and Y were treated as source nodes and target nodes respectively. The edge widths were scaled by the maximum support among the rules covering the source node and target node. Clustering coefficient is calculated to measure whether the nodes tend to cluster together. The clustering coefficient for a node v_i in graph $G = (V,E)$ is calculated as Equation 3.6, where the numerator represents the number of edges within the neighborhoods of v_i, while the denominator is the total number of edges in a complete graph formed by them. The average clustering coefficient of all nodes is taken as the global clustering coefficient $C = \frac{1}{|V|}\sum_{C_i \in V} C_i$.

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E|\}}{k_i(k_i-1)} \tag{3.6}$$

Furthermore, we plotted the change of residue frequency $f(A_i)$ on the mutation sites involved in the output rules. $f(A_i) = \frac{count(A_i)}{count(S_y)}$, where $count(S_y)$ is the number of protein sequences in year y and $count(A_i)$ is the number of amino acid A appeared at site i. The correlation of frequency trend between A_i and A'_j is calculated as Equation 3.7, functioning as a supportive measurement of our results.

$$\begin{aligned} corr(A_i, A'_j) &= \frac{cov(f(A_i), f(A'_j))}{\sigma_{f(A_i)}\sigma_{f(A'_j)}} \\ &= \frac{E[(f(A_i) - \mu_{f(A_i)})(f(A'_j) - \mu_{f(A'_j)})]}{\sigma_{f(A_i)}\sigma_{f(A'_j)}} \end{aligned} \tag{3.7}$$

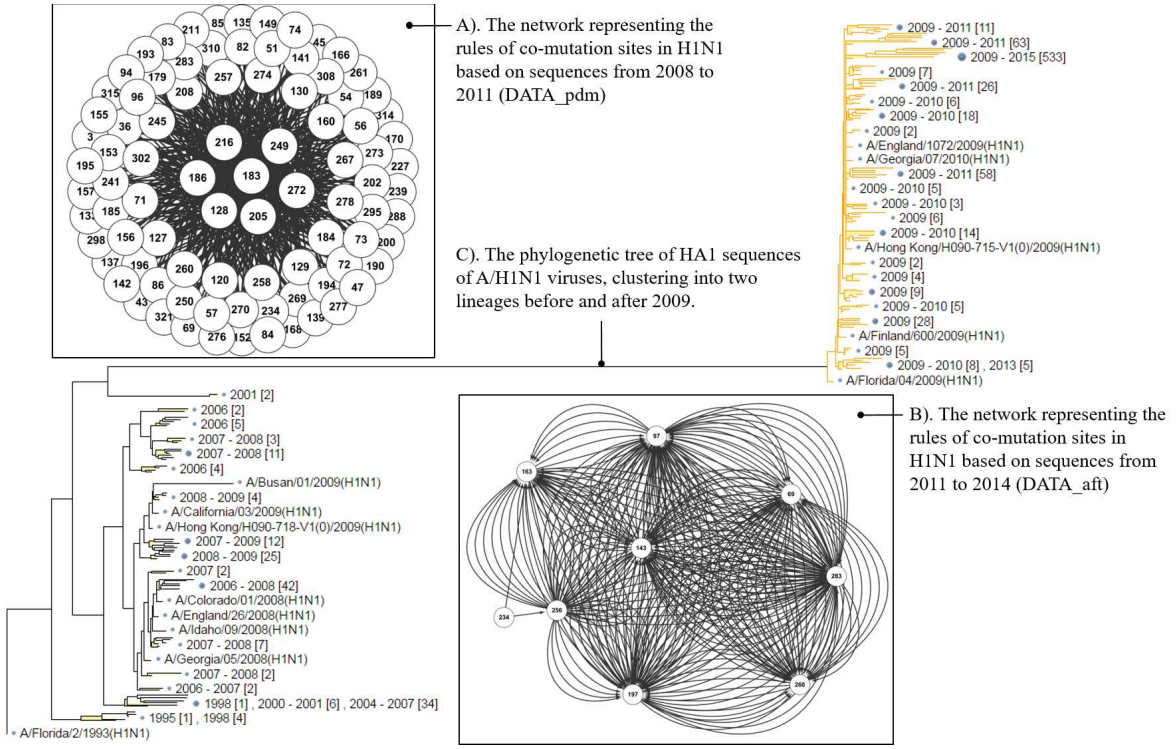


Figure 3.5: The phylogenetic tree and co-occurring mutations of influenza A/H1N1.

3.3.2 Results

Co-occurring mutations on the HA1 protein of influenza A/H1N1, A/H3N2 and B viruses

In this section, only the co-occurring mutations on the HA1 protein, a subunit of HA containing the main receptor binding sites, were analyzed using association rule mining. Results presented in this subsection is based on the publication Chen *et al.* (2016).

The HA1 sequences of human influenza A/H1N1, A/H3N2 and B viruses were retrieved from NCBI as of 31 Jan, 2016 (NCBI, 2014). To make sure mutations can be obtained by comparing HA1 sequences from two conjunction years, sequences from 1976 to 2015 for influenza A/H1N1, from 1968 to 2015 for influenza A/H3N2 and from 1975 to 2015 for influenza B viruses were used. Mutational transactions were obtained from using the *seq2trans* module presented in Figure 3.4. The outputted *transDB* was fed into the association rule mining algorithm. Besides, the HA1 phylogenetic trees for the three subtypes were constructed using the neighbor-joining method and mPAM distance provided by the NCBI tool "Influenza Virus Sequence Tree" based on (Bao *et al.*, 2008). Figure 3.5, 3.6 and 3.7 present the detected co-occurring mutations and the HA1 phylogenetic trees of influenza A/H1N1, A/H3N2 and B viruses respectively.

As shown in Figure 3.5-C, the HA1 phylogenetic tree of influenza A/H1N1 evolved into a different lineage after 2009. Hence, the HA1 sequences of influenza A/H1N1 was divided into two datasets: sequences from 2008 to 2011 was denoted as *DATA_pdm*; sequences from 2011 to 2014 was denoted as *DATA_aft*. The detected co-occurring mutation networks during the influenza A/H1N1 2009 pandemic and after the pandemic are presented in Figure 3.5-A and B.

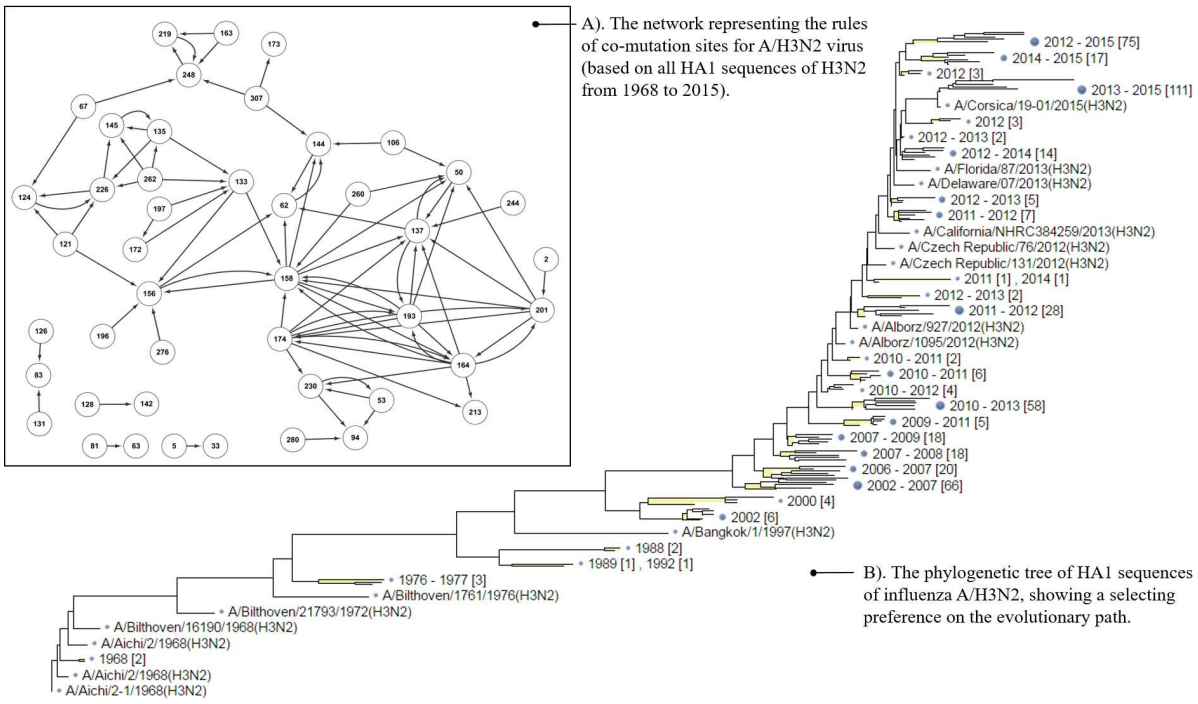


Figure 3.6: The phylogenetic tree and co-occurring mutations of influenza A/H3N2.

It was not surprised that the co-occurring mutation network for the sequences during pandemic (2008-2011) was more dense than after the pandemic (2011-2014). Sites with the highest degree were 128, 183, 186, 205, 216 and 249, potentially driving the antigenic evolution of influenza A/H1N1 to the pandemic strains. In contrast, only a few sites were observed co-mutating frequently.

For influenza A/H3N2, the HA1 phylogenetic tree shows a ladder-like structure in Figure 3.6-B, which indicates the HA1 mutations of influenza B are under more selective pressure. Thus, the support of HA1 co-occurring mutations between two conjunction years (obtained from *seq2trans*) were lower than that of influenza A/H1N1 and B viruses. The most frequent co-mutation sites were 50, 53, 62, 137, 144, 145, 155, 156, 158, 189, 244, 260 and 275. All of them located at the epitope regions of influenza A/H3N2.

The HA1 phylogenetic tree of influenza B viruses (Figure 3.7-C) showed that the B viruses diverged into two co-circulating lineages, namely the B/Yamagata and B/Victoria lineage. Hence, the co-mutations of the two lineages were mined separately (Figure 3.7-A and B). Most HA1 co-mutations of influenza A/H1N1 and A/H3N2 located at the epitope domains (Table 3.4), indicating that those mutation sites might be under strong selective pressure by the human immune system. For influenza B/Yamagata and B/Victoria, only one of the detected mutation site for each lineage located at the epitopes. When mapping the mutation sites onto the HA1 protein structure, it could be observed that the other detected mutation sites also located nearby the epitope domains (Figure 3.8).

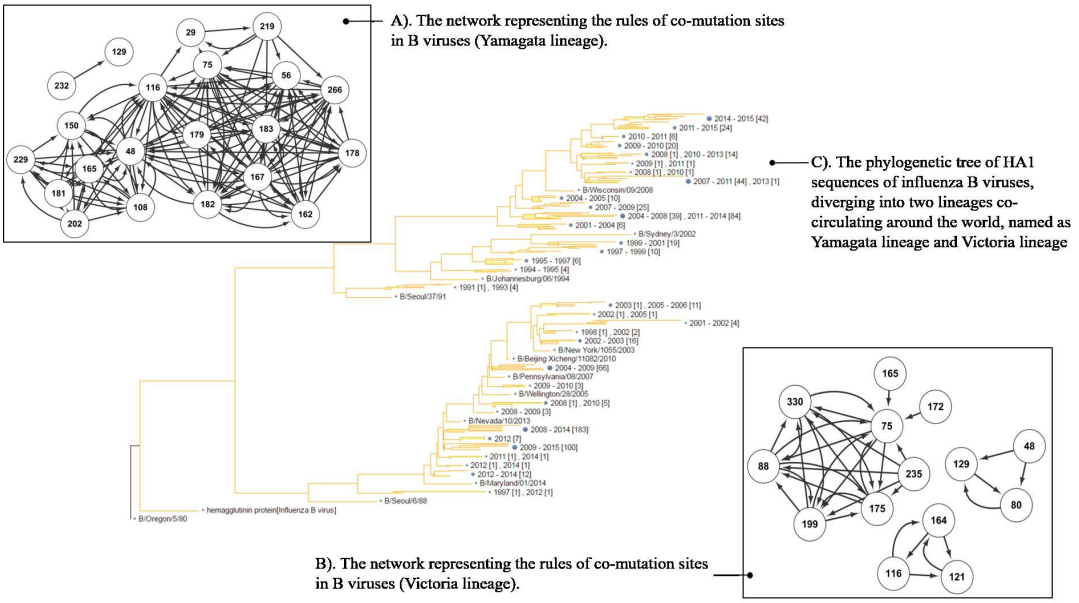
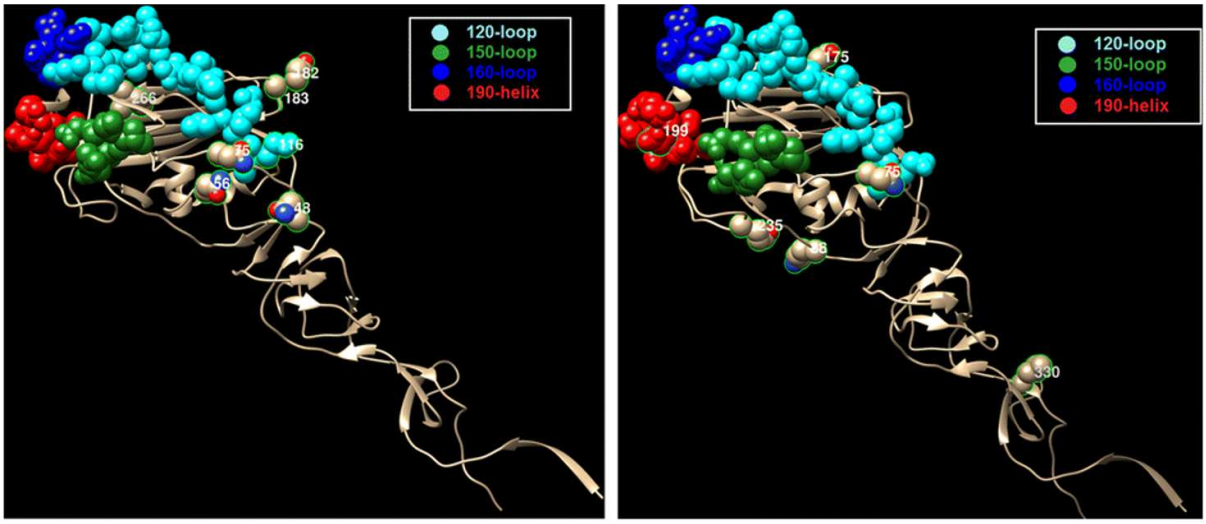


Figure 3.7: The phylogenetic tree and co-occurring mutations of influenza B viruses.

Table 3.4: An overview of extracted rules on the HA1 protein sequences of influenza A/H1N1, A/H3N2 and B viruses using association rule mining.

Dataset	**Detected sites**	**#residues at antigenic sites/Total number of sites**
A/H3N2	50, 53, 62, 127, 144, 145, 155, 156, 158, 189, 244, 260, 275	13/13
A/H1N1(pdm)	128, 183, 186, 205, 216, 249, 272	5/7
A/H1N1(aft)	69, 97, 143, 163, 197, 256, 260, 283	4/8
B/Yamagata	48, 56, 75, 116, 182, 182, 266	1/7
B/Victoria	75, 88, 175, 199, 235, 330	1/6



A). Influenza B/Yamagata lineage

B). Influenza B/Victoria lineage

Figure 3.8: Mapping co-mutation sites detected onto HA1 protein of influenza B/Yamagata and B/Victoria (PDB ID: 4NRJ) (Ni *et al.*, 2014). *The red, blue, cyan and green colors represent 120-loop, 150-loop, 160-loop and 190-helix respectively. The four clusters have been found to cause antigenicity variation, together forming a single large antigenic site with overlapping epitopes. The co-mutation sites are circled and marked with corresponding numbers.*

Sequential intra-mutations of all influenza proteins

Furthermore, to get a more concise subset of co-mutations, the association rules mining based method was improved by integrating the timestamps into mutational transactions. A sequential mutational transaction dataset (sequentialDB) was obtained as stated in Section 3.3.1. To analyze the co-mutations and sequential mutations on all influenza proteins, all protein sequences of human influenza A/H3N2, A/H1N1 and B viruses full-genome set were retrieved from the influenza research database (IRD) as of 30 Jan 2019 (Squires *et al.*, 2012). The influenza A/H1N1 strains were classified as "pandemic" and "non-pandemic", denoted as pH1N1 and epH1N1, based on the sequence similarity with the strains causing the swine flu in 2009. There were 11413, 6127, 258 and 7187 strains of influenza A/H3N2, A/pH1N1, A/epH1N1, and B strains with full-genome set. Only the complete protein sequences were kept for further analyses. Table 3.5 provides an overview of the viral segments, the number of complete protein sequences and lengths for influenza A/H3N2, A/pH1N1 , A/epH1N1 and B viruses.

First, the frequent co-occurring mutations and sequential mutations in each protein were detected using the module *trans2rules* (Figure 3.3). Table 3.6 shows the detected mutations in each protein. Furthermore, we plotted the network of the detected co-occurring mutations and sequential mutations, treating each site as a source node or target node. Clustering coefficient for each network was analyzed to measure the degree to which nodes tend to cluster. Mutations were mapped into protein structures accordingly.

Table 3.5: An overview of the complete protein sequences used for analyses.

	A/H3N2			A/pH1N1			A/epH1N1			B		
Segment	**Protein**	**# Complete**	**Length**	**Protein**	**# Complete**	**Length**	**Protein**	**# Complete**	**Length**	**Protein**	**# Complete**	**Length**
1 PB2	**PB2**	13124	759	**PB2**	6508	759	**PB2**	1430	759	**PB2**	7891	770
2 PB1	**PB1**	11459	757	**PB1**	6453	757	**PB1**	1436	757	**PB1**	7889	752
	PB1-F2	10446	90				**PB1-F2**	1244	57			
	PB1-N40	11466	718	**PB1-N40**	6443	718	**PB1-N40**	1438	718			
3 PA	**PA**	13113	716	**PA**	6512	716	**PA**	1437	716	**PA**	7969	726
	PA-X	13481	252	**PA-X**	6739	232	**PA-X**	1383	252			
4 HA	**HA**	13280	566	**HA**	6626	566	**HA**	1427	565	**HA** Vic	4790	584
										HA Yam	2684	585
5 NP	**NP**	13213	498	**NP**	6550	498	**NP**	1447	498	**NP**	7943	560
6 NA	**NA**	13241	469	**NA**	6581	469	**NA**	1291	470	**NA**	6681	466
										NB	6665	100
7 MP	**M1**	11598	252	**M1**	6556	252	**M1**	1435	252	**M1**	6627	248
	M2	11580	97	**M2**	6546	97	**M2**	1431	97	**BM2**	3144	109
8 NS	**NS1**	11142	230	**NS1**	6516	219	**NS1**	1225	230	**NS1**	4490	281
	NS2	11564	121	**NS2**	6523	121	**NS2**	1964	121	**NS2**	4502	122
Year Range	1968-2018			2009-2017			1995-2017			1993-2016		

Table 3.6: Co-occurring mutations detected in each protein of influenza A/pH1N1, A/epH1N1, A/H3N2 and B viruses. [1]***Green*** *sites locate at the region of PB2 binding with both PB1 and NP.* [2]***Yellow*** *sites locate at PB2cap, binding the cap of a host pre-mRNA molecule.* [3]*HA protein numbering conforms with the H3 numbering system, trimming signal peptides.* [4]***Red*** *sites locate at the epitope domains of the HA protein.* [*]*The HA mutations of influenza A/H3N2 that are consistent with the results using a phylogenic-tree-based method (Ivan* et al., *2017) with distance < 0.77, and the results using solely association rule mining (Chen* et al., *2016)*

Segment	**Protein**	**Influenza A/pH1N1**	**Influenza A/epH1N1**	**Influenza A/H3N2**	**Influenza B**
1 PB2	PB2	**(54, 66, 195, 293,** 731) (299, **456) (344, 354)**	(107, **453, 667)**	-	(115, 301, **382, 383, 397, 442, 468,** 483, **639)**
2 PB1	PB1	-	(177, 327, 372, 375, 383)	(54, 110, 591)	(38, 57, 60, 176, 211, 213, 508, 752)
	PB1-F2	(100, 330, 343, 361, 362)	-	(12, 23, 87)	-
3 PA	PA	-	(272, 277, 321)	(27, 262, 332, 348)	(258, 271, 352, 428, 700)
4 HA (H3 numbering)	HA	**(74, 164, 295)** (97, **143, 197) (162, 216) (163, 256) (185,** 451)	**(82, 94, 208, 266,** 415)	**(164*, 174*, 193*, 201)**	**(108, 150, 166, 230)**
5 NP	NP	(377, 454)	(189, 284, 309, 353) (373, 408)	(52, 280, 312) (343, 423)	(17, 171, 233, 374, 513)
6 NA	NA	(13, **264, 270,** 314) (34, 321, **432)** (77, 81, **449)** (241, 369)	(64, **173)** (234, 382) **(332, 450)**	**(172, 265,** 399, 437)	(42, 148, 198, 219, 235, **389, 392, 436)**
	NB	-	-	-	(21, 31, 43, 86) (53, 57,
7 MP	M1	(80, 192, 230)	-	(219, 227, 230, 166, 205)	-
	M2 (BM2)	(12, 13, 21, 43)	(48, 50)	(24, 52, 89)	(35, 69, 86, 105) (73, 77)
8 NS	NS1	(2, 55, 125, 131)	-	(135, 139)	(97, 108, 110, 114, 115, 119, 137, 194, 279)
	NS2	(2, 83)	-	(20, 40, 55, 89)	(94, 111)

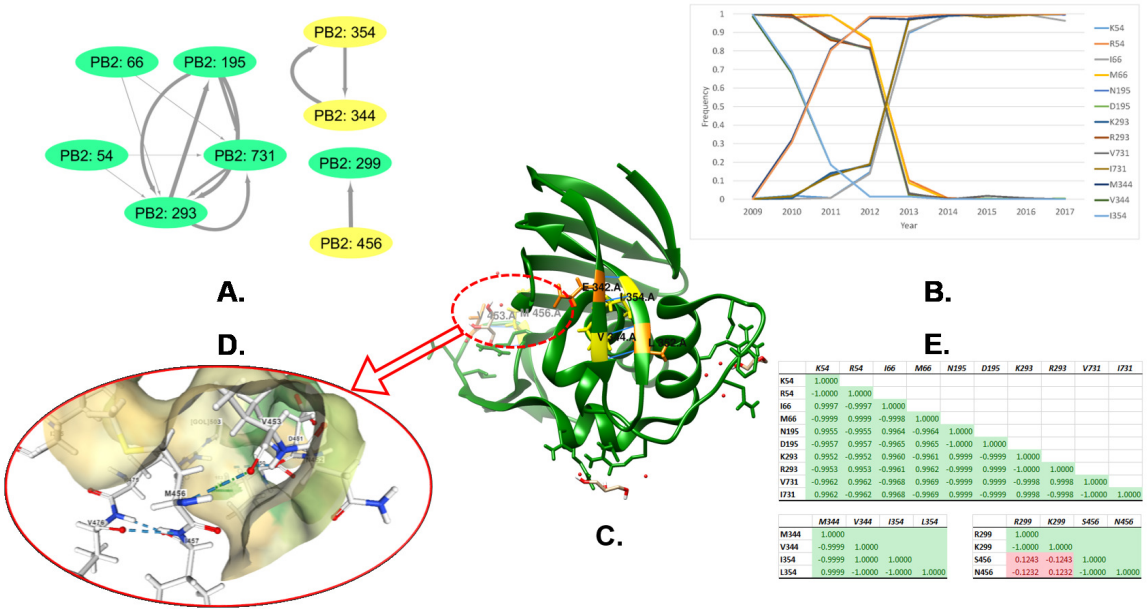


Figure 3.9: Co-occurring mutations on the PB2 of influenza A/pH1N1 viruses. *(A) Groups of sites with co-occurring mutations. The edge width is scaled by the maximum support among the rules. Sites locating at the PB2cap subunit with structure available are colored yellow. (B) The change of residue frequency on the detected sites. Sites within the same group tend to share similar change pattern of residue frequency. (C) PB2cap in complex with the cap analog m7GTP. Residues on the detected sites are colored yellow. Residues forming hydrogen bonds with them are colored orange. Hydrogen bonds are presented with a blue link. (D) The M456 locates at a ligand pocket, forming a hydrogen bond with V453 and Pi interactions with the ligand (colored green).*

Highlights of intra-mutations of influenza A/pH1N1

For influenza A/pH1N1, the most striking co-occurring mutation group locates at the PB2. There are three groups of co-occurring mutations detected in PB2, among which the sub-network (54, 66, 195, 293, 731) has a clustering coefficient of 0.717 (Figure. 3.9A). There is no direct connection among site 54, 66 and 195, but the site 293 tends to mutate following the mutation at any of them. Unfortunately, the subunit protein structures are only available for position 1-37, 318-484 and 535-759 (Consortium *et al.*, 2018). Functional domains of PB2 have not been fully understood yet. Evidence indicated that PB2 contains two regions binding with NP and PB1 respectively (Poole *et al.*, 2004). Sites 54, 66, 195 and 293 locate at the region that can bind with both NP and PB1. As for the site 731, it locates at the binding region with PB1. Other two pairs (344, 354) (299, 456), except 299, locate at the cap-binding domain PB2cap, which binds the cap of a host pre-mRNA molecule. Sites are mapped onto the PB2cap from A/California/07/09 in complex with the cap analog m7GTP (PDB: 5WOP) (Severin *et al.*, 2016; Berman *et al.*, 2000; Burley *et al.*, 2018), as shown in Figure. 3.9C. Residues V344 and I354 form hydrogen bonds with L352 and E342 respectively. Figure 3.9D depicts the GOL ligand pocket where the residue M456 forms a hydrogen bond with V453 (Rose *et al.*, 2018).

We further looked into the residue frequency of those sites (Figure. 3.9B) and calculated the correlation of residue frequency between sites (Figure. 3.9E). The frequency line of M344

almost overlaps with L354. K54, I66, N195, K293, and I731, sharing a similar increasing trend in the frequency. Precisely, the (K54, I66) and (N195, K293, I731) overlap within the group. The results indicate that mutations R54K, M66I, D195N, R293K, V731I, V344M , and I354L may contribute to the fitness of viral strains. The observation is supported by the high correlation of residue frequency (54, 66, 195, 293, 731) and (344, 354) within groups. Interestingly, the rule $456 \Rightarrow 299$ has high support and confidence, but the correlation of residue frequency between the two sites remains low. The second interesting co-mutation pattern is on the HA protein, where 12/14 sites locate at the epitope domain. The site 84 (H3 numbering) is likely to mutate after the mutation at signal peptide site 4 mutates. The residue at HA2 451 is likely to be affected by the mutations on HA1. The rule $NS1 : 2 \Rightarrow 125$ also needs to be highlighted. The mutation D125G on NS1 is reported to benefit viral host adaptation and switching among multiple species. Among the human influenza A/pH1N1 strains, the residue D125 is still dominant. G125 appeared sporadically during 2009-2011 and then not detected.

Frequent sequential HA mutations across subtypes

Similarly, we analyzed the mutation patterns for the influenza A/epH1N1, A/H3N2 and flu B viruses. Similar mutation patterns have been detected. Co-occurring mutation patterns on the HA protein of influenza A/epH1N1, A/H3N2 and B viruses are presented in Figure 3.10. There are five sites on HA of influenza A/epH1N1 tend to mutate simultaneously, namely (82, 94, 208, 266, 415), forming a complete directed network. All the sites locate at the epitope domains, except 451 locating at HA2. This is consistent with the observation among influenza A/pH1N1 strains. The mutations T82K, Y94H, R208K, T266N, and L415I have similar residue distribution, especially since 2003. Furthermore, we compared the HA mutation patterns of B/Yamagata and B/Victoria lineage with a lower threshold, shown in Figure 3.11. There is an insertion N178 at the HA protein of B/Yamagata lineage, site index based on the full-length HA protein of B/Yamagata lineage. Interestingly, the co-occurring mutation pattern of B/Victoria lineage is almost overlapping with that of the whole B virus population, except site 194 at the B/Victoria lineage. For the influenza B/Yamagata lineage, mutations on (63, 95, 144) form a fully connected directed network with a lower threshold.

For the influenza A/H3N2 viruses, all the detected sites are on the epitope domains. Residue distribution of sites 164, 174 and 193 share the closest pattern, and have been reported by a phylogenetic-tree-based method with root-distance < 0.77 (Ivan *et al.*, 2017). Residues at site 201 vary among Asp(D), Phe(F), Asn(N) and Ser(S). The mutation S201N occurred from 1971 to 1989. During 1990-2001, the Ser dominated the viral population again. Subsequently, the mutation S201F was detected and gradually dominated the viral population, with Ser sporadically detected. For the influenza B viruses, the frequency of residues Y166 and D230 overlaps after 2000. The mutation P108A quickly spread among viral population during 2004-2007, but the Pro(P) remains dominant afterward. Mutations G230D and N166Y share a similar frequency distribution. N166Y mutated slighted later than G230D.

Most work analyzed only the co-occurring mutations on the HA protein of influenza A/H3N2.

Table 3.7: Comparing the detected mutations on HA protein of influenza A/H3N2 with pure association rule mining method (Chen *et al.*, 2016) and a phylogenetic tree-based method (Ivan *et al.*, 2017). **Sites overlapped with the results from association rule mining and the phylogenetic-tree-based method are denoted* ***bold*** *and* <u>underlined</u> *respectively.*

	CoSeqMuts (Conf = 0.9, Sup = 8000)		Association Rule Mining	Phylogenetic-tree-based method
	HA1	HA2	HA1	HA1
Group 1	**<u>137, 164, 174, 193, 201, 230</u>**	-	2, 50, 53, 62, 94, **106, 137,** 144, 158, **164, 174, 193, 201,** 213, **230,** 244, 260, 280	62, <u>106,</u> 121, 135, <u>137,</u> 138, 142, <u>144,</u> 145, 156, 158, 189, <u>164,</u> 172, <u>174,</u> <u>193,</u> <u>230,</u> 276
Group 2	**<u>106, 144</u>**	18		

We compared our results with a confidence level 0.90 and support threshold 8000, denoted as CoSeqMuts, with a pure association rule mining method (Chen *et al.*, 2016) and a phylogenetic-tree-based method (Ivan *et al.*, 2017) in Table 3.7. All the sites detected CoSeqMuts on HA1 are also reported either by the pure association rule method or the phylogenetic-tree-based method. It is indicated that our method could report a more concise subset of co-occurring mutations with a lower false positive rate. Besides, it is worth noting that our results cover both the HA1 protein and HA2 subunits, which may cooperate on the viral binding with host cells after cleavage.

Frequent inter-protein mutations across HA and NA

Aside from analyzing the co-occurring mutations in intra-proteins of influenza, our method can also be applied to analyze the mutations across proteins. We analyzed the inter co-occurring mutations in HA and NA, the two membrane glycoproteins of influenza viruses. HA and NA are known to be physiologically interlinked, responsible for the functional balance between the viral binding to and cleavage from host cells. Evidence has shown that HA and NA may have been evolving in a coordinated way, where mutations on one protein may compensate for or enhance the other (Poole *et al.*, 2004).

We presented the intersection of rules with confidence over 0.95 from ARM and top-10 TNS (Table 3.8). Of particular interest is that the human influenza A/pH1N1 showed low HA avidity for glycan receptors and weak enzymatic activity. Our detected mutations (NA:369 $\Leftrightarrow$ HA: 185, 451) and (NA: 41 $\Leftrightarrow$ HA: 69, 143, 216, 260) among the influenza A/pH1N1 strains have indicated that the sites 41 and 369 on the neuraminidase are likely to facilitate the mutations on hemagglutinin, and vice versa. The mutations NA: N369K, HA: S185T and HA: S451N took up a similar quota among the human influenza A/pH1N1 population since 2009. The correlation scores among the frequency of NA: K369 , HA: T202 and HA:N451 are higher than 0.99. Interestingly, the mutation NA: N369K has been reported to enhance surface expression of neuraminidase in an oseltamivir-resistant viral strain (Abed *et al.*, 2014), while the variant with mutation HA: S185T is more likely to avoid immune recognition (Jiménez-

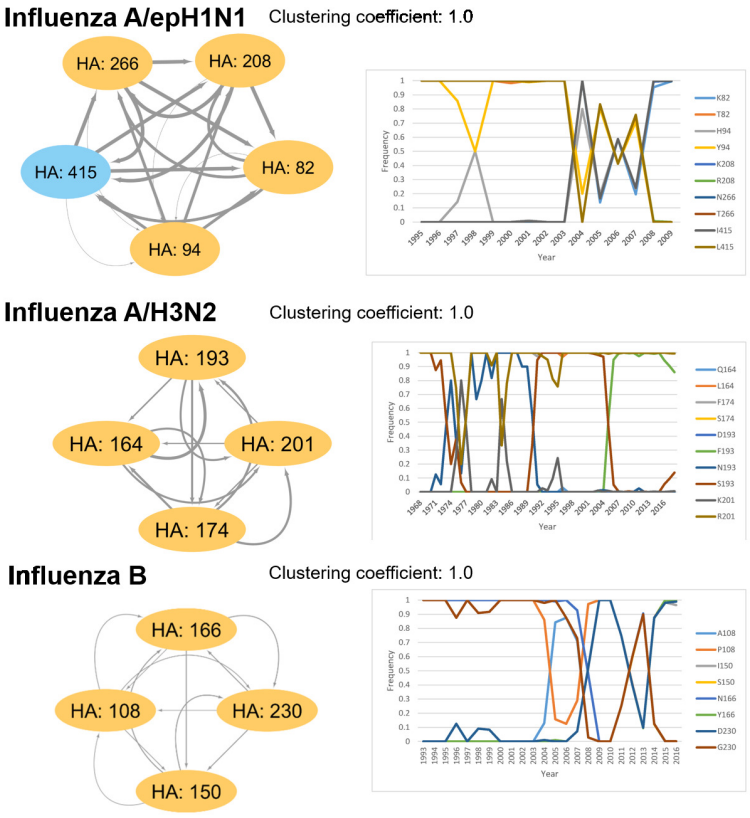


Figure 3.10: Co-occurring mutation patterns on the HA protein of influenza A/epH1N1, A/H3N2 and B viruses. *Sites are indexed based on the H3 numbering (Burke and Smith, 2014). Sites at the epitope domains are colored orange.*

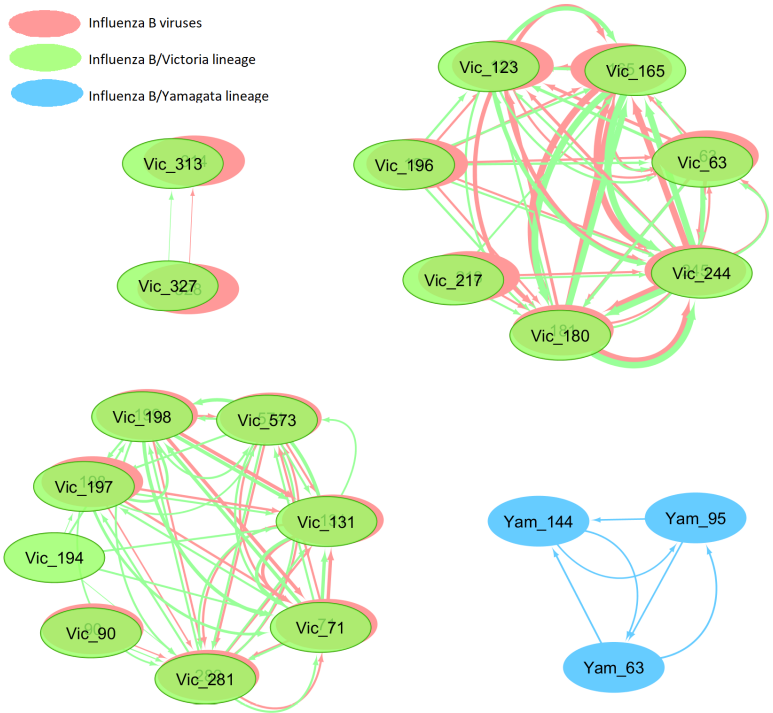


Figure 3.11: Comparison of co-occurring mutation patterns on the HA protein of influenza B/Victoria and B/Yamagata lineage.

Table 3.8: Frequent co-occurring mutations across the HA and NA protein of influenza A/pH1N1 and A/H3N2

Subtype	**A/H3N2**		**A/pH1N1**		**B**	
Protein	HA	NA	HA[1]	NA	HA[1]	NA
Group 1	135[2]	400	185, 451	369	230	187
Group 2			69, 143, 216, 260	41		

[1] The HA sites are indexed using the H3 numbering system.
[2] Sites colored red are on the epitope domain of the HA protein.

Alberto *et al.*, 2013). The mutation S451N occurs at the HA2 subunit. The conformation changes in HA2 may contribute to the fusion between viral and host cell membranes. The previous discussion has been focusing on the compensatory role of N369K with the mutations H275Y and V241I on the neuraminidase. Our results indicate that the mutation N369K may function coordinately with mutations S185T and S451N on the hemagglutinin to enhance the viral binding with host cells while escaping the immune system and the antiviral drug. Unlike the NA mutation N369K, where the residue Lys took the place of Asn gradually among the human influenza A/pH1N1 population, residues at site 41 vary. G41R happened during 2011-2012, but the residue Gly(G) dominated the population again after 2012. The frequency of HA:S86 and HA:N277 has the same trend with residue NA:G41.

3.4 Summary

Predicting the circulating influenza viral strains to guide the manufacturing of flu vaccines remains an open challenge. Sequence based computational models for detecting mutation patterns is promising for profiling the evolution of influenza viruses.

A phylogenetic tree based method was proposed for pairwise co-mutations detection. The method was applied to the HA protein of influenza A/H3N2, and successfully identified dominant mutations responsible for the antigenic evolution of HA. Furthermore, an association rule based method was proposed for co-mutations at multiple sites. The method was applied to the HA protein of influenza A/H1N1, A/H3N2 and B viruses, finding that the co-mutations on HA protein can characterize the evolution of influenza viruses. Noting that the mutation drift of a viral protein is remarkably sensitive to the genomic context and may be affected by the other proteins, the work was improved to screen a concise subset of co-occurring and sequential mutations on all proteins, and across different proteins of influenza viruses.

Computational models provide an alternative to observing whether the mutation hinders or contributes to the viral fitness among the population. Most mutations are deleterious to the virus; some residues dominant the viral population intermittently. It should be interesting to further look into those sites, doing site-directed mutagenesis analysis to measure the effect of a mutation.

The sequence-based computational models proposed in this chapter was intended for, but

not limited to detecting the mutation patterns of influenza viral proteins. The observation of co-occurring mutations, or highly correlated amino acids can facilitate the determination of protein functional domains, shedding light on the design of experiments for further analyses. What can we infer from viral mutations? How distinct are the mutant viral strains from current circulating strains or vaccine candidates regarding to the antigenicity? The next chapter intends to answer those questions by inferring viral phenotype from viral genomic variation. More specifically, a sequence model is proposed to predict the antigenicity of influenza viruses.

Chapter 4

Predicting the antigenicity of influenza viruses from the HA sequences

Antigenic variation is the major way for infectious antigens (including protozoans, bacteria and viruses) to circumvent host immune responses and re-infect the hosts, mainly by altering surface proteins (Webster, 1999). Given the fact that an effective vaccine strain should be antigenically close to the circulating strain, it is significant to monitor and characterize the antigenicity of the circulating influenza viral strains quickly and accurately.

This chapter aims to characterize the antigenicity of influenza viruses from HA sequences only. The proposed sequence model was able to map genetic sequence information to phenotype through supervised learning, and should be applicable to other datasets where the protein sequences are labeled with phenotype.

Chapter 4 is organized as following: Section 4.1 introduces existing computational models for characterizing and predicting the antigenicity of influenza viruses; In Section 4.2, a **C**ontext-**Free** **E**ncoding **S**cheme is introduced for converting protein sequences and sequence pairs into a numeric matrix that can be fed into any downstream learning method for further analyses; In Section 4.3, CFreeEnS is applied to predict the antigenicity of diverse subtypes of influenza viruses from HA sequences. Work of this chapter is mainly based on the publications Zhou *et al.* (2018b,c).

4.1 Methods for predicting the antigenicity of influenza viruses

At present, sequencing of viral genome and testing immunological reaction to candidate vaccines have been a routine work in influenza surveillance. With the enriched viral sequences, a variety of computational models have been proposed to predict antigenic variants of influenza, including genotypic analysis and phenotypic analysis. Accurate prediction of antigenic variants can help facilitate producing efficient vaccines before the spreading of the emerging strains. Genotypic analysis focuses on identifying antigenically important residues by analysing the genetic sequence patterns of influenza. Phenotypic analysis usually involves analyzing im-

munological data (e.g. HI data) to measure the antigenic similarity between the target strain and the reference strain. Another critical task is to infer the antigenic similarity of viral pairs that are not directly compared by HI assays. Integrative analysis endeavor to explore the relationship between genotype and antigenic phenotype by incorporating genetic sequence data and immunological data.

4.1.1 Genotypic Analyses

Given the fact that the antigenicity of influenza viral strains is heavily related to the five epitope domains of HA protein, the number of HA mutations on epitope regions has been taken as an important measurement for the antigenic variation of influenza viruses. Gupta *et al.* proposed $p_{epitope} = \frac{m}{n}$ to measure the antigenic distance between two viral strains, where the m stands for the number of mutations among the n sites in the active epitope. Similarly, $p_{all-epitope} = \frac{m_1}{n_1}$ and $p_{sequence} = \frac{m_2}{n_2}$ are sometimes used as alternatives, measuring fraction of mutations among all epitope regions or the whole sequence. The $p_{epitope}$ worked well in measuring the efficacy of vaccines to influenza A/H3N2 between 1971 and 2004 (Gupta *et al.*, 2006).

Phylogenetic analysis is the most traditional yet useful approach to infer evolutionary history from genetic sequence data. For example, Bedford *et al.* constructed a phylogenetic tree for the HA of influenza A/H3N2 from 1986 to 2002, revealing that the HA has been underlying continuous selective pressure. Also, a single predominant trunk lineage has been highlighted through time, with side branches persisting 1-5 years before extinction (Bedford *et al.*, 2011). Approaches using genetic sequences mainly focus on reconstructing the adaptive evolution path of influenza viruses, along with which mutation patterns are explored, especially antigenically important substitutions.

Conventionally, selective pressure is measured as the ratio of non-synonymous mutations (dN) and synonymous mutations (dS), i.e. dN/dS (Bush *et al.*, 1999; Yang, 2000). It has been observed that adaptive evolution (positive selection) usually bears $dN/dS > 1$. However, the dN/dS ratio is not robust because it is sensitive to deleterious substitutions. To be more specific, when the sequences of isolated viruses are sampled once before been purified by selection, the dN/dS ratio will be greatly affected (Harvey, 2016). To reduce the effects of deleterious substitutions, Bush *et al.* suggested to separate the calculation of internal codon-specific dN/dS from terminal branches (Bush *et al.*, 1999). It has been used as the main approach to investigate individual mutations. Codons on the HA highly preserved, or with dS/dN ratios indicating positive pressure have been detected, providing insights into the analysis of protein functional domains.

Besides, researchers tried to explore the relationship between residues, especially substitutions. The assumption is that residues can cooperate or compensate each other to maintain specific protein function. Tusche *et al.* proposed an improved dN/dS-based methods to detect residue clusters under high positive selection by incorporating spatial distances between protein residues (Tusche *et al.*, 2012). Generally, the approaches based on the dN/dS ratio perform

well in detecting antigenically important positions. However, those approaches are sensitive to the recurrent selection, however, is frequently observed in the evolutionary analyses of human influenza viruses. Statistical analysis usually depends on the changes in the frequency of residues (Lee and Chen, 2004; Gupta *et al.*, 2006; Liao *et al.*, 2008; Shih *et al.*, 2007). Shih *et al.* drew the frequency diagram of residues on the HA1 of influenza A/H3N2, finding that residues sharing similar patterns tend to cluster in antigenic sites. Similarly, Bhatt *et al.* detected a set of mutations by explicitly investigating the frequency of mutations throughout the evolution of human influenza A viruses (Bhatt *et al.*, 2011). Alternatively, machine learning approaches have also been applied to analyze co-occurring mutation patterns in the viral sequences. Substitution matrices (Liao *et al.*, 2008; Yip *et al.*, 2007; Xia *et al.*, 2009; Huang *et al.*, 2009a; Gao *et al.*, 2011; Simonetti *et al.*, 2013; Baker and Porollo, 2016) is serving as the foundation for the machine learning models, based on which various measurements (such as mutual information, information gain, joint entropy) are calculated to measure the correlations between pairwise mutations. For example, Xia *et al.* constructed a site-transition-network (STN) based on mutual information and predict mutations on HA protein, achieving an accuracy of 70% (Xia *et al.*, 2009).

4.1.2 Phenotypic analyses

Multidimensional scaling (MDS) and antigenic cartography form the basis to analyze immunological data, typically HI data (Tzeng *et al.*, 2008). Smith *et al.* pioneered to project HI data of influenza viruses into a two-dimensional map (antigenic map) using conventional MDS, such that the distances between strains represent the HI measurements with least error (Smith *et al.*, 2004). The antigenic map is a free oriented map, which provides a spatial layout representing relative positions of antigens and antisera. This computational approach to analyse binding assay data is called antigenic cartography. It has subsequently been applied to many other pathogens. Following this original analysis on influenza A/H3N2, researchers have extended the cartography to other subtypes, including influenza A/H1N1/pdm09 (Bedford *et al.*, 2014; Garten *et al.*, 2009), influenza B (Barr *et al.*, 2010), avian influenza A/H5N1 (Koel *et al.*, 2014) and even other viruses (e.g. Foot-and-mouth disease virus serotype A (Ludi *et al.*, 2014)).

Given the fact that the HI data is often incomplete and noisy, Cai *et al.* improved the approach by completing the matrix first, named the temporal Matrix Completion-Multidimensional Scaling (MC-MDS) model. This model has been applied to both influenza A/H3N2 and A/H1N1 2009-pandemic strains, showing effectiveness and efficiency in constructing antigenic map (Cai *et al.*, 2012). Antigenic cartography quantifies and visualizes antigenicity relationship between viral strains. The conserved substitutions between clusters represent mutations responsible for the change of viral antigenicity. However, the phenotypic analysis does not include structural information, nor infer the impact of structural changes to antibody binding capability.

4.1.3 Hybrid models integrating genotypic and phenotypic analyses

Hybrid models endeavour to infer antigenic characteristics measured by immunological data from viral sequences. To be more specific, researchers tried to find a way to measure the antigenicity by sequence, which can accurately reflect the antigenic information measured by immunological data (typically HI data). Lee and Chen tried to predict antigenic distances simply based on the number of substitutions in five epitopes (Lee and Chen, 2004). Liao *et al.* improved this approach by grouping amino acid according to their physicochemical properties and counting only substitutions between classes (Liao *et al.*, 2008). The substitution counts were fed into a regression model to predict the antigenicity. Similarly, Huang *et al.* used Shannon entropy of residues as a predictor of antigenicity (Huang *et al.*, 2009a). Recently, Bedford *et al.* proposed a diffusion model to characterize genetic and antigenic evolution of influenza simultaneously (Bedford *et al.*, 2014). Neher *et al.* further proposed a phylogenetic-tree based model and the substitution model to predict HI titers from sequences. An interactive real-time online tool named nextflu was constructed to visualize antigenic properties on the phylogenetic tree (Neher and Bedford, 2015). In addition, the online tool provides convenient way to color the viruses by the number of substitutions in different functional domains (e.g. epitopes, non-epitope regions, receptor binding sites), or by geographic region. With the updates of new data, it can be a useful tool to surveil the evolution of influenza viruses.

4.2 CFreeEnS: a Context-Free Encoding Scheme of protein sequences

In this section, a context-free encoding scheme (denoted as CFreeEnS) is introduced for converting protein sequences and protein sequence pairs into numeric matrices while keeping the most important information of protein sequences. Section 4.2.1 formulates a typical supervised learning problem, showing a typical pipeline of computational modeling. Section 4.2.2 presents how CFreeEnS works in detail, while Section 4.2.3 provides a general method for evaluating a computational model. Work of this section is based on publications Zhou *et al.* (2018b) and Zhou *et al.* (2018c).

4.2.1 A typical pipeline of computational modeling

Figure 4.1 shows a typical pipeline of computational modeling can be divided into four modules: data retrieval, feature engineering, modeling, and evaluation. With a satisfying model performance, the trained model is promising to be deployed for applications. In the feature engineering module, the raw non-numeric dataset is encoded by a numeric matrix so that it can be fed into the downstream learning algorithms, i.e. the modelling module. The performance of a computational model mainly relies on the cooperation of the two parts, i.e. the upstream encoding scheme and the downstream learning algorithm. The effectiveness of a learning algorithm

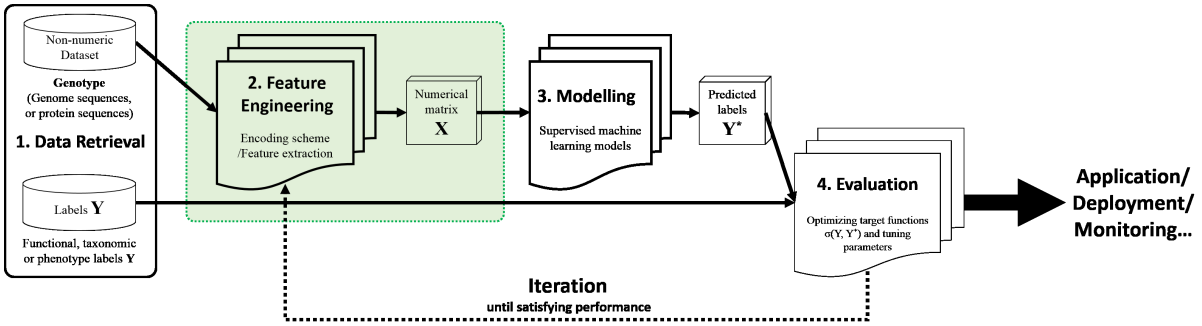


Figure 4.1: A typical pipeline of supervised machine learning models in bioinformatics. @2018 IEEE *1. Data retrieval. Preparing the genotype dataset and the corresponding labels. 2. Feature engineering. Representing the non-numeric raw dataset with numeric features that can be fed into the downstream modelling module. As CFreeEnS contributes to this module, it is highlighted with green shadow. 3. Modelling. Using supervised learning algorithms to predict the labels. 4. Evaluation. Comparing the predicted labels with the true labels to measure the performance of the model, typically by optimizing an error function* $\sigma(Y,Y^*)$*. Iterating the process from feature engineering to evaluation and tuning parameters if necessary until the model achieves a satisfying performance, and then the model is promising to be deployed for applications.*

is largely dependent on the quality of the input, which is the data representation cast by an upstream encoding scheme. Different representations can entangle and hide variant explanatory factors of the data.

As for the application in bioinformatics, usually, the dataset includes genotype information represented by symbolic genome sequences or protein sequences, and phenotype labels about the function or taxonomic name, denoted as Y. However, in the modelling part, most downstream learning algorithms need an input of numeric vectors with equal-length. Thus, an encoding scheme is needed to cast the non-numeric sequence dataset into a numeric matrix X, which can be fed into a downstream supervised learning model to predict the target labels. The predictions are denoted as Y^*. In the evaluation module, the true labels Y and the predicted labels Y^* are compared to measure the performance of the computational model.

To measure the effectiveness of encoding schemes, one way is to compare the overall performances using the same downstream learning method. A good encoding scheme should return a representation keeping the most relevant information about the predicting target and the least noise, which will benefit the predicting accuracy of the downstream learning methods. Implementing expert domain knowledge into the input dataset usually would help improve the design of a suitable encoding scheme, but an encoding scheme with more generic priors instead is more in line with the goal of automating data-driven learning.

4.2.2 CFreeEnS for protein sequences and protein sequence pairs

Herein, a context-free encoding scheme was proposed for both protein sequences with varying lengths and protein sequence pairs, which covers most of the situations in sequence analyses in

bioinformatics.

The method, CFreeEnS, is based on the AAindex database (Kawashima *et al.*, 2007), which is the collection of amino acid indexes and mutation matrices from published work, representing physiochemic and biochemical properties related to the specificity and diversity of protein structures and functions. Currently, the AAindex contains 566 amino acid indexes in AAindex1, 94 substitution matrices in AAindex2 and 47 matrices derived from the statistical pairwise contact potential between amino acids in AAindex3. For encoding protein sequences with varying length to roughly group the proteins, the CFreeEnS encodes the sequences with AAindex1. Likewise, for characterizing more subtle distinctions between proteins, the substitution matrices in AAindex2 are utilized in CFreeEnS.

Figure 4.2 presents how the improved CFreeEnS works. When taking a sequence batch $\boldsymbol{S}$ of m sequences with varying lengths as the input, CFreeEnS encodes each sequence $\boldsymbol{s_i}$ using k amino acid indexes in AAindex1, which represent generic physicochemical and biochemical properties, α-helix, β-strand and turn propensities of amino acids. For the sequence $\boldsymbol{s_i}$ encoded by index j, the outputting numeric vector is denoted as $\boldsymbol{s_i^j}$. The average value v_{ij} is calculated, representing the value of s_i with the property j. After encoded by the k indexes, the sequence s_i is represented by a vector $\boldsymbol{v_i} = [v_{i1}, v_{i2}, ...v_{ij}, ...v_{ik}]$. After stacking the vectors for m sequences, the symbolic dataset is encoded by the numeric matrix $\boldsymbol{X}$ with dimension $m \times k$, which can be fed into a downstream machine learning algorithm together with the label vector $\boldsymbol{Y}$ of length m. When analyzing the substitutions of two protein sequences, pairwise alignment is required before inputting into the encoding module. Taking a batch aligned protein sequence pairs, each sequence pair $\boldsymbol{p_i}$ of length l is encoded with k substitution matrices in AAindex2. For m sequence pairs, CFreeEnS outputs a numeric matrix $\boldsymbol{X}$ with dimension $m \times k \times l$.

Algorithm 2 clarifies how the CFreeEnS encodes a protein sequence or a pairwise protein sequence alignment using k indexes in detail. For a protein sequence s, each residue is replaced by the scores evaluated in the amino acid index idx. The average value v_{idx} of the encoded vector v_s is taken as the evaluation of the sequence. Using k amino acid indexes, the output is saved in a dictionary v where the idx is taken as the key for the v_{idx}. For a protein sequence pair with sequence s_1 and s_2 using a substitution matrix idx, the distance for each pairwise residue a_1 and a_2 is calculated as:

$$d(a_1, a_2) = idx(a_1, a_1) + idx(a_2, a_2) - 2 \times idx(a1, a2)$$

where the $idx(a_i, a_j)$ represents the score in substitution matrix idx for residue a_i and a_j. A penalty is λ is added for a gap. Similarly, using k substitution matrices, the distance vectors are saved in a dictionary v with idx as the keys. As mentioned, there are $k = 94$ substitution matrices in the AAindex database, with subtle distinctions between residues available (Tomii and Kanehisa, 1996), which provides an opportunity to systematically check all substitution scoring matrices. The most effective ones casting the dataset into different space can be selected for the scenario we need to analyze. Traversing each instance in the dataset and stacking the

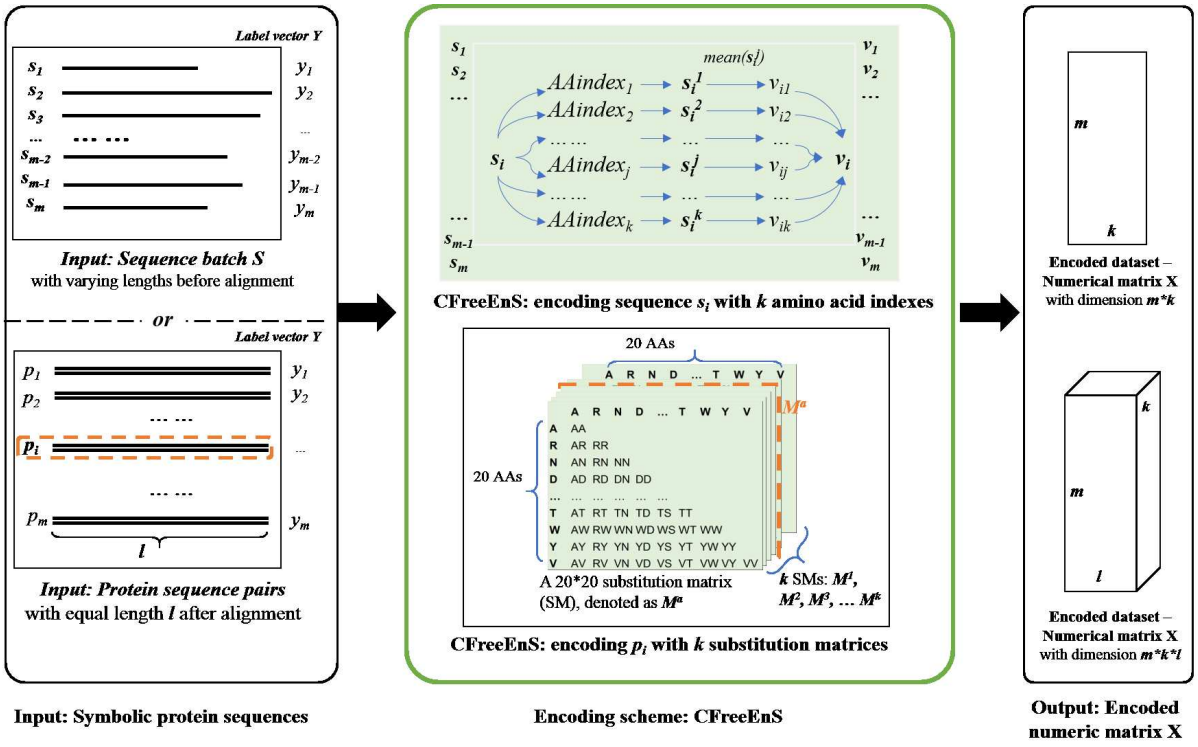


Figure 4.2: The diagram of CFreeEnS for m protein sequences or protein sequence pairs. *For sequences with varying lengths, each sequence s_i can be casted to a numeric vector using a amino acid index $AAindex_j$ in AAindex1. The average score v_{ij} is calculated. Using k amino acid indexes, the protein sequence will be represented by a vector $\mathbf{v_i}$ with length k. For aligned sequence pairs with length l, each pair can be casted to a numeric vector using one amino acid substitution matrix in AAindex2. Using k substitution matrices, each sequence will be represented by a matrix with dimension $l \times k$. For m sequence pairs, CFreeEnS outputs a $m \times k \times l$ matrix.*

value vectors, as illustrated in Figure 4.2, CFreeEnS outputs a numeric matrix of dimension $m \times k$ for m protein sequences, or a matrix of dimension $m \times k \times l$ for m protein sequence pairs. In this way, CFreeEnS can convert symbolic protein sequences and protein sequence pairs into numeric representations that can be fed into downstream learning machine learning models.

Algorithm 2 CFreeEnS for a protein sequence or a pairwise sequence alignment

```
1:  function CFreeEnS(s, idxList)        ▷ Input: s is either a protein sequence or a pairwise sequence alignment; idxList is a list with k index IDs
2:      flag = checkType(s)        ▷ checkType is a function returning 0 if the input s is a protein sequence while 1 is the input is a pairwise sequence alignment.
3:      declare v = {} ▷ v is a dictionary where the keys are the IDs of amino acid indexes for encoding
4:      if flag == 0 then                                   ▷ CFreeEnS for a protein sequence
5:          for idx in idxList do
6:              v_s = []
7:              for j = 1 to len(s) do
8:                  v_s.append(idx.get(s[j])) ▷ Get the score of each residue s[j] from the amino acid index idx
9:              v_idx = v_s.mean()
10:         v[idx] = v_idx
11:     else                                   ▷ CFreeEnS for a pairwise sequence alignment
12:         s_1 = s[0]; s_2 = s[1]
13:         assert len(s_1) == len(s_2)
14:         for idx in idxList do
15:             v_s = []
16:             for j = 1 to len(s_1) do
17:                 a_1 = s_1[j]; a_2 = s_2[j]
18:                 if a_1 == "-" or a_2 == "-" then
19:                     v_s.append(λ)              ▷ Add penalty for gaps in pairwise alignment
20:                 else
21:                     dist = idx.get(a_1, a_1) + idx.get(a_2, a_2) - 2× idx.get(a_1, a_2)     ▷ Get the distance score of pairwise amino acids
22:                     v_s.append(dist)
23:             v[idx] = v_s
24:     return v
```

The encoding scheme has been applied to different tasks of protein classification, as well as measuring the phenotype similarity between proteins, resulting in better performance than other traditional schemes. Details about the application of CFreeEnS on protein classification are presented in Appendix A (Zhou *et al.*, 2018c).

4.2.3 Model evaluation

For each dataset, the model is trained and tested with 10-fold cross validation. Assessment of the performance is based on the average of the following evaluation metrics:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

$$F\text{-}score = 2 * \frac{precision \times recall}{precision + recall} \tag{4.4}$$

Here, TP, TN, FP and FN denote true positive, true negative, false positive and false negative in the confusion matrix obtained from Y^* and $\hat{Y}^*$.

For a dataset of a single subtype, only one substitution matrix is used to encode the dataset. All the available 94 substitution matrices are used for evaluation. And then, those matrices resulting in the optimal predicting model with the highest accuracy are used to encode the combined dataset with various subtypes.

4.3 Predicting the antigenicity of diverse influenza A viruses

4.3.1 Data

The proposed method for predicting antigenicity of influenza viruses does not rely on any subtype-specific feature. Therefore, it is universally applicable to all influenza subtypes. Subsequently, the model is trained and tested on four subtypes which have drawn attention recently, namely influenza A/H1N1, A/H3N2, A/H5N1 and A/H9N2.

Antigenic HAI assay data of the four influenza viruses were collected and used to train computational models for predicting the antigenic distances of influenza viral pairs (Peng *et al.*, 2017). The Archetti-Horsfall distance (dAH) is taken as antigenic distance between a pair of viral strains (Archetti and Horsfall, 1950), which has been reported to be more robust and less dependent on antigenic factors than other measurements (Ndifon *et al.*, 2009). The dAH between viral strains i and j is calculated in Equation (4.5).

$$dAH(i, j) = \sqrt{\frac{H_{ii}H_{jj}}{H_{ij}H_{ji}}} \tag{4.5}$$

where H_{ij} is the HI titer of viral strain i relative to antisera raised against viral strain j. The antigenic distances of viral pairs Y are then discretized into a binary relationship vector Y^* with a threshold of $\theta = 4$ (Liao *et al.*, 2008) as illustrated in Equation (4.6). The estimated antigenic distances $\hat{Y}$ vector can be inferred from X by training regression models, and then discretized

with the same threshold to obtain the estimated binary relationship vector $\hat{Y}^*$.

$$Y^*(i,j) = \begin{cases} 0, & \text{if } d(i,j) < \theta \\ 1, & \text{otherwise} \end{cases} \tag{4.6}$$

Using the dAH measure, distances of 355, 791, 293 and 118 antigenic pairs were calculated for influenza A/H1N1, A/H3N2, A/H5N1 and A/H9N2 viruses, respectively. The percentages of distinct viral pairs in total viral pairs are listed in Table 4.1. The influenza A/H1N1 has approximately equal number of similar and distinct viral pairs, while the influenza A/H9N2 has more distinct pairs, around 68% in all the viral pairs. The imbalance between the similar and distinct pairs in the influenza A/H9N2 dataset may reduce the effectiveness of the predicting method. For the combined dataset, mixing antigenic data from all the four subtypes, the percentage of distinct viral pairs is 52% in all the viral pairs, which means the combined dataset has roughly balanced "similar" and "distinct" viral pairs.

Table 4.1: Datasets for training and testing the predicting model.

Subtype	Number of Sequences	T	D/T	HA1 lengths
H1N1	68	355	0.5	327
H3N2	621	791	0.47	329
H5N1	148	293	0.57	320
H9N2	29	118	0.68	317
Combined	866	1557	0.52	340

[1] *T: Total number of viral pairs;*
[2] *D: The number of antigenic distinct viral pairs;*
[3] *Combined: The combined dataset of H1N1, H3N2, H5N1 and H9N2*

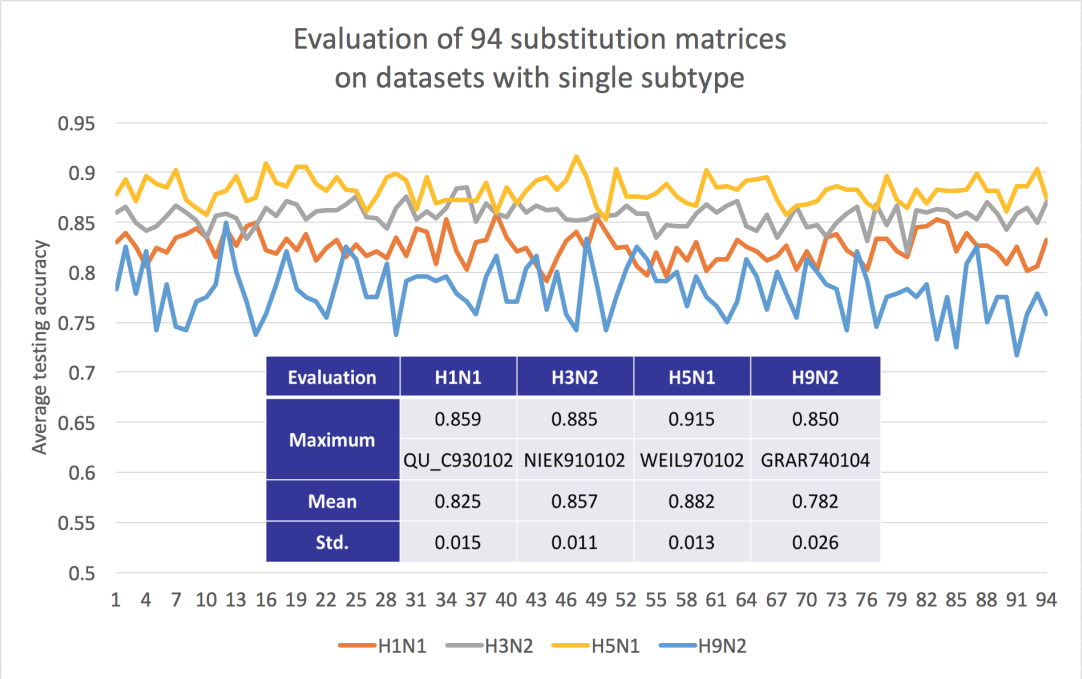
The HA1 protein sequences, the immunologic part of HA protein, of those viruses involved in HAI assays were derived from the Influenza Research Database (Squires *et al.*, 2012). For subtype-specific predictive models, the HA1 sequences were aligned according to subtypes. The lengths of HA1 sequences are 327, 329, 320 and 317 for influenza A/H1N1, A/H3N2, A/H5N1 and A/H9N2 respectively. For a universal model, HA1 sequences of all the four subtypes were mixed before being aligned. The length is 340 after the alignment, which were conducted using MAFFT v7.245 with the FFT-NS-2 progressive strategy (Katoh and Standley, 2013).

4.3.2 Results

Predictions on datasets with single subtype

For each dataset with a single subtype, namely A/H1N1, A/H3N2, A/H5N1 or A/H9N2, all the 94 substitution matrices were used to train a random forest with the same parameters. Each dataset has a distinct substitution matrix resulting in the highest testing accuracy, namely

QU_C930102 for influenza A/H1N1, NIEK910102 for A/H3N2, GRAR740104 for A/H5N1 and WEIL970102 for A/H9N2. The results of testing accuracy are visualized in a line chart (Figure 4.3). Overall, using only one substitution matrix to encode the dataset, the testing accuracy has small standard deviation (<1.5%) in each dataset, except for A/H9N2. The strategy has the best performance on the A/H5N1 dataset with an average testing accuracy of 88.2% (±1.3%), but the worst on the A/H9N2 dataset with the accuracy of 78.2% (±2.6%). The imbalance in the A/H9N2 dataset with 68% distinct viral pairs could partly explain the lower performance.



Evaluation	H1N1	H3N2	H5N1	H9N2
Maximum	0.859	0.885	0.915	0.850
	QU_C930102	NIEK910102	WEIL970102	GRAR740104
Mean	0.825	0.857	0.882	0.782
Std.	0.015	0.011	0.013	0.026

Figure 4.3: Evaluation of all substitution matrices on datasets of single subtype. *The 94 substitution matrices have an average testing accuracy higher than 80% with small standard deviation, except on A/H9N2. Each dataset has a distinct substitution matrix resulting in the highest testing accuracy.*

The best predicting accuracy score for each subtype is greater than 85%, reaching 91.5% on the A/H5N1 dataset. Models obtaining the best performance are based on different substitution matrices, namely QU_C930102 for A/H1N1, NIEK910102 for A/H3N2, GRAR740104 for A/H5N1 and WEIL970102 for A/H9N2. In QU_C930102, the matrix was inferred from the contacts of main chain atoms (Tomii and Kanehisa, 1996). NIEK910102 is a structure-derived correlation matrix considering the amino acid specific main-chain torsion angle distributions (Niefind and Schomburg, 1991). GRAR740104 combines mean chemical distances of properties: composition, polarity, and molecular volume (Grantham, 1974). WEIL970102 is a matrix obtained by subtracting the BLOSUM62 from the WAC matrix (Wei *et al.*, 1997).

In addition, the proposed encoding strategy CFreeEnS was compared with the mutation-counts-based method proposed by Liao *et al.* and regional band-based method proposed by Peng *et al.* on the same datasets. It is worth noting that the methods use not only different encoding schemes, but also distinct training models. To demonstrate that our CFreeEnS is

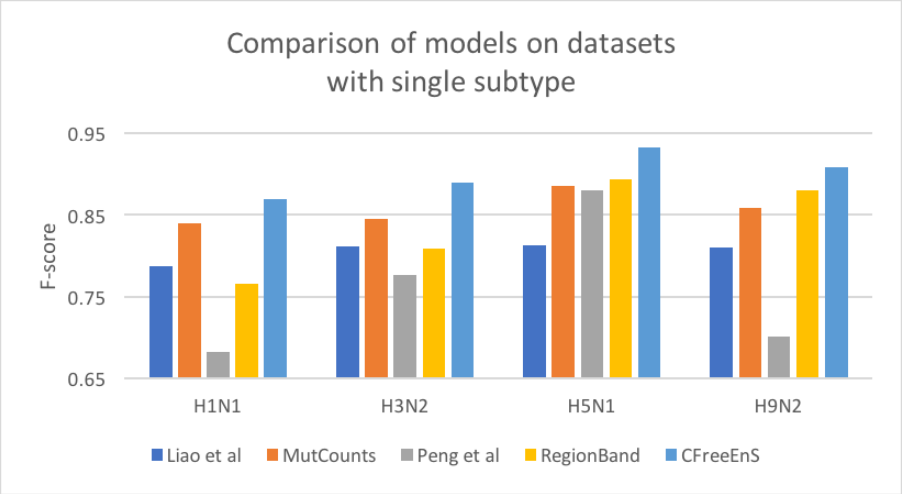


Figure 4.4: Comparing F-score of models on datasets with singe subtype influenza viruse.

more accurate than the subtype-specific handcrafted ones, the mentioned methods in literature were adapted by keeping the upstream encoding scheme and using random forest as the same training model, denoted as MutCounts and RegionBand, respectively.

Figure 4.4 shows the comparison of F-score among five strategies on the four datasets with single-subtype influenza viruses. CFreeEnS obtains the highest F-score among the five strategies on all the four datasets (besides the combined dataset). Accuracy, precision and recall are also evaluated (Table 4.2). Although CFreeEnS sometimes ranks the second or third in precision or recall, it always obtains the highest accuracy and F-score. The experiments demonstrate that our proposed encoding scheme CFreeEnS outperforms subtype-specific features MutCounts and RegionBand in predicting the antigenicity of influenza viruses within the same subtype.

Prediction on the combined dataset with diverse subtypes

The proposed CFreeEnS does not use any subtype-specific information, and thus can be applied to datasets with either one subtype or various subtypes. For a dataset with one subtype, one substitution matrix is enough to encode the dataset. All the available 94 substitution matrices are evaluated. Those with top ranking testing accuracy are used to encode the combined dataset with various subtypes.

For datasets with a single subtype, all the available substitution matrices were traversed. Each dataset has a distinct substitution matrix resulting in the highest testing accuracy, namely QU_C930102, NIEK910102, GRAR740104, and WEIL970102. The four substitution matrices, derived from different properties of amino acids, are selected as the optimal substitution matrices in predicting antigenicity of influenza viruses, denoted as CFreeEnS-4 to be distinguished from CFreeEnS which uses one substitution matrix. With CFreeEnS-4, the 866 viral pairs are encoded as a $866 \times 4 \times 340$ matrix. To feed the data into machine learning models, it was flattened as a 866×1360 matrix, where the 4 feature vectors for each instance were

Table 4.2: Performance comparison among five strategies on four single subtype datasets.

Dataset	Methods	Accuracy	Precision	Recall	F-score
H1N1	Liao *et al.*	0.742	0.717	0.877	0.788
	MutCounts	0.824	0.802	0.884	0.840
	Peng *et al.*	0.661	0.671	0.711	0.683
	RegionBand	0.706	0.669	**0.901**	0.766
	CFreeEnS	*__0.859__	**0.856**	0.887	**0.870**
H3N2	Liao *et al.*	0.784	0.748	0.891	0.812
	MutCounts	0.843	0.841	0.851	0.845
	Peng *et al.*	0.720	0.658	**0.950**	0.777
	RegionBand	0.790	0.763	0.864	0.809
	CFreeEnS	**0.885**	**0.896**	0.882	**0.889**
H5N1	Liao *et al.*	0.753	0.758	0.878	0.813
	MutCounts	0.863	0.859	0.915	0.885
	Peng *et al.*	0.846	0.857	0.908	0.880
	RegionBand	0.858	0.824	**0.978**	0.893
	CFreeEnS	**0.915**	**0.903**	0.965	**0.932**
H9N2	Liao *et al.*	0.708	0.816	0.819	0.810
	MutCounts	0.775	0.823	0.914	0.859
	Peng *et al.*	0.633	**0.888**	0.601	0.702
	RegionBand	0.804	0.818	0.954	0.880
	CFreeEnS	**0.850**	0.860	**0.964**	**0.908**

* The highest scores among five strategies on each dataset are colored **red**.

stacked by column. Here, random forest was used with the same restrictions on maximum depth of trees, i.e. 9.

The inconsistency of auto-selected substitution matrix indicates that different properties may dominate the viral antigenicity in different subtypes of influenza viruses. To improve the prediction in diverse subtypes, all those properties are taken into account to encode the combined dataset. The increases of predicting accuracy compared with MutCounts and RegionBand are 14.8% and 9.5% respectively, indicating that cross-subtype properties have been captured by the encoding scheme CFreeEnS. Further experiments on transfer learning have supported that the properties captured in one subtype of influenza can also work well in predicting the antigenicity of other subtypes of influenza.

Table 4.3 presents the performance comparison among five strategies on the combined dataset. With 10-fold cross-validation, the average testing accuracy of CFreeEnS-4 on the combined dataset is 84.6%, higher than the second highest accuracy of 75.1% using the regional band-based method.

The proposed encoding scheme CFreeEnS outperforms current methods that handcraft subtype-specific features when applied to predicting the antigenicity of influenza viruses, especially in the combined dataset with various subtypes. By systematically checking all the available substitution matrices, which consider different properties of amino acids, it is found that properties related to the structures of amino acids or contacts between amino acids can

Table 4.3: Performance comparison among five strategies on the combined dataset.

Dataset	Methods	Accuracy	Precision	Recall	F-score
Combined	Liao *et al.*	0.739	0.716	0.879	0.789
	MutCounts	0.698	0.675	**0.944**	0.781
	Peng *et al.*	0.741	0.757	0.800	0.775
	RegionBand	0.751	0.723	0.912	0.807
	CFreeEnS-4	*__0.846__	__0.837__	0.900	__0.867__

* The highest scores among five strategies on each dataset are colored **red**.

help improve the prediction in the combined dataset.

To be more specific, besides fundamental properties such as composition, polarity and molecular volume, information about contacts of main chain atoms and amino acid specific main-chain torsion angle distribution can help improve the predicting accuracy. This is consistent with our knowledge that different viral subtypes share major protein structures. The shared properties which affect the antigenicity of diverse influenza subtypes may give insights into the mechanisms of virulence of the influenza viruses. Another interesting finding is that the substitution matrices used in different subtypes are distinct. It suggests that the amino acid properties dominating the antigenicity of influenza viruses may vary from subtype to subtype.

Transfer learning: predicting the antigenicity of an emerging unknown subtype of influenza A virus

To check whether the knowledge gained in one subtype can be applied to the other subtype, experiments about transfer learning across subtypes were conducted. To be more specific, a random forest using one subtype was trained. Subsequently, the trained model was tested on a different subtype of which not a single viral strain has been used in the training. For example, a model was trained on the influenza A/H1N1 dataset, and then tested on the influenza A/H3N2, A/H5N1, A/H9N2 datasets respectively.

The accuracies of transfer learning using the three encoding schemes (i.e., MutCounts, RegionBand and CFreeEnS) are shown in Figure 4.5. We can observe that CFreeEnS outperforms the other two encoding schemes in every experiment. The highest prediction accuracy is 84.3% when the model is trained on the A/H1N1 dataset and tested on the A/H5N1. The experiments of transfer learning indicate that CFreeEnS can encode generic properties conserved across subtypes. In addition, it gives a high accuracy in predicting the antigenicity of influenza A/H5N1 (83.3%) even with small training dataset like A/H9N2 (only 118 sequence pairs as training instances). The full result of comparison is available in Table 4.4. In some experiments, RegionBand has moderately better performance in recall. Overall, however, CFreeEnS has higher F-scores. Integrating the regional band-based handcrafted features into the encoding scheme might further improve the performance of prediction. Figure 4.6 presents the learning curves of the random forest regressor trained on different datasets, which shows that our models do not suffer the over-fitting problem.

Table 4.4: Evaluation metrics for the performances of three encoding schemes on transfer learning.

TrainSet	TestSet	Method	Accuracy	Precision	Recall	F-score
H1N1	H3N2	MutCounts	0.660	0.649	0.758	0.697
		RegionBand	0.624	0.585	0.934	0.720
		CFreeEnS	*0.722	0.776	0.650	0.707
	H5N1	MutCounts	0.723	0.773	0.778	0.775
		RegionBand	0.752	0.726	0.961	0.826
		CFreeEnS	0.843	0.868	0.878	0.873
	H9N2	MutCounts	0.657	0.837	0.670	0.736
		RegionBand	0.758	0.770	0.959	0.854
		CFreeEnS	0.814	0.865	0.885	0.875
H3N2	H1N1	MutCounts	0.619	0.686	0.545	0.606
		RegionBand	0.627	0.626	0.774	0.692
		CFreeEnS	0.685	0.767	0.599	0.673
	H5N1	MutCounts	0.662	0.712	0.756	0.732
		RegionBand	0.746	0.738	0.909	0.815
		CFreeEnS	0.805	0.769	0.978	0.861
	H9N2	MutCounts	0.636	0.779	0.706	0.741
		RegionBand	0.777	0.789	0.952	0.863
		CFreeEnS	0.797	0.818	0.931	0.871
H5N1	H1N1	MutCounts	0.637	0.624	0.824	0.710
		RegionBand	0.647	0.626	0.861	0.725
		CFreeEnS	0.763	0.773	0.797	0.785
	H3N2	MutCounts	0.721	0.683	0.865	0.762
		RegionBand	0.670	0.621	0.929	0.744
		CFreeEnS	0.733	0.703	0.839	0.765
	H9N2	MutCounts	0.640	0.823	0.666	0.727
		RegionBand	0.740	0.759	0.949	0.843
		CFreeEnS	0.805	0.802	0.977	0.881
H9N2	H1N1	MutCounts	0.583	0.596	0.734	0.651
		RegionBand	0.624	0.601	0.909	0.723
		CFreeEnS	0.676	0.688	0.734	0.710
	H3N2	MutCounts	0.684	0.690	0.750	0.709
		RegionBand	0.622	0.585	0.934	0.719
		CFreeEnS	0.765	0.782	0.756	0.769
	H5N1	MutCounts	0.723	0.806	0.725	0.759
		RegionBand	0.791	0.757	0.977	0.853
		CFreeEnS	0.833	0.802	0.967	0.877

* The highest scores among three encoding schemes are colored red.

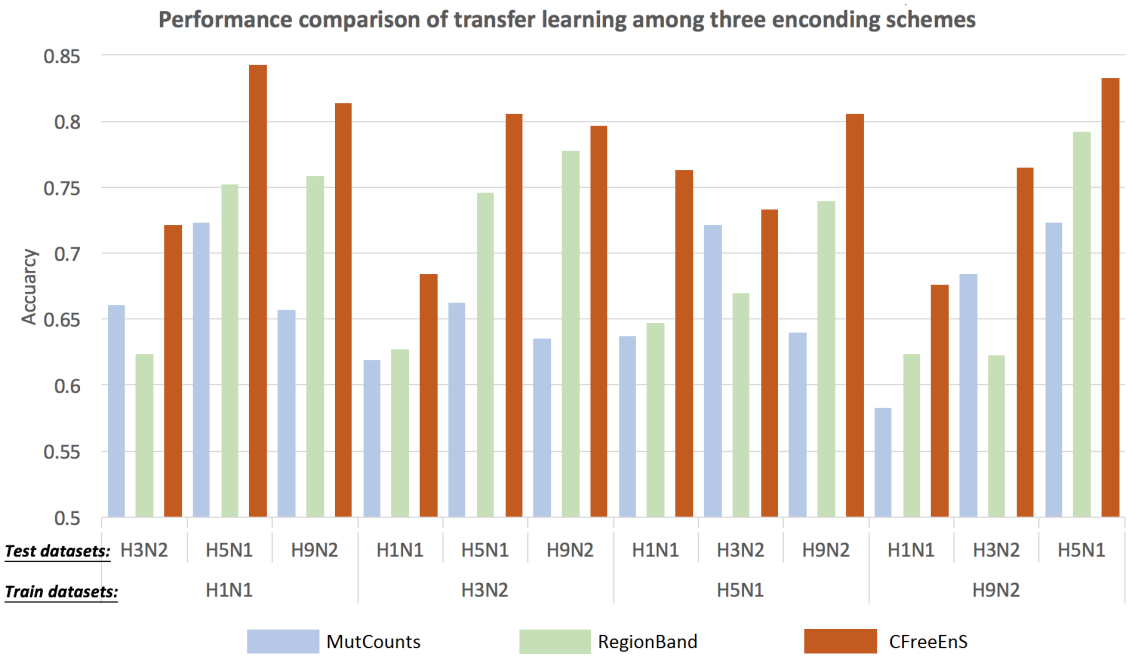
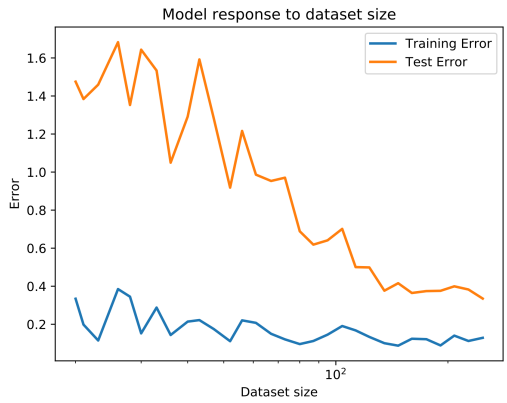


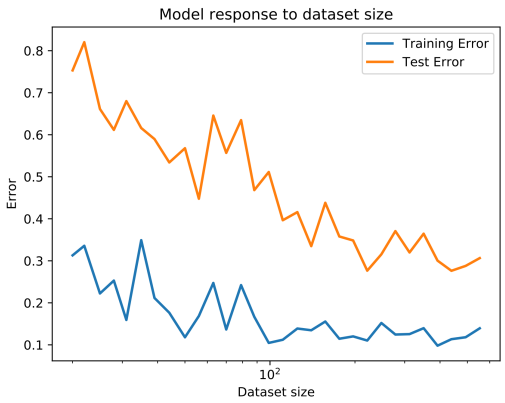
Figure 4.5: Accuracy scores of transfer learning using three encoding schemes: MutCounts, RegionBand and CFreeEnS. *MutCounts: features that are used in the method proposed by Liao et al.; RegionBand: features that are used in the method proposed by Peng et al.. All the models use random forest as a downstream learning method.*

4.4 Summary

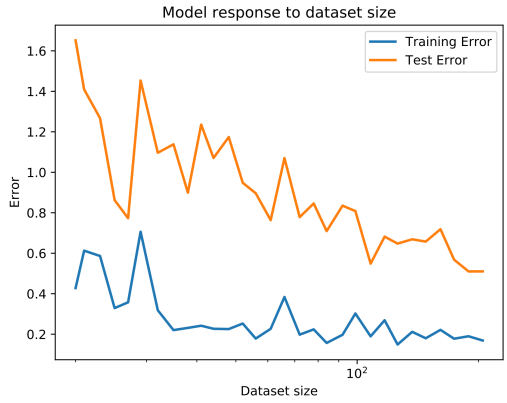
CFreeEnS is a protein representation heavily depends on the AAindex database, i.e. the known properties of amino acids. Therefore, features selected by the downstream learning models are easy to interpret through the analysis of variable importance. It has been demonstrated that with features created by CFreeEnS, protein functions can be predicted with high accuracy. However, CFreeEnS is not good at disentangling more abstract features, or providing a new angle to explain the relationship between genotype and phenotype. The *m-NGSG*, inspired from NLP, although not as good as CFreeEnS on the mentioned tasks, provides a novel perspective to treat the biological sequences. The abstract features generated by *m-NGSG* can be taken as new properties of a group of amino acids. Graphic representations of protein sequences are also interesting, providing visual qualitative inspection of sequences, but they are not efficient in describing long protein sequences (Wen and Zhang, 2009). Different representations of protein sequences may disentangle or hide different aspects. An encoding scheme capturing the known generic properties of amino acids can help automate the process of constructing features and facilitate annotating protein functions. For profiling other aspects of proteins, e.g. predicting the protein folds, computational predictions using other novel representations may give more insights. Besides, the CFreeEnS, free from dependence on carefully designed features, is applicable to encoding different protein sequence pairs into a numeric matrix. It is promising for other applications in bioinformatics measuring the phenotype similarity from sequences, such as the neutralization escape of HIV-1 virus (Wu, 2014).



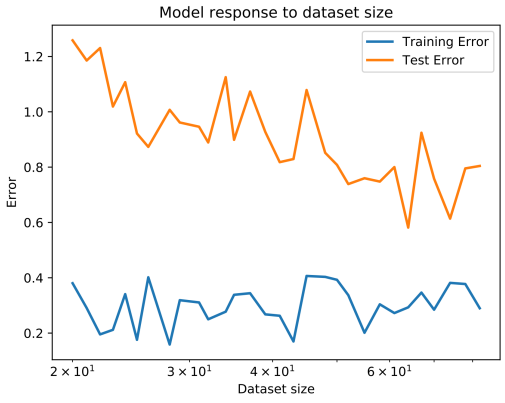
(a) The influenza A/H1N1 dataset encoded using QU_C930102.



(b) The influenza A/H3N2 encoded using NIEK910102.



(c) The influenza A/H5N1 encoded using GRAR740104.



(d) The influenza A/H9N2 encoded using WEIL970102.

Figure 4.6: Learning curve of the random forest regressor trained on different datasets.

Chapter 5

Structure-based analysis to quantify viral binding preference with host cells

Genomic variation of influenza viruses can be observed and predicted from sequence models (Chapter 3). With the enrichment of immunological assays data, the antigenicity of viral strains can also be inferred (Chapter 4). However, determinants of the virulence of influenza viruses remain an open challenge. Receptor binding specificity is an important aspect to profile the viral strains. This chapter focuses on structural analyses on the HA protein, quantifying the receptor binding specificity of a viral strain with known HA protein sequence. Section 5.1 gives a brief introduction on the structural basis of viral binding with host cells. Section 5.3 investigates the HA mutations of a new emerging influenza A/H7N9 strain, giving insights into how the mutations change the protein structure and the receptor binding specificity of the protein. Work of this chapter is main based on the publication Zhou *et al.* (2018a).

5.1 Structural basis of viral binding with host cells

Hemagglutinin (HA), a homotrimer transmembrane glycoprotein of the influenza viruses, functions in the process of viral binding and endocytosis of host cells. The HA enables the virus to enter a host cell by recognizing sialic acids that are attached to the host cell transmembrane proteins. Figure 5.1 depicts how the HA binding with host cells, as well as structures HA and sialic acid (Daniels *et al.*, 1987; Lazniewski *et al.*, 2017). Each monomer of HA consists of the HA1 and HA2 subunits after the proteolytic cleavage of the HA0 precursor. The HA1 attaches to the influenza viral particles through the receptor binding domains, while the HA2 facilitates the membrane fusion of the viral particle with the endosomal lipid bilayer. Sialic acid serves as the receptor in a chain that is attached to a host cell surface protein.

The receptor binding sites (RBSs) of all subtypes of HA protein locate at the globular domain of the HA1 subunit.The RBSs forms a shallow pocket, where the four conserved residues Y98, W153, H183 and Y195 (named after the H3 numbering system (Burke and Smith, 2014)) serve as bases of the structure. There are three structural domains nearby the four conserved

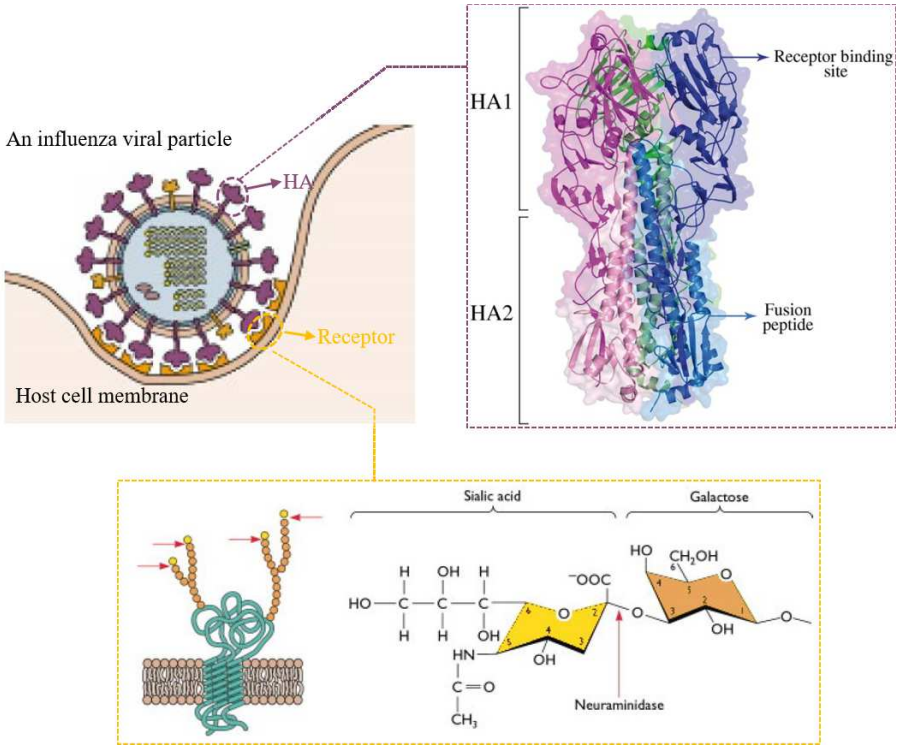


Figure 5.1: An illustration of an influenza viral particle binding with host cells. *The structures of HA protein and the host cell receptor have been zoomed in. For the homotrimer HA protein (purple), each monomer (colored blue, magenta and green respectively) consists of an HA1 and HA2 subunit, containing the receptor binding sites and fusion peptide respectively. Sialic acid serves as the receptor in a chain that is attached to a cell surface protein embedded in the plasma membrane. A typical structure of sialic acid linked with galactose via an $\alpha_{2,3}$ linkage is shown in the yellow box.*

residues, namely the 130-loop, 190-helix and 220-loop, serving as the basic binding pocket for sialic acid. The length and amino acid compositions in the RBSs vary between strains and can the determinants of the preferential receptors.

Mammalian membranes are composed of glycolipids, glycoproteins, glycophospholipid anchors and proteoglycans. Some glycans are terminated with a galactose connected to the sialic acids. The main sialic acid in human is the N-acetylneuraminic acid, connected with galactose via the $\alpha_{2,3}$ or $\alpha_{2,6}$ glycosidic bond, shown in Figure 5.2. Both types of sialic acid have been discovered in human bronchial and lung tissues. The $\alpha_{2,6}$-linked glycans distribute mainly in the upper respiratory tract while the $\alpha_{2,3}$-terminated glycans are main in the lower respiratory tract (Ibricevic *et al.*, 2006; Xiong *et al.*, 2014). Besides, more $\alpha_{2,3}$-linked glycans are observed in the lungs of children than adults. For the avian species, both $\alpha_{2,3}$- and $\alpha_{2,6}$-linked glycans have been observed in the respiratory and intestinal tract, but mainly the $\alpha_{2,3}$-terminated glycans (Edinger *et al.*, 2014).

The avian influenza virus strains preferentially bind to sialic acids with the $\alpha_{2,3}$ linkage, while strains circulating among humans often prefer binding with the $\alpha_{2,6}$-linked sialic acid. Therefore, the glycan with $\alpha_{2,3}$ linkage is also named as the "avian receptor", while the $\alpha_{2,6}$-terminated glycan is often called the "human receptor". Efficient transmission of the avian influenza viruses among humans often requires the recognition of $\alpha_{2,6}$-linked glycans, which

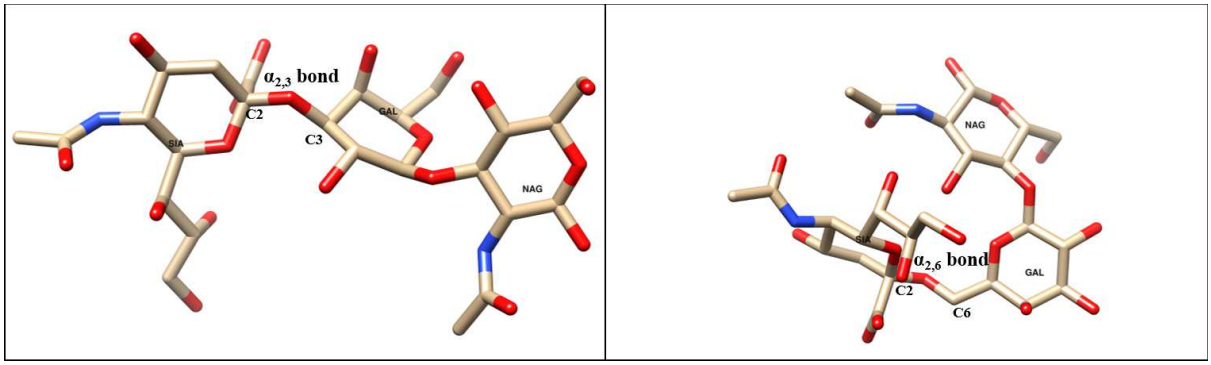


Figure 5.2: Structure of sialic acids recognized by the HA protein of influenza viruses. *The left presents the avian-type receptor, where the SIA and GAL is linked by the $\alpha_{2,3}$ bond, while the right presents the human-type receptor, where the SIA and GAL is linked by the $\alpha_{2,6}$ bond.*

allows the virus to replicate in the upper respiratory tract and trigger host pathways (e.g. cough, sneezing) leading to efficient transmission. Besides, the ability of viral binding with the $\alpha_{2,3}$-linked glycans is one determinant for the infection in the lower respiratory system (e.g. the lung), leading to more severe symptoms and even death. Thus, the preference of receptors is taken as one key factor influencing the host specificity and virulence of influenza.

The human respiratory tract is covered with diverse long sialylated glycans. Restricted to the availability and the quantity of the receptors, the most popular ligands co-crystallized with the HA protein are the receptor analogs, typically linear and short. For example, the linear sialopentasaccharides LSTa (NeuAcα2-3Galβ1-3GlcNAcβ1-3Galβ1-4Glc) and LSTc (NeuAcα2-6Galβ1-4GlcNAcβ1-3Galβ1-4Glc), which are isolated from human milk, are the avian and human receptor analogs respectively. Other analogs, including the 3'SLNLN (NeuAcα2-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAc) and 6'SLNLN (NeuAcα2-6Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAc), the shorter sialotrisaccharide 3'SLN (NeuNAcα2-3Galβ1-4GlcNAc) and 6'SLN (NeuNAcα2-6Galβ1-4GlcNAc), are popularly used to assess the binding preferences of influenza viruses. It should be noted that analyzing the HA co-crystallized structure with receptor analogs is an simplification of the problem, and receptors may not bind exact the same way as the receptor analogs.

5.2 Methods to identify genetic markers for cross-species transmission or virulence determinants

5.2.1 Animal models

Parallel adaptive animal models (e.g. mice, ferrets, swine, non-human primates etc.) have been significant yet expensive helpers in observing the evolution, host tropism, pathogenicity and transmissibility of influenza viruses *in vivo*, identifying repeated and probably gain-of-function mutations (Ping *et al.*, 2011; Narasaraju *et al.*, 2009). The applicable advantages and disadvantages of the most popular animal models have been summarized in Table 5.1.

Table 5.1: Advantages and disadvantages of different animal models.

Species	Advantages	Disadvantages
Mice	✓ Small size ✓ Low cost (animals, housing) ✓ Homogeneous responses - inbred, pathogen free ✓ Availability of molecular biology/immunology reagents ✓ Pathology of viral pneumonia caused by highly pathogenic viruses (1918 H1N1, HPAI H5N1) similar to humans	× Seasonal influenza virus strains need adaptation in order to achieve efficient replication and virulence × Respiratory tract anatomy and receptor distribution different from humans × Not suitable for study of live-attenuated vectored vaccines × Not suitable for transmission experiments
Ferret	✓ Respiratory tract anatomy and receptor distribution similar to humans ✓ Human-like clinical signs and pathology of disease ✓ Human and avian influenza virus isolates replicate without prior adaptation ✓ Suitable for transmission experiments	× Limited availability of molecular biology/immunology reagents × Host response variability—genetically outbred × Need to confirm seronegativity to influenza × Systemic disease different than in humans × Genome not annotated × Practical considerations—use of high number of animals per group very expensive
Pig	✓ Human and avian influenza virus isolates replicate without prior adaptation ✓ Availability of molecular biology/immunology reagents	× Host response variability—genetically outbred × Need to confirm seronegativity to influenza (maternal antibodies might be problematic) × Seem to mount an abnormal response in certain heterologous challenges, which has not been observed in humans or other species × Practical issues—big size, husbandry requirements
NHP	✓ Respiratory tract anatomy and receptor distribution similar to humans ✓ High similarity to the human immune system ✓ Susceptible to non-adapted human strains ✓ Broad availability of molecular biology/immunology reagents	× Lack of clinical signs upon infection with seasonal strains × Host response variability—genetically outbred × Need to confirm seronegativity to influenza × Ethical concerns × Prohibitively expensive × Very experienced personnel and highly specific facilities needed × Variable degree of permissiveness for influenza virus infection and clinical signs

Bouvier and Lowen reviewed their applicable scenarios and protocols (Bouvier and Lowen, 2010). Researchers endeavored to find genetic markers for virulence, or other virulence-related traits.

Some researches work on highlighting genes related to the transmissibility among hosts. It is obvious that human-to-human transmission test is infeasible. Alternatively, human influenza viral strains also show efficient transmissibility among ferrets and guinea pigs, which bear similar receptors distribution with humans (Bouvier and Lowen, 2010). Thus, they are often used to examine transmissibility-related genome signatures. For instance, Maines *et al.* analyzed the transmissibility of swine-origin 2009 A/H1N1 through respiratory droplets by establishing ferrets and mice model compared with the seasonal A/H1N1 (Maines *et al.*, 2009). The results show that the 2009 A/H1N1 strain caused increased morbidity but transmitted less efficiently than the seasonal strain. Further, they tested the binding specificity of the viral HA by a dose-dependent direct receptor-binding and human lung tissue-binding assays to explore the mechanism. Herfst *et al.* looked into the mutant A/H5N1 viruses, showing that 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 (Herfst *et al.*, 2012).

In addition to analyzing existing strains, the emerging gain-of-function (GoF) studies attempt to highlight mutations that enhance airborne transmissibility among ferrets . The GoF studies include any experiments that generate a phenotype by altering genotypes. Virologists

use gain- and loss-of-function experiments to help reveal the genetic makeup of viruses and the genotype-phenotype relationship but under much debate (Duprex *et al.*, 2015). For example, Chou *et al.* analyzed which segment of the 2009 pandemic strain conferred the transmissibility by making reassortment between the pandemic A/California/04/09(H1N1) strain and another poorly transmissible A/Puerto Rico/8/34(H1N1) and then tested transmissibility in guinea pigs (Chou *et al.*, 2011). They found that the M segment should be responsible for the enhanced transmission among human. Other traits of HA related to transmissibility identified by GoF studies include preference to different types of sialic acid linkages (Yamada *et al.*, 2006), glycosylation on the globular head (Herfst *et al.*, 2012; Imai *et al.*, 2012) and HA stability. Similarly, Yen *et al.* found the NA segment may also played an important role in facilitating human-to-human transmission. Results of Massin *et al.* emphasized the importance of polymerase activity on mammalian transmission. Substitutions on PB2 (e.g. E627K) can confer the influenza viral adaptation to mammalian cells (Long *et al.*, 2013; Jagger *et al.*, 2010; Zhu *et al.*, 2010). It is worth noting that these traits of proteins are only indicative of transmissibility in the specific animal models, but not sufficient or necessary for the transmissibility among mammalians (Mehle and Doudna, 2009b; Herfst *et al.*, 2010).

Pathogenicity-related genome signatures, including drug-resistant markers, are also under investigation. For instance, Watanabe *et al.* showed an increased pathogenicity in macaques when conferred the avian-type receptor binding ability by HA-222D and HA-222G (Watanabe *et al.*, 2011). The substitution NA-H275Y results viral resistance to oseltamivir (Baz *et al.*, 2010; Bloom *et al.*, 2010). There is also evidence showing cooperative or compensating mutations (Gong *et al.*, 2013). Again, we should bear in mind that even when phenotypic traits of interest can be identified, clear genetic markers for these traits are only present in some cases.

However, experiments in animal models are costly and time-consuming. Such experiment cannot exhaust all genetic variants of influenza virus that associated with transmissibility and pathogenicity. Findings from experiments and the data generated in experiments provide valuable information for further computational analysis. The development of rapid phenotype assessment providing more evidence indicating the genotype-phenotype relationship will greatly help the computational modelling analysis.

5.2.2 Computational models

Benefited from the development of high throughput sequencing techniques and accumulated experimental data, computational models have been an attractive complement to experimental studies. However, the Achille's heel is the lack of established validating framework for the numerous computational methods, resulting in the concerns about accuracy of the predictions. The experimental results of animal models provide valuable information on the relationship between genotype and phenotype. By integrating the laboratory genotype-phenotype assessment and large-scale viral sequences, it is possible to delineate the biochemical traits of the viral protein and predict phenotype from sequences.

There have been tools collecting the findings and highlight those interpretable residues of an input influenza protein sequence. For example, the influenza A/H5N1 influenza genetic changes inventory compiled by CDC supports international surveillance on the highly pathogenic strain (https://www.cdc.gov/flu/avianflu/h5n1-genetic-changes.htm). Besides, the FluSurver enables the researchers to quickly access the residue annotations given an influenza protein sequence (http://flusurver.bii.a-star.edu.sg/). Large-scale sequences available makes it possible to get clues from past strains. The reconstruction of evolutionary pathways helps to understand the evolutionary mechanism that allows adaptive substitutions controlling host tropism and pathogenicity.

Besides, great efforts have been made in annotating signatures for cross-species transmission and increased virulence, which can facilitate early detection of potential pandemic strains. For example, Miotto *et al.* identified a catalog of sites as human-to-human transmission markers by comparing a set of human-transmissible and avian isolates (Miotto *et al.*, 2008). Eng *et al.* made use of the reported genomic signatures to profile influenza viral proteins and then classified the avian and human influenza viral strains using the SVM (Eng *et al.*, 2014). Similarly, Qiang and Kou employed energy-feature-vectors into a protein sequences and constructed an artificial neural network (ANN) model to discriminant avian and human influenza A viruses (Qiang and Kou, 2010).

However, those sequence-based computational models require solid validation against experimental data, particularly for revealing the context dependency of these substitutions before the models coming into service for informing policy-making. One debate on machine learning algorithms, especially the inner workings of neural networks, is that they perform like a black-box in the sense that they can approximate any function without giving any insights into the structure of the function being approximated. In terms of the application in biology, the estimated functions for prediction may not have substantial biological meaning.

Structural analysis, benefited from the enriched protein structure database, may compensate for that by analysing the impact of substitutions on protein structure, simulating the protein-ligand binding process and thereby optimizing therapies (Koday *et al.*, 2016). For example, Su *et al.* combined molecular docking and molecular dynamics simulation to analyze conformation changes of drugs binding to NA of H7N9 under the mutation R289K, finding that this particular mutation may contribute to the enhanced drug resistance of influenza A/H7N9 (Su *et al.*, 2013). Similarly, Kannan and Kolandaivel; Pan *et al.* investigated the binding preference of A/H1N1 HA protein with different host cell receptors, giving insight on the enhanced virulence by the HA mutation D222G (Kannan and Kolandaivel, 2016; Pan *et al.*, 2012).

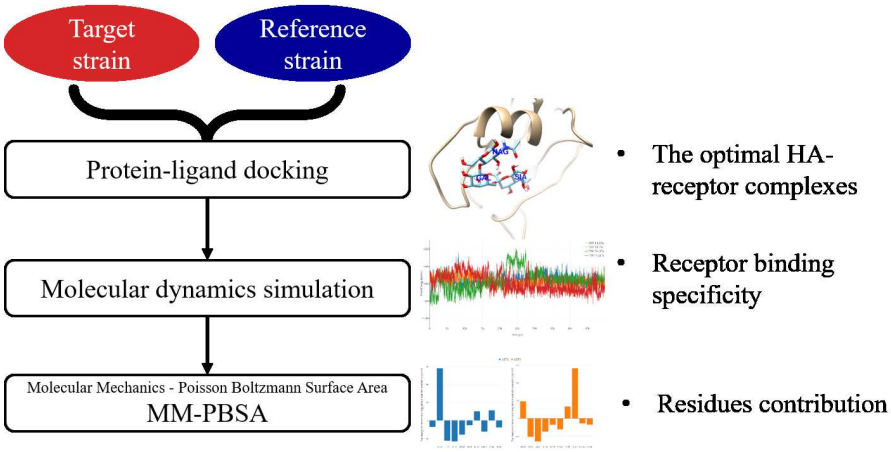


Figure 5.3: The workflow for structural analyses on HA-receptor bindings.

5.3 The receptor binding specificity of a novel influenza A/H7N9 virus

Influenza viruses are undergoing continuous and rapid evolution. Experimental studies have revealed several host and virulence markers, indicating differential host binding preferences which can help estimate the potential of causing a pandemic. The fatal influenza A/H7N9 has drawn attention since the first wave of infections in March 2013, and raised more grave concerns with its increased potential to spread among humans.

The fact that infections of influenza A/H7N9 in poultry are subclinical makes it challenging to surveil its spreading among poultry and to measure the risk of human infection (Xiang, 2016). Most cases have been identified following their exposure to live birds or live poultry markets. Even with small clusters of human infection cases being detected, the WHO declared low likelihood of human-to-human transmission of avian influenza A/H7N9, with the support of current epidemiological and virological evidence. One case reported on 22 Feb 2017 caught our attention. A patient declared no exposure to any live birds, live poultry markets or others with suspicious illnesses (Shen and Lu, 2017). The potential of the newly emerging strain to circulate among humans should be measured. This strain isolated from this patient is named influenza A/Taiwan/1/2017(H7N9), abbreviated as TW17.

In this section, the mutations and receptor binding specificity of the novel influenza A/H7N9 are systematically analyzed using computational approaches. Section 5.3.1 highlights the mutations, functional markers and binding regions of the novel influenza A/H7N9 strain. Section 5.3.2 focuses on the HA protein, which bears the most mutations. Figure 5.3 shows an overview of structural analyses on HA-receptor bindings. First, the optimal conformations with host cell receptors are predicted from molecular docking. And then, the changes of viral binding preference and the residue contributions are modeled using molecular dynamics simulation (see section 5.3.3). The results give insight into how the mutations change the protein structure, especially the functional domains of the protein.

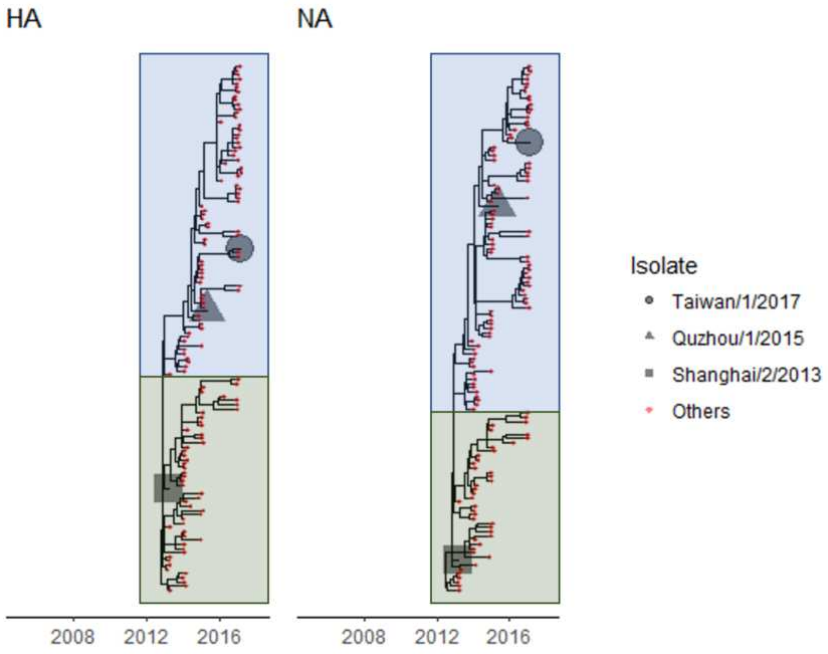


Figure 5.4: The HA and NA phylogenetic trees of influenza A/H7N9. *Two major clusters are denoted using blue and green boxes.*

5.3.1 Functional markers of the novel influenza A/H7N9 strain

Selecting representative strains from phylogenetic trees

To select representative strains of influenza A/H7N9, the phylogenetic trees were constructed for all proteins of influenza A/H7N9. All 125 available isolates of influenza A/H7N9 as of 8 April 2017 were retrieved from GISAID EpiFlu Database (Shu and McCauley, 2017). The coding sequences were obtained and aligned using the influenza virus sequence annotation tool (Bao *et al.*, 2007) and MUSCLE (Edgar, 2004). The phylogenetic trees were inferred using BEAST (Bouckaert *et al.*, 2014) and assessed with TRACER (Rambaut *et al.*, 2014). The maximum clade credibility trees were selected as output and visualized using ggtree (Yu *et al.*, 2017). Figure 5.4 shows that the HA and NA of recent circulating influenza A/H7N9 strains could be from two different clusters, colored with blue and green boxes respectively. No conspicuous positive selection was observed in each cluster. The viral strain isolated from the patient claiming no exposure to any live bird or live poultry market is influenza A/Taiwan/1/2017(H7N9) (ID:EPI_ISL_248778), abbreviated as TW17. The strain influenza A/Shanghai/02/2013(H7N9) from the first outbreak of influenza A/H7N9 in 2013, and the strain influenza A/Quzhou/1/2015(H7N9) bearing the highest HA sequence similarity with TW17 were selected as reference strains, abbreviated as SH13 and QZ15 respectively. Besides, the phylogenetic trees indicate that the HA and NA of current circulating HA and NA strains may be originated from the SH13.

Mutations, suspicious functional markers and receptor binding regions of the circulating influenza A/H7N9

Mutations of all proteins of TW17 were obtained by comparing with the reference strain SH13. Table 5.2 summarizes the phenotypically or epidemiologically interesting mutations and their potential functional domains. The annotations were retrieved from FluSurver (http://flusurver.bii.a-star.edu.sg/). There are many mutations locating at the viral oligomerization interfaces or regions binding with small ligands, suggesting possible significant change of viral binding with small ligands, especially with host cell receptors. It is not surprising that the HA protein has the most mutations. Other interesting mutations are the M2:E2D, PA:G66S, and PB2:M570I, E627K, locating at the viral oligomerization interfaces or regions binding with small ligands. Another PB2 mutation I292V has been reported to be associated with the host specificity shift of influenza viruses. The NA mutations M26I, M72I, Y166H, A210V, S242P, R289K and N322S locate at the viral oligomerization interface, among which S242P and R289K are also involved in the process of human immune responses.

Since the HA of TW17 bears the most mutations which cover the processes of both host cell recognition and immune response, the impact of HA mutations on the TW17 strain was investigated in detail. Four insertions RKRT after site 337 and 15 mutations were observed. Among the observed mutations, 14 locate at viral oligomerization interfaces, including A130P, S136N, I138T, L235Q and I335V which have been reported to be related to antigenic shifts or mild drug resistance (Philpott *et al.*, 1990). Besides, the mutation S136N and I138T introduce a new potential N-glycosylation site with pattern NGTR(136-139). The potential of NNTY at site 493 to be N-glycosylated is increased as predicted by NetNGlyc (Gupta and Brunak, 2001). Glycosylation, an important way for the influenza viruses to evolve, has been reported to be associated with host specificity, virulence and immune response (Wu *et al.*, 2017a). The likely introduced two glycosylation patterns are worth noting.

Furthermore, a carbon probe based approach was applied to identify putative ligand binding sites. The tool SITEHOUND was used to calculate an affinity map for the carbon probe and then cluster the points with favorable interaction energies (Hernandez *et al.*, 2009). The cluster with the highest total interaction energy is consistent with known equivalent receptor binding domain (RBD) in H3, including mutations A143V and L235Q which are associated with antibody recognition and host receptor binding. Hence, the equivalent RBD of H3 was used as tentative binding regions to dock host receptor analogs to HA protein, namely 130-loop (139-146), 190-helix (192-204), 220-loop (228-237) and some conserved residues (106, 152, 160-162).

Table 5.2: Mutations of the influenza TW17 strain compared with the reference SH13 strain.

Impact	Protein	Mutations
Viral oligomerization interfaces or binding small ligands	HA	I56T, A130P, S136N, I138T, A143V, K182E, L235Q, M245I, A310T, I335V, G338A, E396A, E403K, S499R
	M2	E24D
	NA	M26I, M72I, Y166H, A210V, S242P, R289K, N322S
	PA	G66S
	[a]PB2	M570I, E627K
Host receptor binding	HA	L235Q, E396A
Host specificity shift	PB2	I292V, E627K
Glycosylation	HA	S136 N, I138T
Antibody recognition sites	HA	I56T, A130P, S136 N, I138T, A143V, L235Q, I335V, E396A, S499R
	NA	S242P
Drug binding	NA	S242P, R289K

[a]Best reference hit strain for PB2 is the influenza A/Duck/Guangdong/E1/2012(H10N8)

5.3.2 Molecular docking predicting the optimal conformations with host receptor analogs

The structures of HA proteins and host receptor analogs used for protein-ligand docking

To conduct protein-ligand docking and molecular dynamics simulation for analyzing the binding specificity of the three representative strains (SH13, QZ15, and TW17), the structures of HA protein and host cell receptors are needed.

The avian and human receptor analogs, denoted as LSTa and LSTc, were obtained from PDB: 5E2Z (Rose *et al.*, 2016) and 2YP3 (Lin *et al.*, 2012) respectively. Crystal structures of SH13 HA with the LSTa (PDB ID: 4N5K) and LSTc (PDB ID: 4N60) could be found in the Protein Data Bank (Xu *et al.*, 2013). However, extra four residues have been observed in the HA protein of the new TW17 strain (Section 5.3.1), which makes it difficult to get reliable HA structure of TW17 through simple manual substitutions and energy minimization. Therefore, the HA structures of the other representatives were predicted from homology modeling using SWISS model (Biasini *et al.*, 2014; Arnold *et al.*, 2006).

First, the primary sequences were first searched with BLAST (Altschul *et al.*, 1997) and HHBlits (Remmert *et al.*, 2012) among the SWISS-MODEL template library (Kiefer *et al.*, 2009). Three templates with the highest quality among the 1218 results were selected for model construction. Models were built based on the target-template alignment using ProMod3 (Šali and Blundell, 1993). The model constructed with the highest Global Model Quality esti-

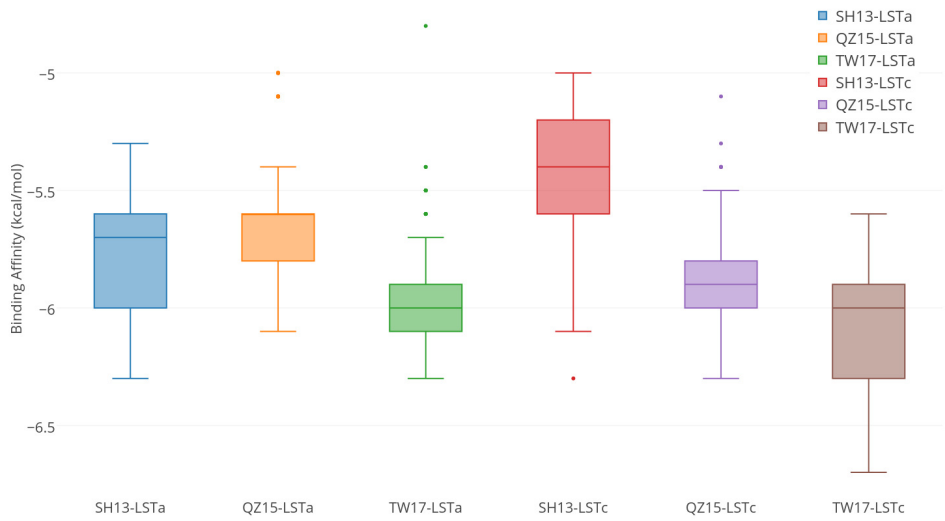


Figure 5.5: Binding affinity of host receptor analogs with the H7N9 HA proteins. *SH13-LSTa stands for the docking of LSTa to the HA protein of SH13 strain and so forth.*

mation (GMQE), a quality estimation combining properties from the target-template structure alignment, was selected for molecular docking (Benkert *et al.*, 2011).

The predicted optimal HA-receptor conformations and receptor binding preference inferred from molecular docking

With the HA structures and host receptor analogs, molecular docking can be used to estimated the pose of conformations and roughly compare the binding affinity of complexes. The molecular docking experiments were conducted using QuickVina 2, an improved version of QuickVina by optimizing the local search of docked conformation candidates (Alhossary *et al.*, 2015). The receptor analogs LSTa and LSTc were docked respectively to the receptor binding domain of the H7N9 HA protein (sites 106, 139–146, 152, 160–162, 192–204, 228–237). All available rotatable bonds of receptor analogs, LSTa and LSTc, were activated to ensure flexibility. The top conformation with the optimal binding affinity among 500 independent docking experiments was selected for molecular dynamics simulation. Results of binding affinities for each group of experiments are presented in Figure 5.5.

Comparing the SH13-LSTc, QZ15- LSTc and TW17-LSTc, it was observed that the HA protein acquired enhanced ability to bind both LSTa and LSTc as the viral strain evolved from SH13 to QZ15 and TW17. The observations were supported by Student's T-test, shown in Table 5.3, which was used to assess the significance of group mean difference of docking. Generally the HA of avian influenza virus has binding preference to LSTa. Both types of receptors (LSTa and LSTc) exist in human respiratory tract. The LSTa is abundant in human lower respiratory tract, while the LSTc is abundant in the upper. The enhanced binding ability to LSTc suggests easier infection to human and a higher potential of airborne transmission, which may partially explain the current epidemic wave.

Furthermore, the best and worst conformations were superimposed to check the binding

Table 5.3: T-test for HA-receptor docking experiments (N= 500)

Group 1	Group 2	[a]Mean Difference (kcal/mol)	99% Confidence Interval (kcal/mol)	*p*-value (one-tailed)
TW17-LSTa	SH13-LSTa	−0.205	(− 0.237, − 0.173)	< 0.0001
TW17-LSTc	SH13-LSTc	− 0.655	(− 0.696, − 0.614)	< 0.0001
SH13-LSTc	SH13-LSTa	− 0.345	(−0.307, − 0.383)	< 0.0001
TW17-LSTc	TW17-LSTa	−0.105	(− 0.141, − 0.069)	< 0.0001

[a]Mean difference = Mean (Group 1) – Mean (Group 2)

poses. The receptor binding analogs in the optimal and the worst complexes are colored pink and blue respectively. Figure 5.6-A and C show the SH13 and TW17 HA complexes with LSTa. The optimal complexes of HA-LSTa have SIA towards the 220-loop, while the worst poses have SIA towards the 130-loop. In contrast, the optimal poses of LSTc in complex with HA have SIA close to the 130-loop and the other way around for the worst poses (shown in Figure5.6-B and D). The superimposition also demonstrates that the reliable poses can be differentiated from the non-reliable ones from docking scores.

However, there have been concerns about the accuracy of predicted binding energy by molecular docking (Ramírez and Caballero, 2016). Docking score functions have advantages in searching for optimal conformations of ligands at efficiency, but less good at describing the binding energy than atomic scale force fields. Therefore, the optimal docked complexes were selected to conduct molecular dynamics simulation for analyzing residue-ligand interactions.

5.3.3 Molecular dynamics simulation revealing residues that contributing the the enhanced binding with host cell receptors

Molecular dynamics simulation

To observe HA-receptor interaction and the change of residue contributions, the HA complexes of SH13 and TW17 with LSTa and LSTc, namely SH13-LSTa, SH13-LSTc, TW17-LSTa and TW17-LSTc, were subject to molecular dynamics simulation for 50 nanoseconds. The simulation process was conducted using GROMACS (Abraham *et al.*, 2015). Figure 5.7 describes a general process for conducting molecular dynamics simulation. The crystal structures of SH13-LSTa and SH13-LSTc were obtained from PDB (PDB ID: 4N5K and 5N60), while the complex structures of TW17-LSTa and TW17-LSTc were predicted from the molecular docking (described in section 5.3.2). The AMBER99SB-ILDN force filed was used to describe the system (Lindorff-Larsen *et al.*, 2010), generating topology files from PDB structures. All complexes were then solvated under the explicit TIP3P water model in a cubic box. Afterwards, a solvated, electroneutral system was assembled by adding counter ions (Jorgensen *et al.*, 1983).

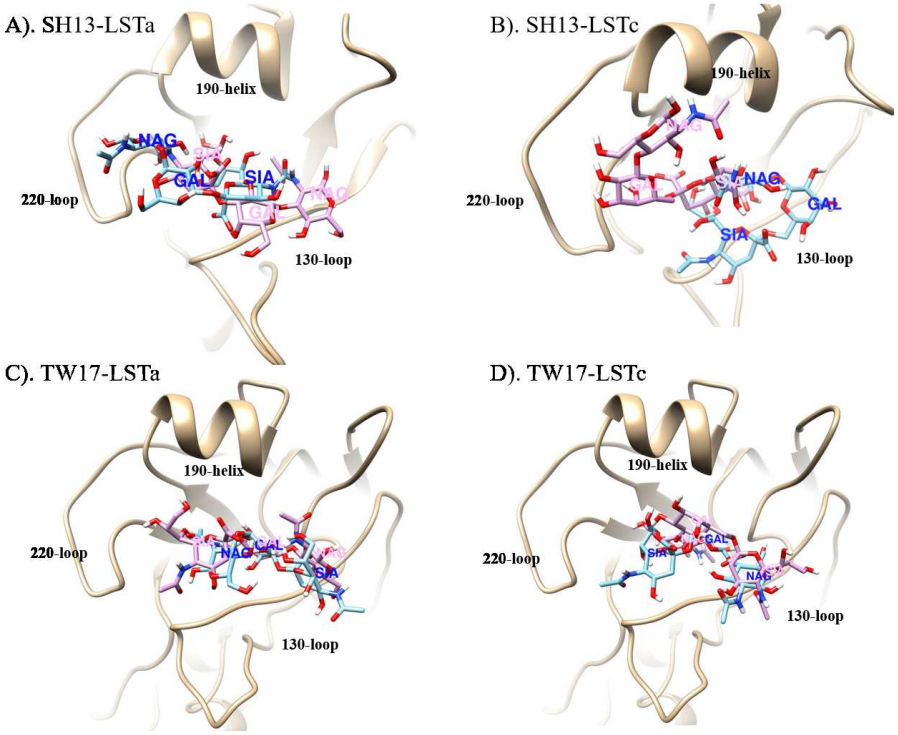


Figure 5.6: Superimpose the best and worst conformations of each HA-receptor binding predicted from molecular docking.

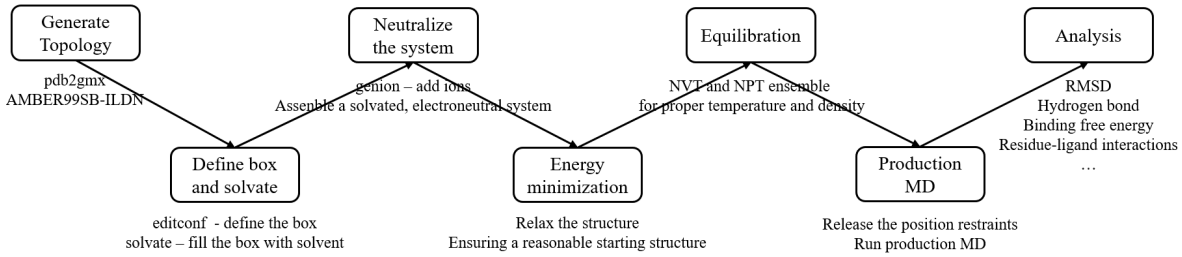
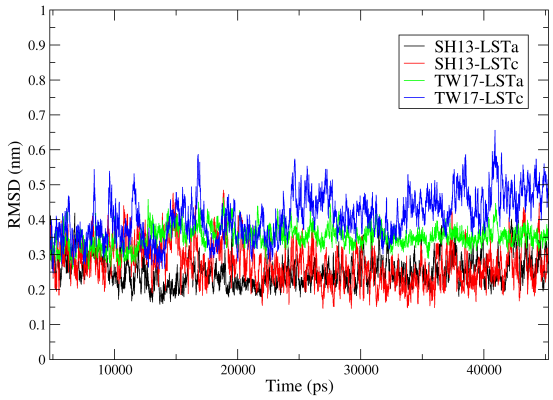


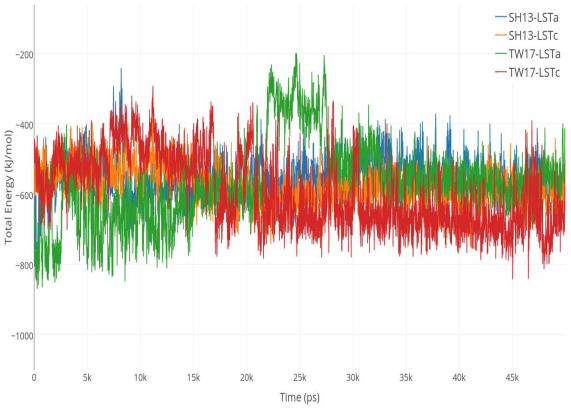
Figure 5.7: Typical steps for conducting molecular dynamics simulation.

And then, the steepest-descent energy minimization was applied for each complex to relax the system, ensuring a reasonable starting system. Position restrains for both NVT and NPT equilibration were conducted for 100 ps with modified Berendsen thermostat and Parrinello-Rahman pressure coupling (Berendsen *et al.*, 1984; Parrinello and Rahman, 1981). Temperature, pressure, density and total energy were all well equilibrated before running the production MD simulation for 50 ns.

The root-mean-square deviation (RMSD) of C_α atoms for each HA-LSTa/LSTc complex was visualized in Figure 5.8a. As observed, all the trajectories fluctuated within a small range. The TW17-LSTc complex has the largest fluctuation among the four trajectories, but still in a small range (around 0.45 nm). Besides, the VdW energy, electrostatic energy and total interaction energy of the four complexes were investigated. The Molecular Mechanics–Poisson Boltzmann Surface Area (MM-PBSA) was applied to estimate the binding free energy ΔG_{bind},



(a) Monitoring the RMSD of C_α atoms for SH13/TW17-LSTa/LSTc from the starting coordinates.



(b) Comparing the HA-SIA total vacuum MM energy for SH13/TW17-LSTa/LSTc during the whole MD simulation process.

Figure 5.8: Analysis of RMSD and total vacuum MM energy during the whole MD simulation.

which is composed of the binding free energy in the vacuum phase $\Delta G_{bind,vac}$ and the solvation free energy $\Delta G_{bind,solv}$ as shown in Equation (5.1). The solvation free energy is defined as the difference of solvation values of complex, receptor and ligand (Equation 5.2). The solvation energy for each component is composed of polar and non-polar energy, denoted as ΔG_{polar} and $\Delta G_{non-polar}$ respectively (Equation 5.3. The ΔG_{polar} is derived from the Poisson Boltzmann (PB) equation and the $\Delta G_{non-polar}$ is computed by the solvent accessible surface area (SA). The calculations of binding free energy was implemented using the *g_mmpbsa* package (Kumari *et al.*, 2014).

$$\Delta G_{bind} = \Delta G_{bind,vac} + \Delta G_{bind,solv} \tag{5.1}$$

$$\Delta G_{bind,solv} = \Delta G_{solv,complex} - (\Delta G_{solv,receptor} + \Delta G_{solv,ligand}) \tag{5.2}$$

$$\Delta G_{solv} = \Delta G_{polar} + \Delta G_{non-polar} \tag{5.3}$$

Figure 5.8b compares the total energy of the four complexes. The total energy of TW17-LSTa fluctuates obviously around 25 ns. Each trajectory fluctuated within a small range after 30 ns. Therefore, the last 20 ns frames were selected for further analysis. Table 5.4 lists the average total binding energy values of the four HA-receptor complexes. The mutant HA of TW17 reached the largest binding energy with the human receptor analog LSTc at -664.779 kJ/mol. Besides, the mutant HA had enhanced binding with both LSTa and LSTc. The binding energy with LSTa and LSTc were increased by 69.67 kJ/mol and 13.58 kJ/mol respectively. As to the binding preference, both the HA protein of SH13 and TW17 had higher binding energy with LSTc, indicating binding preferences for LSTc. The binding preference of SH13 with LSTc may partially explain the first the avian influenza virus outbreak among human in 2013. What's more, the enhanced binding with LSTa and LSTc is consistent with the outbreak and the high pathogenicity of the circulating strain in early 2017.

To further test this hypothesis, another 50 ns of molecular dynamics simulation was con-

Table 5.4: Average total binding energy (kJ/mol) of the HA-LSTa/LSTc complexes.

	LSTa	LSTc	[a]ΔE_1
SH13	− 541.559	− 595.111	+ 53.553
TW17	− 555.135	- 664.779	+ 109.644
[b]ΔE_2	+ 13.576	+ 69.667	

[a]Binding preference of HA protein: $\Delta E_1 = \Delta E_{HA,\ LSTa} - \Delta E_{HA,\ LSTc}$

[b]Difference of HAs binding to receptors: $\Delta E_2 = \Delta E_{SH13,\ receptor} - \Delta E_{TW17,\ receptor}$

Table 5.5: Interacting residues and the number of hydrogen bonds between host receptor analogs and the HA proteins of influenza A/H7N9.

	SH13	TW17	[1]#H-bond
LSTa	143, 163, [2]**164**, 204	139, 140, 141, 142, 144, 145, **164**, 199, 235	7/13
LSTc	139, **140**, **143**, **145**, 163, 192	106, **140**, 141, 142, **143**, **145**, 164, 192, 235	7/9

[1] *Numbers of hydrogen bonds between the host cell receptor analogs and the HA proteins: SH13/TW17.*
[2] *Conserved binding residues are highlighted as* ***bold****.*

ducted. Appendix B shows the comparison of two rounds of 50 ns molecular dynamics simulation for each system and the average total binding energy. Both rounds of simulation indicated that the mutant TW17 HA had enhanced binding with LSTa and LSTc, a binding preference with LSTc, and the largest average binding energy among the four systems. The SH13 strain also had a binding preference for LSTc.

Binding free energy and residue-ligand interaction energy calculation

It has been reported that the hydrogen bonds generally do not contribute much to the interaction energy, but plays a significant role in host specificity. The residues involved in the HA-receptor interaction and the number of hydrogen bonds for each system are listed in Table 5.5. The mutant HA of TW17 formed more hydrogen bonds and had more residues interacting with both LSTa and LSTc than the HA of SH13. The interacting bonds are visualized in Figure 5.9. Residues H192, L235 and S236 on the HA of SH13 interacted with both LSTa and LSTc, the site numbers have been converted to equivalent sites in the TW17 HA protein. Residues R139, T140, G142 and N164 on the HA of TW17 interacted with both LSTa and LSTc. Residue A143 on the HA of SH13 interacted with LSTa, but in the HA of TW17, V143 interacts with LSTc. Furthermore, for each snapshot during the whole MD simulation, the total interaction energy was decomposed to observe the residues contributions. The average contribution of each residue to the total binding energy was calculated. Only the residues involved in the HA-receptor interactions in the optimally docked complexes were focused. Figure 5.10 shows the average contribution of each residue. As seen, the residue R139 on the HA of TW17 enhanced the viral binding with both LSTa and LSTc. Besides, the binding preference was

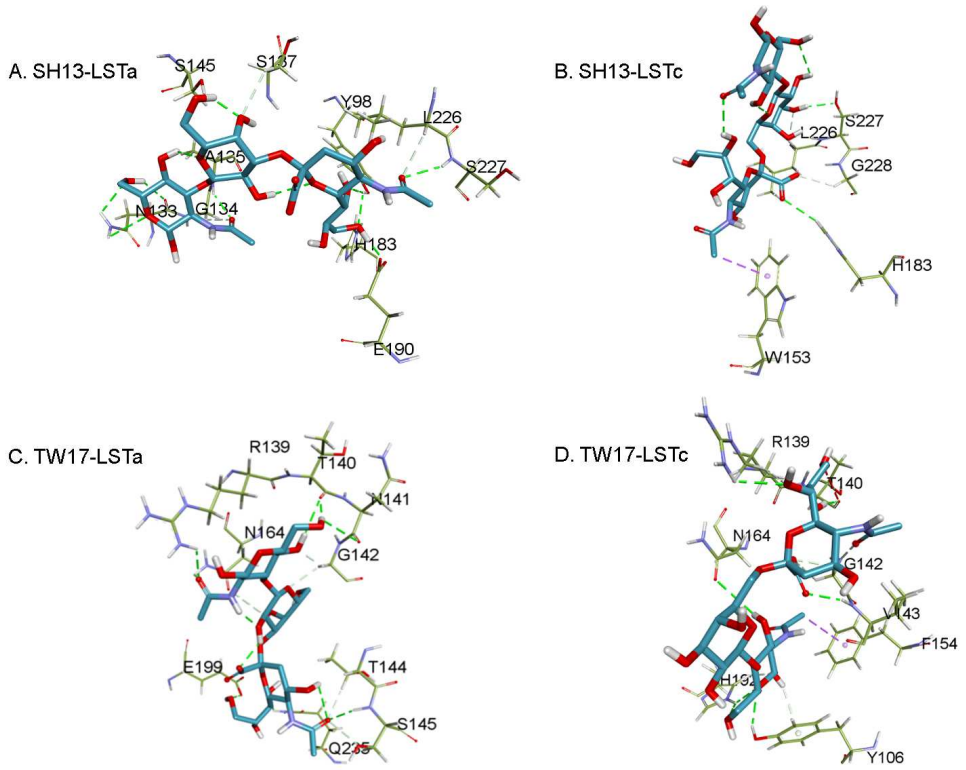


Figure 5.9: Visualization of HA-receptor interactions in the optimally docked complexes of SH13-LSTa, SH13-LSTc, TW17-LSTa and TW17-LSTc.

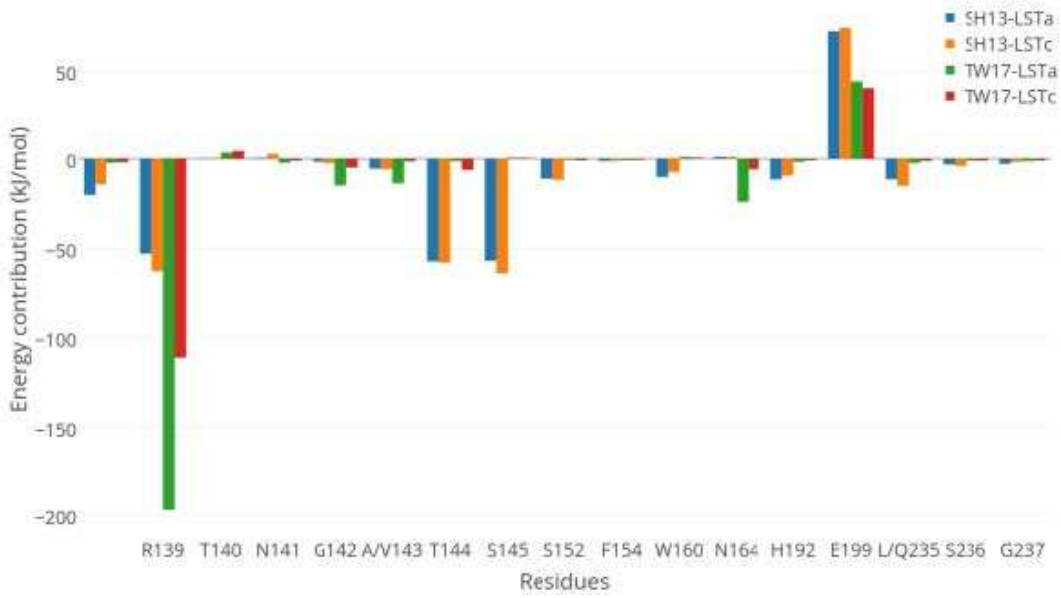


Figure 5.10: Average energy contribution of residues that involved in receptor-ligand interactions in the optimally docked complexes.

quantified by the difference of residue contribution to binding with LSTa and LSTc, clarified as Equation (5.4).

$$\Delta\Delta G_1 = \Delta G_{HA-LSTa} - \Delta G_{HA_LSTc} \tag{5.4}$$

Residues on the HA of SH13 strain showed small difference of energy contribution to LSTa and LSTc ($<$ 10 kJ/mol). For instance, the R139 had mild preference for LSTc. In contrast, the mutant HA of the TW17 strain had more residues showing clear binding preferences. The residues R139, V143, N164 showed preference for LSTa while the K202 showed preference for LSTc. The residue R139, bearing a clear preference for LSTa, enhanced binding of TW17 with both LSTa and LSTc. The top 10 residues showing binding preference to either LSTa and LSTc are highlighted in Figure 5.11a.

Similarly, the impact of a mutation is quantified as the difference of residue contribution to the total interaction energy before and after mutation, defined as Equation (5.5). Figure 5.11b shows the top 10 residues that either strengthen or weaken the HA binding with receptors after mutations. As seen, the residue R139 and K202 significantly strengthen the HA binding with LSTa and LSTc respectively (with $|\Delta\Delta G| > 100$ kJ/mol). R139 locates at a new potential N-glycosylation site introduced by the mutation S136N and I138T, which may explain the increased binding with receptors. E199 and K202, locating at the 190-helix where no mutations were observed, enhanced the binding with LSTc. The mutation K182E near the 190-helix might have affected contributions of E199 and K202 to the HA-LSTc binding.

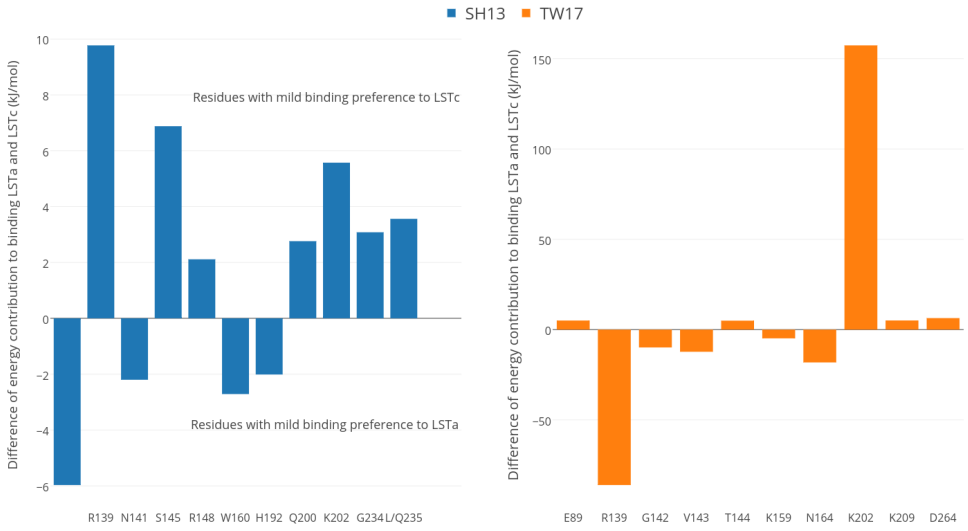
$$\Delta\Delta G_2 = \Delta G_{SH13-receptor} - \Delta G_{TW17-receptor} \tag{5.5}$$

The obtained results are novel and specific to the influenza A/Taiwan/1/2017(H7N9) strain, providing deeper understanding of the impact of HA mutations and the mechanism of receptor recognition. In addition, the process should be applicable to analyse the impact of other mutations on the binding to small ligands.

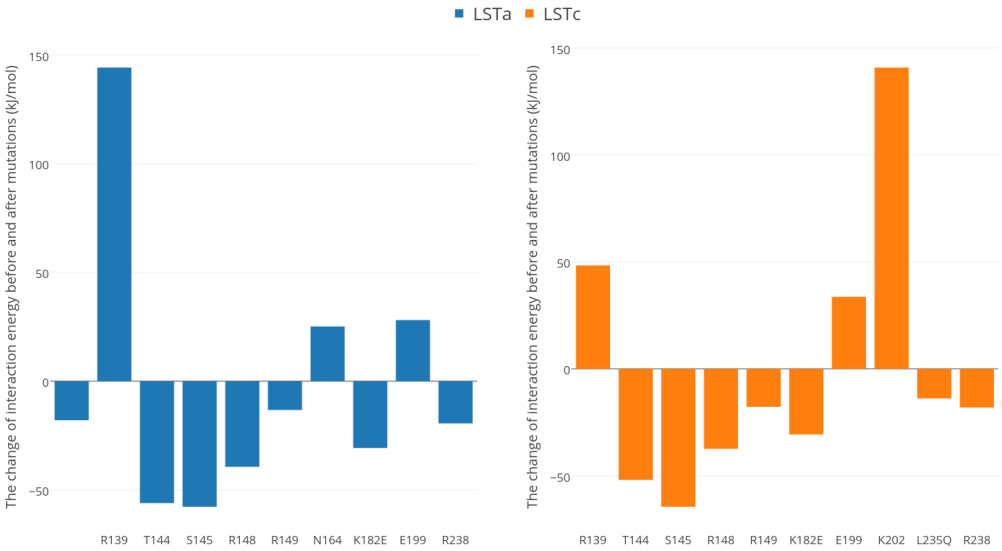
5.4 Summary

It has been taken as a well-accepted knowledge that the HA protein is a key determinant for the host specificity of influenza viruses (Lazniewski *et al.*, 2017). However, the mechanism of adaptive mutation leading the shift of host-specificity still needs to be explored. Structure-based analysis on the HA-receptor binding is an alternative to quantify the receptor binding preference of influenza viruses.

This chapter takes a new emerging influenza A/H7N9 strain as the subject, conducting both sequence and structural analyses to assess its potential in circulating among humans. The sequence analysis highlighted mutations in protein functional domains of influenza viruses. Molecular docking and molecular dynamics simulation revealed that the HA enhanced the binding with both avian and human receptor analogs. Besides, the decomposition of residue contributions revealed several residues being responsible for the change of receptor binding



(a) The top 10 residues in HA that show binding preference for LSTa or LSTc.



(b) The top 10 residues in HA that strengthen or weaken the HA binding with LSTa or LSTc.

Figure 5.11: Comparing residues contribution to HA-receptor binding energy.

preference. The obtained results are specific to the mentioned strain, shedding light on the impacts of HA mutations and the mechanisms of receptor recognition. What is even more important is that the pipeline of analysis is applicable to analyzing the impacts of other mutations on the binding of proteins with small ligands, helping to quantify the receptor binding specificity of influenza viruses.

Chapter 6

Profiling the evolution, mutations and receptor binding specificity of viral strains from outpatients and inpatients in Singapore

This chapter presents an example of profiling the evolution, mutations and receptor binding specificity of influenza viruses. The influenza A/H1N1, A/H3N2 and B viruses causing mild infections among outpatients and severe infections among inpatients during 2012-2015 in Singapore were sequenced and computationally investigated, using both sequence analyses and structural analyses.

6.1 Introduction

6.1.1 The burden of influenza infections in Singapore

Pandemics and seasonal epidemics of influenza have made substantial social and economic impacts in Singapore. Table 6.1 summarizes the reported infections and deaths in Singapore raised by influenza pandemics. With the improvement of public health care, the mortality during flu pandemics has been reduced, but the infections remain high. The infections of influenza during non-epidemics are also under surveillance. There is an estimation of $630,000$ influenza infections a year, causing $520,000$ doctor visits, $315,000$ days of sick absence from work (Ng *et al.*, 2002). Influenza-associated hospitalizations in tropical regions have also been monitored. The hospitalization rate in Singapore was about 28.3 and 29.6 per $100,000$ person-year during 2004-2008 and 2010-2012 respectively (Ang *et al.*, 2014). The age distribution of the hospitalized patients followed a J-shaped pattern, where the elders (>75 years of age) and the infants (<6 months of age) showed >47 times and >26 times higher hospitalized rate than the adults ($25-44$ years old). The high risk of server infections to the elders and the very young

Table 6.1: Infections and deaths in Singapore during the past influenza pandemics.

Pandemics	Infections	Death	References
1918 A/H1N1	–	3500	Lee *et al.* (2008)
1957 A/H2N2	$> 77,000$	680	Lee *et al.* (2008)
1968 A/H3N2	$> 127,000$	> 540	Lee *et al.* (2008)
2009 A/H1N1	$> 270,000$	18	Cutter *et al.* (2010)

highlights the importance of an effective vaccination program.

6.1.2 The necessity for regionally focused influenza study

The WHO recommends the composition of vaccines every six-month for the following influenza season, based on which the national vaccine regulatory agencies will facilitate the development, producing and licensing of influenza vaccines. The recommendations are based on the data reported from WHO collaborating centers and collected in the WHO Global Influenza Surveillance and Response system (Duncan *et al.*, 2012).

However, the recommended vaccines composition may mismatch with the circulating strains, leading to the reduced vaccine efficacy. For example, during 2012-2013, vaccination was more effective against influenza A/H1N1 and B viruses than against the influenza A/H3N2 (Ho *et al.*, 2014). The low effectiveness of vaccination to influenza A/H3N2 could be partly explained by the observation that 84% of the 206 clinical positive influenza A/H3N2 samples collected in 2009-2013 mismatched with the recommended vaccine candidate strain A/Perth/16/2009(H3N2) for the 2009-2012 influenza seasons (Lee *et al.*, 2015b). Besides, the study alerted that a patient with comorbidities could experience more severe illness if receiving a sub-optimal vaccine containing mismatches. Meanwhile, Lee *et al.* observed the circulating A/H3N2 strain in Singapore had a distinct mutation pattern on epitope domains from the Northern and Southern hemisphere. Such findings support the rational and necessity for regionally focused surveillance and more customized seasonal influenza compositions to improve the protection against locally circulating viral strains.

The Surveillance Program for Influenza under the Ministry of Health (MOH), Singapore, is a part of the WHO international laboratory-based surveillance network, providing weekly update and review on the influenza incidences among the suspicious patients in the community (Ang *et al.*, 2016). The first surveillance study of influenza incidences during 1972-1986 reported that the influenza A outbreaks occurred every year, while the influenza B outbreaks occurred at a longer interval (16-24 months) (Doraisingham *et al.*, 1988). Some surveillance studies have also been carried out in military setting (Lee *et al.*, 2011; Seah *et al.*, 2010; Yap *et al.*, 2012) and university setting (Tan *et al.*, 2015; Virk *et al.*, 2014, 2017). Similar temporal pattern of the influenza A and B incidences have been detected by those studies. Another surveillance study was carried out in hospital setting to monitor antiviral drug resistance of influenza A/H3N2 viruses in 2009-2013 (Lee *et al.*, 2015a). Nonetheless, only a few of these surveillance studies linked to genetic information, with one study reported an analysis of 34

full-genomes of 2009 pandemic influenza A/H1N1 viruses on campus (Virk *et al.*, 2017).

Nowadays, sequencing becomes easier and cheaper. Molecular surveillance of influenza viral strains at the genomic level is feasible and promising for helping uncover mutational patterns, profile viral binding specificity, and give insights into the design of vaccines. Such analyses should be carried out systematically and regularly. As part of this effort, two influenza genome datasets during 2012-2015 in Singapore were sequenced analyzed. One genome dataset was obtained from outpatients with mild influenza and seeking medical attention at the University Health Centre in the National University of Singapore (NUS), while the other dataset was obtained from inpatients with severe influenza and admitted to hospitals across Singapore. Computational analyses were conducted using both the nucleotide/protein sequences and inferred 3D protein structures. Molecular insights and patterns about the surveilled viruses, including the evolutionary, mutational and receptor binding preference, were uncovered in the background of vaccine strains and some other genomes available publicly.

6.2 Influenza viral samples and genome sequencing

Nasal swabs were collected from outpatient subjects between 20 to 55 years old attending the NUS University Health Center (UHC) at the National University of Singapore (NUS) or inpatient subjects from local Singapore hospitals, including Communicable Disease Center (CDC), Tan Tock Seng Hospital (TTSH), National University Hospital (NUH), and Singapore General Hospital (SHG). QuickNavi^{TM} – Flu rapid diagnostic test kit (Denka Seiken) was used to screen for influenza infected subjects and directly determine the type of the influenza virus infecting the subjects.

6.2.1 RNA extraction and reverse transcription

Totally, there were 47 influenza positive nasal samples, 27 of which obtained from outpatient subjects and 20 from inpatient subjects. Table 6.2 lists the IDs and collection date of the samples in this study, where the IDs of outpatient specimens start with G2 (indicating the use of G-II breath sampler machine for collecting samples) while the IDs of the inpatient specimens start with the abbreviation of the hospital name. The study on these specimens were approved by the National Healthcare Group (NHG) Domain Specific Review Board (DSRB) with reference number 2011/01883 and Nanyang Technological University (NTU) Institutional Review Board (IRB) with reference number IRB-2015-12-023. Written informed consents were obtained from the participants before sample collection.

Influenza virus from each specimen was propagated in MDCK cells (ATCC, CCL-34). The tissue culture of infective fluid was harvested after 72 hours. Viral RNAs were then extracted from the MDCK-culture fluid using QIAamp Viral RNA Mini Kit (Qiagen, Valencia, CA, USA). Afterwards, reverse transcription was performed to produce viral cDNAs using SuperScript^{TM} III Reverse Transcriptase (Invitrogen, Carlsbad, USA) and universal primer for

Table 6.2: An overview of the collected influenza positive samples from outpatient and inpatient subjects during 2012-2015, including IDs, collection date, subtype/lineage. *Samples of the same subtype/lineage are grouped with the same color.*

Institution	Sample ID	Date	Subtype/Lineage	Sample ID	Date	Subtype/Lineage	Sample ID	Date	Subtype/Lineage
NUS University Health Centre (UHC)	G2-5.1	2/12/2013	A/H3N2	G2-19.1	13/2/2014	A/H1N1	G2-31.1	3/4/2014	A/H3N2
	G2-6.1	2/12/2013	A/H3N2	G2-20.1	13/2/2014	A/H1N1	G2-34.1	21/4/2014	B/Yamagata
	G2-7.1	26/11/2013	B/Victoria	G2-22.1	13/2/2014	A/H1N1	G2-36.1	10/6/2014	B/Yamagata
	G2-8.1	9/12/2013	A/H3N2	G2-23.1	13/2/2014	A/H1N1	G2-43.1	18/6/2014	B/Yamagata
	G2-9.1	9/12/2013	A/H3N2	G2-24.1	18/2/2014	B/Yamagata	G2-44.1	2/7/2014	A/H1N1
	G2-10.1	23/12/2013	A/H3N2	G2-25.1	6/3/2014	A/H1N1	G2-46.1	8/7/2014	A/H3N2
	G2-13.2	15/1/2014	B/Victoria	G2-26.1	6/3/2014	A/H3N2	G2-51.1	12/8/2014	A/H3N2
	G2-14.1	15/1/2014	B/Victoria	G2-27.1	6/3/2014	A/H1N2	G2-52.1	27/8/2014	A/H3N?
	G2-15.1	16/1/2014	B/Victoria	G2-29.1	21/1/2014	A/H3N2	G2-63.1	14/11/2014	A/H3N2
Communicable Disease Centre (CDC)	CDC-64	2012	A/H3N2	CDC-91	2012	A/H3N2	CDC-149	2012	A/H3N2
	CDC-73	2012	A/H3N2	CDC-109	2012	A/H3N2	CDC-204	2012	A/H3N2
	CDC-85	2012	A/H3N2	CDC-126	2012	A/H3N2	-		
	CDC-90	2012	A/H3N2	CDC-148	2012	A/H3N2			
National University Hospital (NUH)	NUH-6	2012	A/H3N2	-					
Singapore General Hospital (SGH)	SGH-A	30/7/2012	A/H3N2	SGH-D	29/1/2013	A/H3N2	SGH-G	10/9/2014	A/H1N1
	SGH-B	2/3/2012	A/H3N2	SGH-H	28/3/2015	A/H3N2	-		
	SGH-C	6/12/2012	A/H1N1	SGH-F	5/7/2014	A/H1N1			
Tan Tock Seng Hospital (TTSH)	TTSH-13A	12/7/2012	A/H3N2	TTSH-69	28/9/2012	B/Yamagata	-		

influenza A (Lee *et al.*, 2013) or influenza B (World Health Organization, 2014) respectively. The reverse transcription was carried out at 50°C for 30 min, followed by enzyme inactivation at 95°C for 1 min. Subsequently, 40 cycles of polymerase chain reaction (PCR) was performed, each consisting of denaturation at 95°C for 15 sec, followed by annealing and extension at 72°C, using Biometra T-Personal Thermal Cycler (Biometra, Goettingen, Germany) or ABI 2400 thermal cycler (Applied Biosystems, CA, USA).

For influenza A positive samples, one-step qPCR (Lee *et al.*, 2013) was performed to differentiate A/H1N1 and A/H3N2 samples. For the influenza B samples, conventional two-step RT-PCR was performed using lineage-specific primers (World Health Organization, 2014) to differentiate the Victoria and Yamagata lineages; and amplified products of 284 bp and 388 bp, respectively were visualized by agarose gel electrophoresis.

The subtype/lineage for each sample has also been listed in Table 6.2. Figure 6.1 presents an overview of the subtypes/lineages of the collected samples. In the 27 outpatient samples, there are 11 influenza A/H3N2, 6 influenza A/H1N1, 1 influenza A/H1N2, 1 influenza A of unknown subtype H3Nx, 4 influenza B/Victoria and 4 influenza B/Yamagata.

6.2.2 NGS and Sanger for genome sequencing

Viral cDNAs were sequenced using the Sanger method and Illumina MiSeq Next Generation Sequencer (NGS). For the influenza A virus, NGS was used to determine the full genomes following an optimized protocol for the influenza virus (Lee *et al.*, 2016). The Sanger method

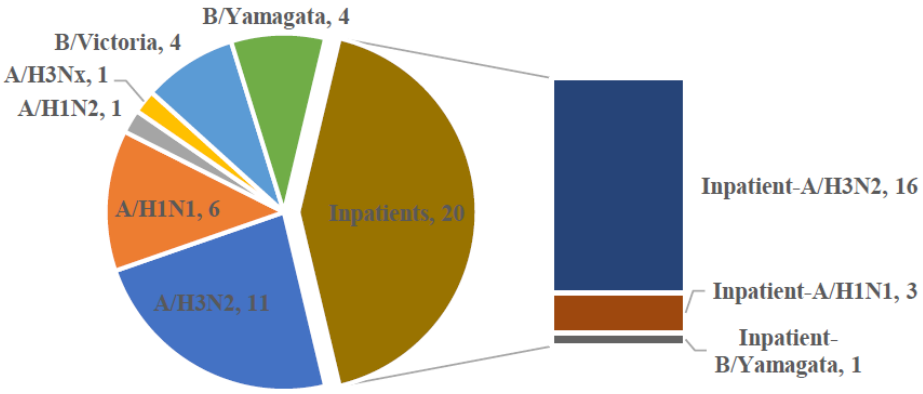


Figure 6.1: Subtypes of collected samples.

was used to sequence all HA1 of influenza A viruses and sequence the genomes of influenza B viruses. Influenza A/H3N2 primers (Lee *et al.*, 2013), influenza A/H1N1 primers (Deng *et al.*, 2015) and 18 sets of influenza B primers (Tewawong *et al.*, 2015) were used for sequencing. An addition set of primers (Chi *et al.*, 2005) was used for sequencing the HA1 gene of influenza B viruses. The HA1 segments from Sanger sequencing were used to validate the NGS data. Overall, only a number of nucleotide mismatches were observed in the minority of the alignments between HA1 from Sanger and HA from NGS. If a mismatch was observed, nucleotide symbol in NGS sequence was substituted with nucleotide symbol in associated Sanger sequence. Then, each influenza virus segment was input to NCBI Influenza Virus Sequence Annotation Tool (Bao *et al.*, 2007) for identifying potential sequence errors as well determining the coding regions. Errors in some segments were fixed by performing another sequencing with Sanger method or removing low quality regions observed in the chromatograms.

6.3 Molecular evolution study of sampled sequences

6.3.1 BLAST analysis

Then, to obtain insights about the mixture of the sampled genomes, the hemagglutinin (HA) and neuraminidase (NA) nucleotide segments of the influenza genomes being studied (except the NA of G2-52.1 isolate that was failed to be sequenced using both NGS and Sanger methods due to unknown reason) were BLASTed against influenza virus sequences in GenBank. The top BLAST hits for the HA and NA of a particular queried genome were usually from different influenza virus strains in the database (except for sample G2-10.1 and G2-44.1), although the HA and NA of the virus hits were often available in GenBank. Further inspection verified that the associated NA of top virus hit given by BLASTing HA, if available, was usually scored lower than the top NA hit and vise versa.

The other segments of the outpatient samples G1-27.1 and G1-52.1, which are of are of A/H1N2 subtype and unknown A/H3Nx subtype, are also investigated. For sample G1-27.1, the

Table 6.3: The WHO recommended vaccine strains from 2010 to 2020.

	A/H1N1	A/H3N2	B	B (For quadrivalent vaccines)
2010-2011	A/California/7/2009 (H1N1)pdm09-like virus	A/Perth/16/2009 (H3N2)-like virus	B/Brisbane/60/2008-like virus	-
2011-2012				-
2012-2013		A/Victoria/361/2011 (H3N2)-like virus	B/Wisconsin/1/2010-like virus	-
2013-2014			B/Massachusetts/2/2012-like virus	B/Brisbane/60/2008-like virus
2014-2015		A/Texas/50/2012 (H3N2)-like virus		
2015-2016		A/Switzerland/9715293/2013 (H3N2)-like virus	B/Phuket/3073/2013-like virus	
2016-2017		A/Hong Kong/4801/2014 (H3N2)-like virus	B/Brisbane/60/2008-like virus	B/Phuket/3073/2013-like virus
2017-2018	A/Michigan/45/2015 (H1N1)pdm09-like virus			
2018-2019		A/Singapore/INFIMH-16-0019/2016 (H3N2)-like virus	B/Colorado/06/2017-like virus (B/Victoria/2/87 lineage)	B/Phuket/3073/2013-like virus (B/Yamagata/16/88 lineage)
2019-2020	A/Brisbane/02/2018 (H1N1)pdm09-like virus	A/Kansas/14/2017 (H3N2)-like virus		

viral genome contains PB2, PB1, PA, HA and M segments potentially from influenza A/H1N1, while the NP, NA and NS segments may originate from influenza A/H3N2. For sample G1-52.1, the BLAST results indicate that other segments besides HA were also closely related to influenza A/H3N2. Therefore, these segments of G1-52.1 were used for reconstructing the phylogenetic tree of influenza A/H3N2.

6.3.2 Phylogenetic tree analyses

Influenza genomes collected from 2009 to early 2018 were retrieved from NCBI and GISAID. The HA sequences for each subtype were grouped by year and then clustered using CD-HIT to select representatives at a similarity threshold of 0.97 (Li and Godzik, 2006). The genomes with those representative HA sequences were selected as the representative strains for analyses. The phylogenetic tree for each segment was constructed using the sequences from the representatives and the WHO recommended vaccine strains. Table 6.3 lists the recommended components for the trivalent and quadrivalent influenza vaccines from 2010 to 2020. The nucleotide sequences for each segment of influenza A/H1N1, A/H3N2 and B viruses were aligned using MUSCLE (Edgar, 2004) before using the BEAST (Bouckaert *et al.*, 2014) to infer time-based phylogenetic trees. Afterwards, the maximum clade credibility (MCC) time trees were produced with 10% burn-in and visualized with the ggtree package (Yu *et al.*, 2017).

Figure 6.2 shows the HA and NA phylogenetic trees of influenza A/H1N1, A/H3N2 and B viruses. As seen, for each subtype, the outpatient strains from G2 study and the inpatient strains from hospital study do not cluster into different clades, but tending to be mixed. For influenza A/H3N2, a group of inpatient cases were clustered together and close to the 2014-2015 vaccine strain influenza A/Texas/50/2012(H3N2) in both the HA and NA phylogenetic trees. The phy-

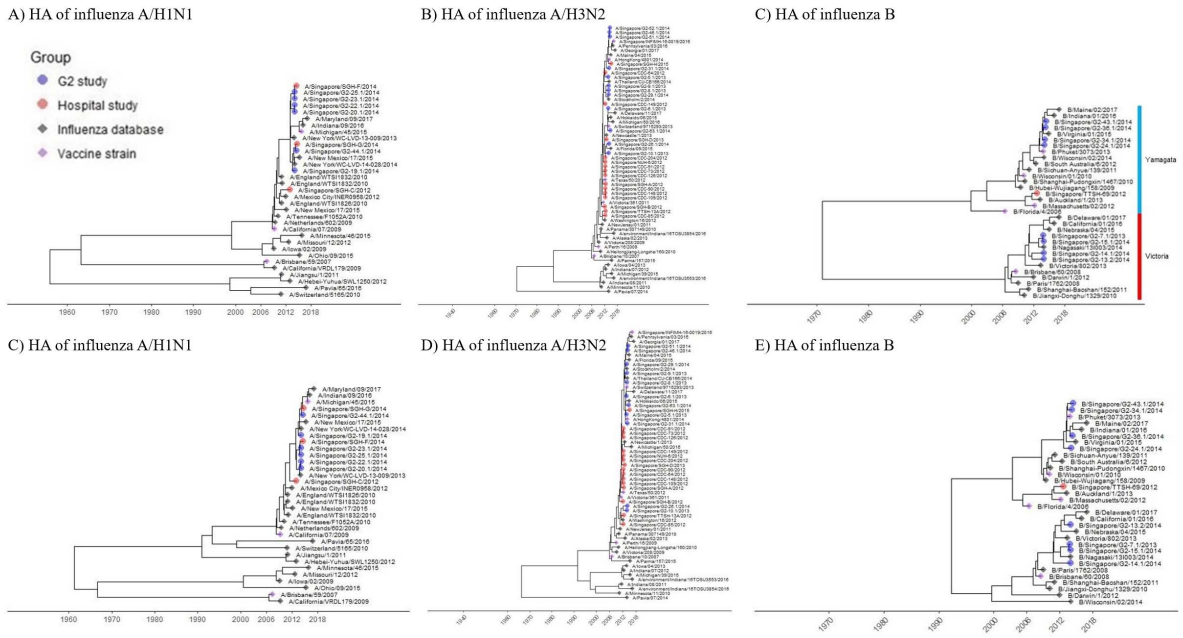


Figure 6.2: The HA and NA phylogenetic trees of influenza A/H1N1, A/H3N2 and B viruses.

logenetic trees constructed from other segments of influenza A/H3N2 suggested the same inference. For the influenza A/H1N1, both the inpatient and outpatient strains can be rooted from the influenza A/California/7/2009(H1N1), the dominant strain for the 2009 pandemic. For influenza B, the inpatient strain TTSH-69 originated from the 2013-2015 vaccine strain influenza B/Massachusettes/2/2012(Yamagata). The outpatient strains of B/Yamagata were mainly close to the 2015-2018 vaccine strain influenza B/Phuket/3037/2013, while the outpatient strains of B/Victoria (except the PA of G2-7.1 isolate) were close to the influenza B/Brisbane/60/2008 (a component of quadrivalent influenza vaccine from 2010-2012 and 2013-2018).

6.3.3 Phenotypically or epidemiologically interesting mutations and genomic signatures discriminating outpatient and inpatient samples

Coding sequences (CDSs) for the main influenza proteins, i.e., PB2, PB1, PA, HA, NP, NA, M1 and NS1, were identified using the Annotation Tool available in NCBI Influenza Virus Resources (Bao *et al.*, 2007), translated to proteins and then aligned using MUSCLE (Edgar, 2004).

Phenotypically or epidemiologically interesting mutations

The protein sequences were compared with the closest reference among vaccine strains using FluSurver (`https://flusurver.bii.a-star.edu.sg/`) to detect phenotypically or epidemiologically interesting mutations. For A/H1N2, which is not a part of vaccination program, the automatic best hit for each segment from FluSurver was used. For PB1, PB2, PA, HA and M2, the reference strain is A/Michigan/45/2015(H1N1). For NP, NA, NS1 and

NS2, the reference strain is A/Switzerland/9715293/2013(H3N2). For M1, the reference sequence is M1 A/California/07/2009(H1N1). The NP and NS2 sequences are identical with that of A/Switzerland/9715293/2013(H3N2). For A/H1N1, the reference strain was the influenza A/Michigan/45/2015(H1N1)pdm09; for A/H3N2, the reference strains were the influenza A/Singapore/INFIMH-16-0019/2016(H3N2) for NP and A/Hong Kong/4801/2014(H3N2) for other proteins. For B/Victoria and B/Yamagata, the references were B/Brisbane/60/2008 and B/Phuket/3073/2013 respectively. All the reference strains are compositions of quadrivalent vaccines for use in the 2017-2018 northern hemisphere or the 2018 southern hemisphere. Table 6.4 summarizes the mutations with the highest interest level, residing in functional domains or known to be related with glycosylation, receptor binding, drug resistance, virulence and etc.

For the sequenced influenza A/H1N1 strains, two mutations, namely N179S (H1N1pdm numbering 162; H3 numbering 162) in HA and K363N in NA, are known to be at viral oligomerization interfaces and can affect the binding with small ligand(s). Besides, N179S is at an antibody recognition site in HA protein. The mutation removes a potential N-glycosylation site, while the mutation K386N in NA protein creates a new potential N-glycosylation site. Notably, both mutations are not specific to inpatient isolates. Mutation N179S is observed in all isolates, while mutation K386 could be observed in SGH-C, G2-19.1, G2-20.1 and G2-22.1 isolates.

For influenza A/H3N2, more mutations of interest level 3 were observed and they were detected in HA, NA and PB2. The HA mutations T144V/A, A154S, S160N and K176T (128, 138, 144 and 160 respectively in H3 numbering) and NA mutation S245N were mostly observed in either inpatient strains or outpatient strains. The mutation N61S (H3 numbering 45) in HA were observed in most outpatient and inpatient isolates, except CDC-85 and TTSH-13A. The other mutation suggested by FluSurver was the D701N in PB2. It was only observed in the isolate SGH-H, while the PB2 of the rest of the isolates contained D in position 701.

For influenza B strains, two mutations Y180N and D211N were observed in HA, which are known to be determinants for host specificity and create a new potential N-glycosylation site for HA protein. The mutation D211N on HA was exclusively observed in the sequenced B/Yamagata strains. Meanwhile, two mutations N463D and T465A are predicted to disturb a potential N-glycosylation pattern. Notably, the mutations mutation Y180N on HA, N463D and T465A on HA were exclusively observed in the only inpatient influenza B isolate TTSH-69. For the influenza A/H1N2 isolate G2-27.1, a potential reassortment from influenza A/H1N1 and A/H3N2, the best references for each segment detected by FluSurver were recorded. The PB1, PB2, PA, HA and M1 and M2 were predicted to be originated from influenza A/H1N1, while the rest segments were predicted to be from A/H3N2. The mutation N179S on HA removes while D151N on NA creates a potential N-glycosylation site.

Table 6.4: Mutations of the positive influenza samples with high interest level. *F1: binding with small ligands; F2: viral oligomerization interfaces; F3: antibody recognition sites; F4: remove a potential N-glycosylation site; F5: create a new potential N-glycosylation site; F6: T-cell epitope presented by MHC molecules; F7: host specificity shift; F8: virulence; F9: drug resistance; F10: drug binding.*

Type/Subtype	Protein	Mutation	Phenotype									
			F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
A/H1N1	HA	N179S	✓	✓	✓	✓						
	NA	K386N	✓	✓			✓					
A/H3N2	HA	N24K		✓		✓		✓				
		N61S	✓	✓		✓						
		T144V/A	✓	✓	✓	✓			✓			
		A154S	✓	✓	✓				✓	✓		
		S160N	✓	✓	✓		✓					
		K176T	✓	✓	✓		✓		✓			
	NA	I222V	✓	✓	✓						✓	✓
		S245N	✓	✓	✓		✓				✓	✓
	PB2	D701N							✓	✓		
B	HA	Y180N	✓	✓	✓		✓		✓			
		D211N	✓	✓	✓		✓		✓			
	NA	N463D	✓	✓	✓	✓						
		T465A	✓	✓		✓						
A/H1N2	HA	N179S	✓	✓	✓	✓						
	NA	D151N	✓	✓			✓				✓	✓

Sites discriminating the inpatient and outpatient strains

For each protein alignment of influenza A/H1N1, A/H3N2 and B viruses, each site was considered as a feature. A OneRule classification for discriminating the inpatient and outpatient strains was conducted independently (Holte, 1993). The accuracy of classification is taken as the power of a site discriminating the inpatient and outpatient strains.

Figure 6.3 visualizes the accuracy of using each site to classify the inpatient and outpatient strains. For influenza B, all isolates of B/Victoria were from the outpatients. There was only one B/Yamagata isolate from inpatient (TTSH-69). Thus, only the B/Yamagata isolates were used in the classification. The sites with high power classifying inpatient strains from the outpatient strains were labeled in the figure. For influenza A/H1N1, 6 sites had an excel predicting power ($> 80\%$), including PB1-41, PB1-43, PB1-695, PA-73, NA-40 and NS1-111; for influenza A/H3N2, 15 sites had a predicting power over 70%, including PB2-588, PA-272, PA-668, PA-669, PA-675, HA-161 (H3 number: 145), HA-175 (H3: 159), HA-241 (H3: 225), HA-505 (H3: 489), NA-151, NA-221, NA-251, NA-267, NA-392 and NS1-26. For influenza B, the limited number of positive sample (only one inpatient isolate) restricted the significance of the accuracy of classification. Overall, there were more sites in HA and NA than in other proteins that could discriminate the single inpatient sample from the others.

6.4 Receptor binding specificity of sampled viruses

6.4.1 Molecular docking analyses of HA-receptor binding

Representative strains for influenza A/H1N1, A/H3N2, B/Victoria and B/Yamagata were selected from the phylogenetic tree analyses, including the inpatient and outpatient strains, the corresponding vaccine strains. To compare the receptor binding specificity of those strains, the structures of HA proteins were predicted from homology modeling. The host receptor analogs were isolated from structures co-crystallized with HA protein.

Table 6.5 listed the selected strains. Fist, the HA1 sequence for each strain was searched with BLAST (Altschul *et al.*, 1997) and HHBlits (Remmert *et al.*, 2012) against the SWISS-MODEL template library (Kiefer *et al.*, 2009). And then, an atomic-resolution model was constructed based on the sequence alignment using ProMod3 (Biasini *et al.*, 2013).The predicted structure with the highest Global Model Quality Estimation (GMQE) was selected as the HA protein structure (Benkert *et al.*, 2011). The optimal matching templates and the GMQE scores for the HA1 proteins of representative isolates were listed in Table 6.5.

The avian and host receptor analogs 3'SLN and 6'SLN were separated from the HA co-crystallized structures (PDB ID: 3UBQ and 3UBN respectively) (Xu *et al.*, 2012b). All rotatable bonds were activated before conducing molecular docking with the predicted HA1 protein structures using QuickDock2 (Alhossary *et al.*, 2015). A grid box covering the binding sites, namely the 130-loop (140-loop for flu B), 190-helix, 220-loop (240-loop for flu B) and some conserved sites, was predefined specifically for each HA protein (Wang *et al.*, 2007). The

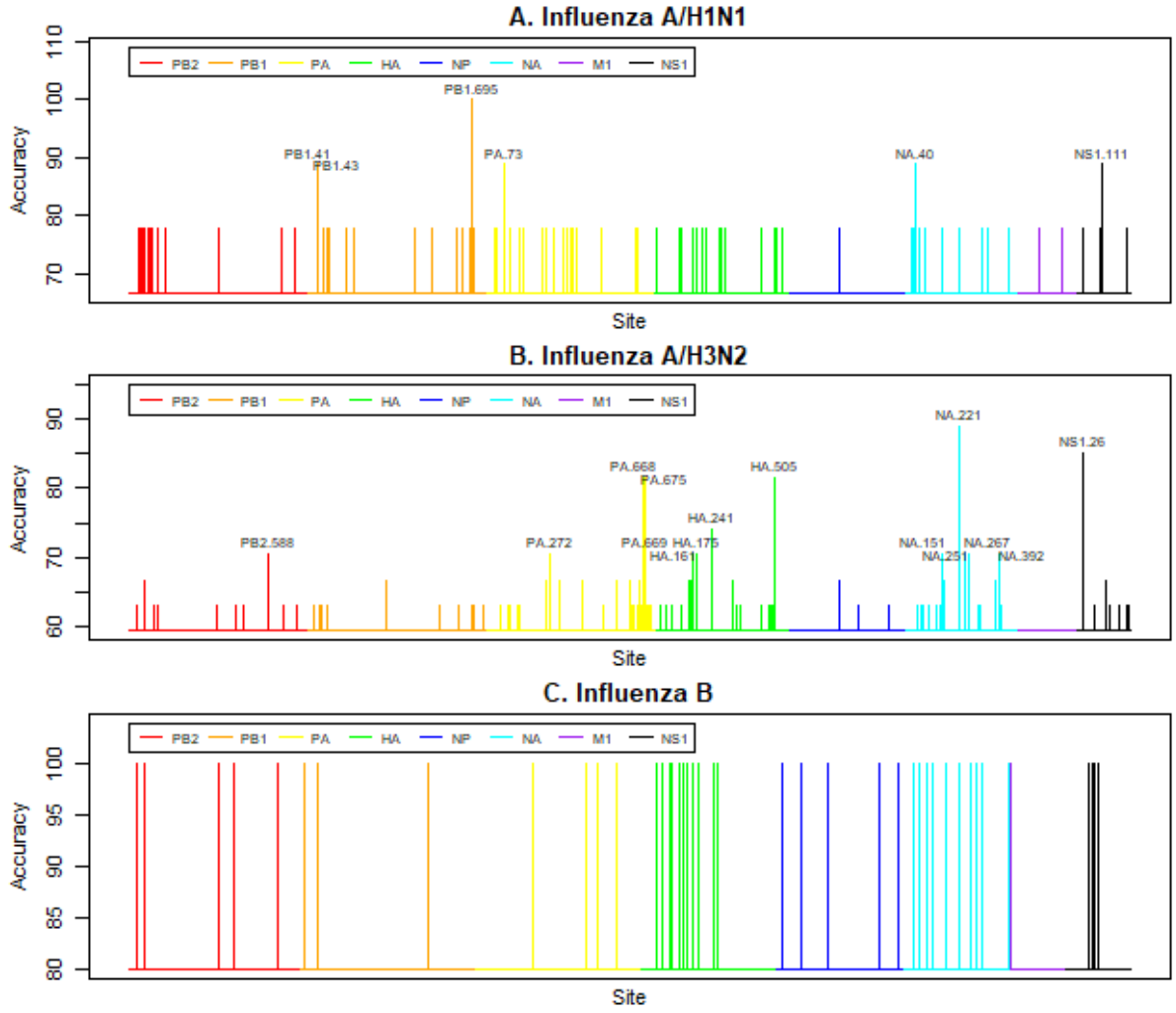
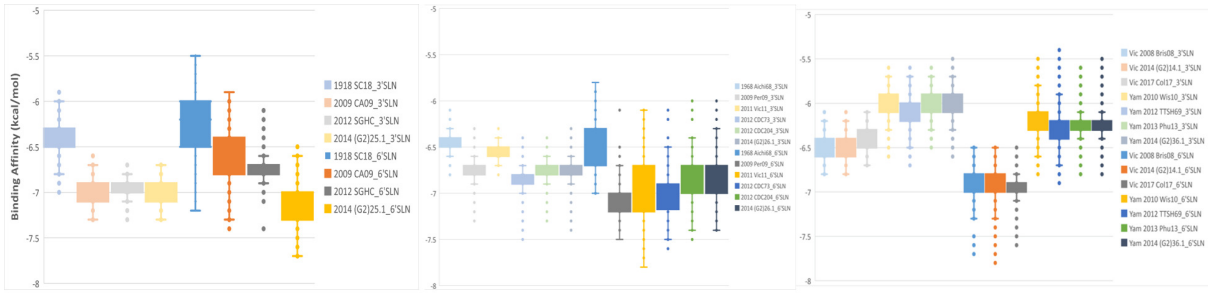


Figure 6.3: The power of sites in discriminating inpatient and outpatient strains.

Table 6.5: Representative strains and the predicted optimal matching template for the HA proteins 280 of influenza A/H1N1, A/H3N2 and B viruses.

Subtype or lineage		Representative strains							
		Outpatient	Template (GMQE)	Inpatient	Template (GMQE)	Vaccine	Template (GMQE)	Others	Template (GMQE)
A/H1N1		G2-25.1	4LXV (0.99)	SGH-C	4LXV (0.99)	CA09	3ZTN (0.99)	SC18	6D8W (0.95)
A/H3N2		G2-26.1	4WE8 (0.94)	CDC-204	4WE8 (0.94)	Vic11	4WE8 (0.95)	Aichi68	1HA0 (0.97)
				CDC-73	4WE8 (0.95)	Per09	4WE8 (0.93)	-	-
B	Yamagata	G2-36.1	4M40 (0.98)	TTSH-69	4M40 (0.99)	Phu13	4M40 (0.98)	Wis10	4M40 (0.98)
	Victoria	G2-14.1	4FQK (0.99)	-	-	Col17	4FQK (0.98)	Bris08	4FQK (0.99)



A) Influenza A/H1N1 B) Influenza A/H3N2 C) Influenza B

Figure 6.4: Binding affinity of host cell receptors with the HA1 of the representative (A) influenza A/H1N1, (B) influenza A/H3N2 and (C) influenza B viruses.

avian receptor analog 3'SLN and the human receptor analog 6'SLN was docked to the receptor binding domain of HA structures respectively. Each docking experiment was repeated independently for 1000 times. Figure 6.4 presents the binding affinities for each group of experiments in box plots. Experiments of HA1-3'SLN binding are shown in lighter color, while the HA1-6'SLN bindings are shown in darker color. Besides, the student T-test was conducted to compare the group mean difference. Results of the student's T-test at confidence level 99.9% are shown in Table 6.6.

For influenza A/H1N1, the HA1 protein acquired enhanced ability to bind both the avian and human receptor analogs from 1918 to 2014, especially the binding with human receptor analogs. From the results, it seems that the G2-25.1 strain from 2014 has a binding preference for the human receptor analog while the SGH-C strain from 2012 has a binding preference for the avian receptor analog. To check the binding preferences of each isolate, we performed student T-test to analyze the significance of group mean difference. All representative strains had significant difference binding with 3'SLN and 6'SLN, with p-value less than 0.001. The G2-25.1 strain had binding preference with 6'SLN while other strains (SGH-C, CA09 and SC18) had binding preference with 3'SLN.

For influenza A/H3N2, strains from 2009 to 2014 did not show obvious enhanced binding ability with either the avian or human receptor analog. HA proteins of the influenza A/H3N2 representatives binding with 6'SLN showed larger variance than binding with 3'SLN. The binding with human receptor analog (6'SLN) was even slightly weakened from 2011 onwards. The harmfulness of A/H3N2 might be mainly attributed to the mutations and mis-matching with the vaccine strains. The inpatient strain CDC-73 showed slightly stronger binding ability with both avian and human receptor analogs than the G2-26.1 and the CDC-204 strain. Results of the student-T test indicated that all the representative strains of influenza A/H3N2 had a binding preference with the human receptor analog (3'SLN) at 99.9% confidence level. It is worth noting that all strains showed larger variance in docking experiments when binding with 6'SLN than 3'SLN.

All the representative strains of flu B had preference binding with the human receptor analog (6'SLN), among which all the Victoria lineage isolates showed stronger binding ability with both avian and human receptor analogs than the isolates from the Yamagata lineage. Rep-

Table 6.6: The group mean differences of the HA1 proteins of the representative influenza A/H1N1, A/H3N2 and B binding with 3'SLN and 6'SLN.

Subtype		Strain	3SLN		6SLN		P(T<=t) two-tail
			Mean	Variance	Mean	Variance	
A/H1N1		SC18	-6.4137	0.0379	-6.2607	0.1388	1.94E-29
		CA09	-7.0108	0.0129	-6.59	0.0660	3.50E-291
		SGH-C	-6.9731	0.0148	-6.7354	0.0229	1.10E-242
		G2-25.1	-7.0176	0.0131	-7.1316	0.0441	5.41E-48
A/H3N2		P0	-6.4360	0.0083	-6.4908	0.0667	3.50E-10
		P14	-6.5386	0.0334	-6.7433	0.0541	7.37E-95
		Per09	-6.7553	0.0144	-7.0885	0.0267	0
		Vic11	-6.5876	0.0056	-6.9068	0.1184	3.40E-135
		CDC-73	-6.8135	0.0165	-7.0311	0.0437	5.50E-142
		CDC-204	-6.7787	0.0150	-6.8567	0.0426	4.57E-24
		G2-26.1	-6.7665	0.0190	-6.8401	0.0423	1.66E-20
B	Vic	Bris08	-6.5068	0.0178	-6.9225	0.0206	0
		G2-14.1	-6.5093	0.0180	-6.9396	0.0256	0
		Col17	-6.3634	0.0186	-6.9533	0.0201	0
	Yam	Wis10	-6.0016	0.0292	-6.2213	0.0226	7.80E-168
		TTSH-69	-6.1004	0.0295	-6.2469	0.0504	1.76E-56
		Phu13	-5.9653	0.0274	-6.2367	0.0216	5.60E-245
		G2-36.1	-6.0041	0.0295	-6.2351	0.0220	1.00E-182

resentatives within each lineage maintained the binding ability with both receptors without significant increasing or decreasing.

6.4.2 A/H1N1 receptor binding modes

The comparison of molecular docking suggested that most of the sampled Singapore influenza strains had binding preference for the human receptor analog (6'SLN) except for the influenza A/H1N1 inpatient isolate SGH-C collected in 2012. For SGH-C, the average binding energy of 1000 ligand poses was -6.7354 kcal/mol for 6'SLN compared to -6.9731 kcal/mol for 3'SLN (Table 6.6). Therefore, the HA-3'SLN/6'SLN complexes of influenza A/H1N1 generated from molecular docking studies were further analyzed to identify the structural characteristics of the observed HA-receptor bindings, aiming to investigate possible reasons for the higher binding affinity of SGH-C with the avian receptor.

First, the ligand pose with the lowest energy conformation predicted by QuickDock2 (Alhossary *et al.*, 2015) was selected for each H1-receptor complex. All HA structures were aligned to C-α atoms. The putative contacts and solvent accessible surface area (ASA) of docked complexes were investigated afterwards for the H1N1 hemagglutinin (HA1 domain) of viral isolates G2-25.1 and SGH-C and reference strains A/California/04/2009 (CA09; PDB ID 3UBQ and 3UBN (Xu *et al.*, 2012b)) and A/Brevig Mission/1/1918 (BM18; PDB ID 4JUH and 4JUJ (Zhang *et al.*, 2013)).

Putative hydrogen bonds and the solvent accessible surface areas (ASA) of ligand (L), target

Table 6.7: Solvent Accessible Surface Area of HA-receptor complexes.

Strain (Target)	Component	Receptor (Ligand)	
		3'SLN	6'SLN
G2-25.1	Target (T)	16860.8	16860.8
	Ligand (L)	863.58	842.08
	Complex (C)	16762.2	16742.3
	L+T – C	962.23	960.57
Buried (%)	(L+T – C)/C*100%	5.74	5.74
SGH-C	Target (T)	16881.2	16881.2
	Ligand (L)	847.06	866.26
	Complex (C)	16767.4	16791.9
	L+T – C	960.89	955.65
Buried (%)	(L+T – C)/C*100%	5.73	5.69
CA09	Target (T)	16685.7	16568
	Ligand (L)	906.33	807.32
	Complex (C)	16893.2	16568.9
	L+T – C	698.89	806.39
Buried (%)	(L+T – C)/C*100%	4.14	4.87
BM18	Target (T)	16912.9	16908.1
	Ligand (L)	879.44	828.54
	Complex (C)	17182.7	16938.3
	L+T – C	609.59	798.36
Buried (%)	(L+T – C)/C*100%	3.55	4.71

(T) and docked complex (C) were also calculated using PyMOL version 1.8.4 (Schrodinger, 2015). The buried surface area (BSA) of each complex was estimated as Equation (6.1).

$$BSA = ASA_L + ASA_T - ASA_C \tag{6.1}$$

Default parameters were used for estimation. For example, the distance between the donor and the acceptance atoms was 3.6Å, the ASA probe radius was 1.4Å. Table 6.7 compares the ASA for each component and the percentage of buried surfaces, while Table 6.8 lists the number of putative hydrogen bonds between the HA protein and host cell receptors.

The HA of G2-25.1 has an identical fraction of buried surface area as of 5.74% when binding with 3'SLN and 6'SLN. For SGH-C, the HA has a larger fraction of buried surface area when binding with 3'SLN (5.73%) than with 6'SLN(5.69%). For both reference strains CA09 and BM18, the HA experimentally determined complexes have larger buried surface area with the human receptor analog than with the avian receptor analog. Besides, the buried surface areas of current Singapore strains (G2-25.1 and SGH-C) with both types of host cell receptors are larger than the reference strains (CA09 and BM18).

Although it is a well-established knowledge that influenza virus recognizes host cells through binding with sialylated glycan receptors, there were observations that the influenza viruses were also able to infect cells with sialic acid receptors been removed (Stray *et al.*, 2000; Nicholls *et al.*, 2008). Therefore, the contacts between each receptor moiety and the HA protein were

Table 6.8: Number of putative hydrogen bonds between the HA protein and host cell receptors.

Strain	Ligand residue	Receptor	
		3'SLN	6'SLN
G2-25.1	SIA	5	7
	GAL	3	1
	NAG	2	3
	Total	10	11
	GAL(%)	30	9.1
SGH-C	SIA	6	6
	GAL	3	0
	NAG	2	4
	Total	11	10
	GAL(%)	27.3	0
CA09	SIA	5	6
	GAL	0	1
	NAG	0	0
	Total	5	7
	GAL(%)	0	14.3
BM18	SIA	3	6
	GAL	0	1
	NAG	0	1
	Total	3	8
	GAL(%)	0	12.5

investigated. In Table 6.8, the number of hydrogen bonds formed with each of the three receptor moieties are decomposed. As seen, the sialic acid (SIA) moiety is heavily utilized in all four complexes, forming over 50% hydrogen bonds. All strains form the most number of hydrogen bonds with SIA than GAL and NAG when binding with both the avian and human receptor analogs. All strains form more hydrogen bonds with SIA in the complex with 6'SLN, except for SGH-C which form the same number of hydrogen bonds with SIA in the 3'SLN and 6'SLN. Besides, the contributions of GAL in forming hydrogen bonds have been investigated. The GAL in 3'SLN forms a high fraction of hydrogen bonds in the complexes of both Singapore strains G2-25.1 (30%) and SGH-C (27.3%). In contrast, the complexes of the two reference pandemic strains have none hydrogen bonds between the GAL in 3'SLN. The distinction of GAL's contribution in 3'SLN and 6'SLN in forming HA bonds with the HA protein provides a potential new binding mode for viral recognition with host cell receptors. Also, the preference of SGH-C for 3'SLN may partially be explained by the observation that its HA forms no hydrogen bonds with the GAL in 6'SLN, while forming 3 hydrogen bonds with the GAL in 3'SLN.

Figure 6.5 compares the HAs binding with the avian receptor among the Singapore representative strains (SGH-C and G2-25.1) and the pandemic representatives (CA09 and BM18), while Figure 6.6 compares the HAs binding with the human receptor. For SGH-C and G2-25.1, the predicted structures of the complexes with the lowest binding energy are presented in figures. For CA09 and BM18, the high quality complex structures with the avian and human

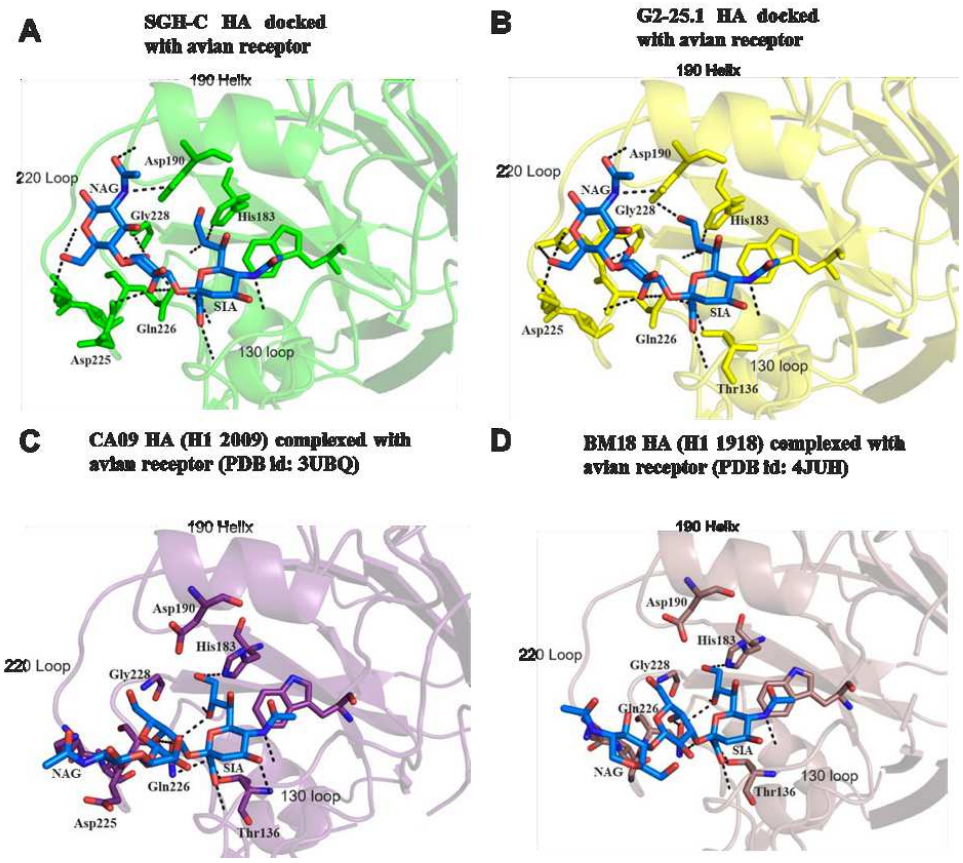


Figure 6.5: H1 HA with avian receptor analogs. *(A) HA of SGH-C docked with 3'SLN. (B) HA of G2-25.1 docked with 3'SLN. (C) HA of CA09 complexed with 3'SLN (PDB ID: 3UBQ). (D) HA of BM18 complexed with 3'SLN (PDB ID: 4JUH).*

receptors are available in PDB, with a resolution of 2-3 Å. The HA proteins are shown with cartoon representatives, while the receptors are represented as sticks. The receptor binding domains are mainly composed of the 130-loop, the 190-helix and the 220-loop as indicated. Putative hydrogen bonded contacts, which were identified by the PyMOL polar contacts, are indicated as black dashed lines.

As seen in Figure 6.5, the predicted complexes of Singapore strains (G2-25.1 and SGH-C) contain more hydrogen bonds and the avian receptor binds deeper into the binding groove than the pandemic representatives (CA09 and BM18). Both the SIA (the innermost ring) and NAG (the outermost ring) form potential hydrogen bonds with the residue Asp190 (D190), which is absent in the HA of CA19 and BM18. The GAL in 3'SLN forms potential hydrogen bonds with key residues Asp225 (D225), Gln226 (Q226), Gly228 (G228). However, in the HA of BM18 and CA09, GAL only forms few hydrogen bonds with Q226; The observations are consistent with the hydrogen bond analyses presented in Table 6.8.

Figure 6.6 illustrates binding with human receptor analog. The docked receptors to Singapore strains are colored pink (Figure 6.6A and B) while those in previous pandemic strains CA08 and BM18 are colored green (Figure 6.6C). Typically, the avian receptor binds in an extended conformation, with SIA facing inwards to the binding groove (Figure 6.5). In com-

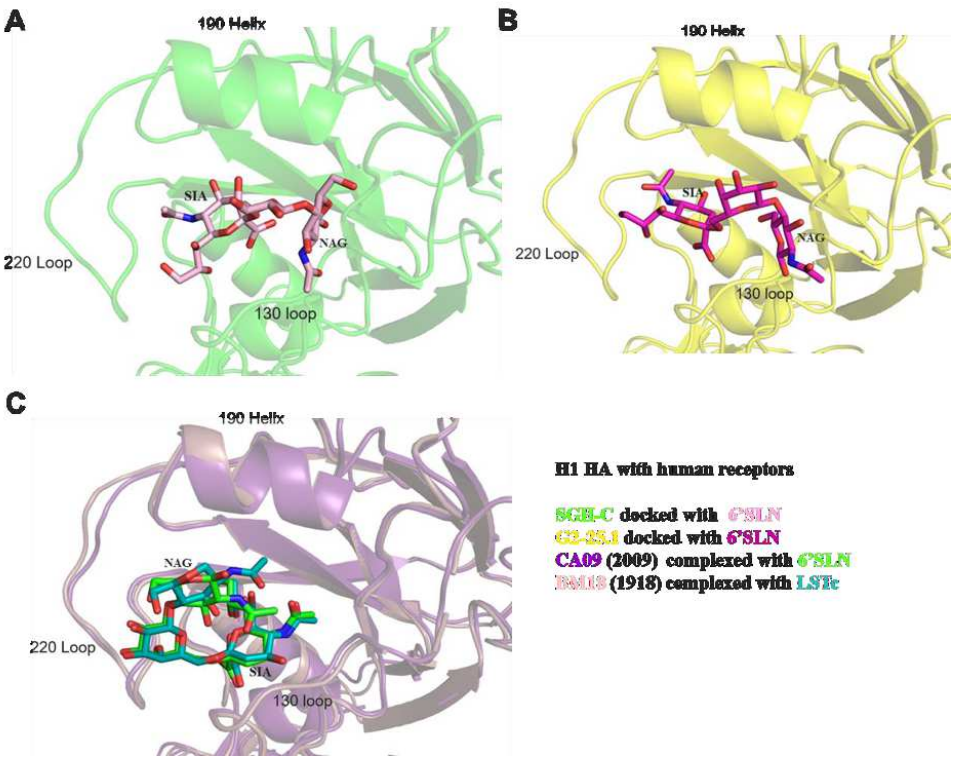


Figure 6.6: H1 HA with human receptor analogs. *(A) HA of SGH-C docked with 6'SLN. (B) HA1 of G2-25.1 docked with 6'SLN. (C) Superimposed HA1 of CA09 and BM18 complexed with 6'SLN and LSTc respectively. CA09 is shown in dark purple and BM18 in light pink (PDB ID: 3UBN and 4JUJ respectively).*

parison, 6'SLN binds in the cis conformation (Figure 6.6C). A novel ligand pose has been observed in G2-25.1 and SGH-C. For SGH-C (Figure 6.6A) and G2-25.1 (Figure 6.6B), the SIA faces towards the binding groove near the 220-loop instead of the 130-loop (Figure 6.6C). Besides, the NAG faces downwards (towards the 130-loop) instead of upwards (towards the 190-helix) in the HA of CA09 and BM18 (Figure 6.6C).

The buried surface area upon complex formation is used here as a measure of the goodness of fit of the receptors (ligands) for respective targets (HA1 domain of viral strains). It is reasoned that, a receptor that has high affinity (low binding energy), makes favorable contacts with the receptor binding site (RBS) in the HA1 globular head domain. The receptor binding site consists of the 190 helix, 220 Loop and 130 Helix that form a binding groove (Figure 6.5 and Figure 6.6). For ligands with low binding energy, it is generally expected more of the ASA is buried after complex formation compared to ligands that fit poorly into the binding groove. Also, the low energy ligand conformation may form ideal geometries and donor acceptor distance for hydrogen bonding. Hence, it overall fits better into the binding groove. Upon complex formation, generally, the ligand with more favourable interactions will have a higher percentage of buried surface area as a fraction of the total complex surface area, compared to a receptor with a poorer fit or one that the HA protein has lesser affinity.

6.5 Summary

Influenza A/H1N1, A/H3N2 and B viruses that caused mild infections among outpatients and severe infections among inpatients in Singapore during 2012-2015 were isolated and characterized. The evolution, mutations and receptor binding specificity with host cells were investigated systematically.

As a result, there was a strong impression that the local Singapore strains were a main part of global virus circulation. The BLAST output suggested that reassortment events among local and global strains within each of influenza A/H1N1, A/H3N2 and B might not be uncommon. Particularly, one viral genome containing gene segments originated from both influenza A/H1N1 and A/H3N2 was documented. Furthermore, the phylogenetic tree of each gene segment showed that the outpatient and inpatient cases were generally overlapped with the global and vaccine representatives, except for a cluster of inpatients infected with influenza A/H3N2 viruses that were closely related to vaccine strain A/Texas/50/2012(H3N2). Several protein sites were found critical in discriminating the inpatient and outpatient cases. For example, the site 695 in PB1 and site 221 in NA achieves the highest accuracy for influenza A/H1N1 and A/H3N2, respectively. Some of these sites have known functional importance, while some are novel reported. Finally, an enhancement of HA binding with both the avian and human host cell receptors was observed, suggesting viral evolution towards dual receptor binding specificity. Besides, the binding ability to each receptor was relatively stable for the hemagglutinin of influenza B.

Overall, the research constructs a pipline for profiling important characteristics of influenza viruses, including the evolution, mutations and receptor binding specificity. The computational analyses can help facilitate the genomic surveillance on the circulating influenza viruses, while the findings extend knowledge about the circulating Singapore strains. Further functional assays and crystallized protein-ligand structures will be useful to corroborate these insights.

Chapter 7

Discussion and conclusion

7.1 General discussion

The goal of this dissertation is to develop a computational model to predict the virulence of new emerging influenza viral strain. Predicting antigenic variants and detecting virulence-related sites are important aspects towards the goal.

In Chapter 3, a phylogenetic tree based method was proposed for pairwise co-mutations detection. The method was applied to the HA protein of influenza A/H3N2, and successfully identified dominant mutations responsible for the antigenic evolution of HA. Furthermore, an association rule based method was proposed for co-mutations at multiple sites. The method was applied to the HA protein of influenza A/H1N1, A/H3N2 and B viruses, finding that the co-mutations on HA protein can characterize the evolution of influenza viruses. It is worth noting that the mutation drift of a viral protein is remarkably sensitive to the genomic context and may be affected by the other proteins. Therefore, the work was improved to screen a concise subset of co-occurring and sequential mutations on all proteins, and across different proteins of influenza viruses.

In Chapter 4, an encoding scheme CFreeEnS was proposed for universe subtype of influenza viruses. By systematically checking all the available substitution matrices, which consider different properties of amino acids, it was found that properties related to the structures of amino acids or contacts between amino acids can help improve the prediction in the combined dataset.

In Chapter 5, preliminary structural analysis has been applied to explore the impact of HA mutations on the host receptor binding preference. Such simulation-generated information should be integrated to provide a comprehensive description of viral capability of host infection and drug resistance. A new emerging influenza A/H7N9 strain was taken as the subject, both sequence and structural analyses were applied to assess its potential in circulating among humans. The sequence analysis highlighted mutations in protein functional domains of influenza viruses. Molecular docking and molecular dynamics simulation revealed that the HA enhanced the binding with both avian and human receptor analogs. Besides, the decomposition

of residue contributions revealed several residues being responsible for the change of receptor binding preference. The obtained results are specific to the mentioned strain, shedding light on the impacts of HA mutations and the mechanisms of receptor recognition. What is even more important is that the pipeline of analysis is applicable to analyzing the impacts of other mutations on the binding of proteins with small ligands, helping to quantify the receptor binding specificity of influenza viruses.

Chapter 6 presented an example of profiling the evolution, mutations and receptor binding specificity of influenza viruses. The influenza A/H1N1, A/H3N2 and B viruses causing mild infections among outpatients and severe infections among inpatients during 2012-2015 in Singapore were sequenced and computationally investigated, using both sequence analyses and structural analyses. The research constructed a pipline for profiling important characteristics of influenza viruses, including the evolution, mutations and receptor binding specificity. The computational analyses can help facilitate the genomic surveillance on the circulating influenza viruses, while the findings extend knowledge about the circulating Singapore strains. Further functional assays and crystallized protein-ligand structures will be useful to corroborate these insights.

7.2 Future directions

The contagious influenza virus infection has been circulating in various species around the world, causing high burden to the society. This dissertation paid much attention to the viral mutations, using both sequence and structural computational models to detect and predict viral mutations, especially mutations leading to antigenic variations and increased virulence. Functional assays and crystallized protein-ligand structural analyses could be one direction for the community to consolidate the predicted characteristics of influenza viruses from computational models.

Also, it should be interesting to integrate the sequence analyses and structural analyses on the antigenicity and receptor binding specificity of influenza, serving as a preliminary assessment tool for characterizing the virulence of influenza viruses. Besides, similar to analyzing the receptor binding specificity using molecular docking and MD simulation, it is promising to take the drug resistance of influenza viruses into account.

To date, influenza vaccines are the most effective ways to prevent the infection from influenza viruses. However, the effectiveness of vaccines varies, which is heavily dependent on the match of antigenicity between the circulating strains and the vaccine components. The antigenic drift of influenza raised the challenges of updating influenza vaccines, especially when the strains are difficult to isolate or characterized through HAI assays.

Current influenza vaccines are strain-specific, including only three or four strains dominating the circulation among human populations. Researchers are endeavoring to discover universal influenza vaccines, which could provide long-lasting protection against multi-subtype of influenza viruses for pan-group. Instead of focusing on the protein regions under high evo-

lutionary pressure (e.g. the globular head of HA protein), the internal viral proteins, M2e, NA, stalk domains and conserved regions of HA are potential targets for a universal influenza vaccine (Krammer and Grabherr, 2010).

No universal influenza vaccine is approved for public use, but there are some achievements made in clinical trail. The National Institute of Allergy and Infectious Disease (NIAID) released a universal influenza vaccine trail named "M-001", which contains antigenic sequences shared by several different influenza viral strains. In May 2019, BiondVax announced the pivotal Phase 3 clinical trail for assessing the safely and effectiveness "M-001" against multi-strain and multi-subtype of influenza (Gabisonia *et al.*, 2019).

Integrating artificial intelligence (AI) into the system for characterizing the antigenicity and drug resistance of influenza strains is one way to facilitate the selection of influenza vaccine candidates. Using AI to design flu vaccines is an alternative shot. The development of new drugs takes countless time and cost before reaching human trails. AI can facilitate the process for screening drugs and designing vaccines (Davies, 2019; Mak and Pichika, 2018). In July 2019, researchers claimed the success of AI in developing a promising new influenza vaccine, which is also the first AI-developed vaccine of the world to begin a 12-month clinical trail. This is a milestone for both AI and vaccination, which also points out one direction for future work.

Bibliography

Abed, Y. *et al.* (2011). Role of permissive neuraminidase mutations in influenza a/brisbane/59/2007-like (h1n1) viruses. *PLoS pathogens*, **7**(12), e1002431.

Abed, Y. *et al.* (2014). Impact of potential permissive neuraminidase mutations on viral fitness of the h275y oseltamivir-resistant influenza a (h1n1) pdm09 virus in vitro, in mice and in ferrets. *Journal of virology*, **88**(3), 1652–1658.

Abraham, M. J. *et al.* (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1**, 19–25.

Air, G. M. (2012). Influenza neuraminidase. *Influenza and other respiratory viruses*, **6**(4), 245–256.

Akand, E. H. and Downard, K. M. (2018). Identification of epistatic mutations and insights into the evolution of the influenza virus using a mass-based protein phylogenetic approach. *Molecular phylogenetics and evolution*, **121**, 132–138.

Alhossary, A. *et al.* (2015). Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics*, **31**(13), 2214–2216.

Allen, M. P. *et al.* (2004). Introduction to molecular dynamics simulation. *Computational soft matter: from synthetic polymers to proteins*, **23**, 1–28.

Altschul, S. F. *et al.* (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, **25**(17), 3389–3402.

Amano, Y. and Cheng, Q. (2005). Detection of influenza virus: traditional approaches and development of biosensors. *Analytical and bioanalytical chemistry*, **381**(1), 156–164.

Ang, L. W. *et al.* (2014). Influenza-associated hospitalizations, singapore, 2004-2008 and 2010-2012. *Emerg Infect Dis*, **20**(10), 1652–60.

Ang, L. W. *et al.* (2016). Characterization of influenza activity based on virological surveillance of influenza-like illness in tropical singapore, 2010-2014. *J Med Virol*, **88**(12), 2069–2077.

Archetti, I. and Horsfall, F. L. (1950). Persistent antigenic variation of influenza a viruses after incomplete neutralization in ovo with heterologous immune serum. *Journal of Experimental Medicine*, **92**(5), 441–462.

Arinaminpathy, N. and Grenfell, B. (2011). Dynamics of glycoprotein charge in the evolutionary history of human influenza. *PLOS ONE*, **5**, 1–7.

Arnold, K. *et al.* (2006). The swiss-model workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, **22**(2), 195–201.

Baker, F. N. and Porollo, A. (2016). Coeviz: a web-based tool for coevolution analysis of protein residues. *BMC bioinformatics*, **17**(1), 119.

Bao, Y. *et al.* (2007). Flan: a web server for influenza virus genome annotation. *Nucleic acids research*, **35**(suppl 2), W280–W284.

Bao, Y. *et al.* (2008). The influenza virus resource at the national center for biotechnology information. *Journal of virology*, **82**(2), 596–601.

Barr, I. G. *et al.* (2010). Epidemiological, antigenic and genetic characteristics of seasonal influenza a (h1n1), a (h3n2) and b influenza viruses: basis for the who recommendation on the composition of influenza vaccines for use in the 2009–2010 northern hemisphere season. *Vaccine*, **28**(5), 1156–1167.

Barrero, P. R. *et al.* (2011). Genetic and Phylogenetic Analyses of Influenza A H1n1pdm Virus in Buenos Aires, Argentina. *Journal of Virology*, **85**(2), 1058–1066.

Baum, L. G. and Paulson, J. C. (1991). The n2 neuraminidase of human influenza virus has acquired a substrate specificity complementary to the hemagglutinin receptor specificity. *Virology*, **180**(1), 10–5.

Baz, M. *et al.* (2010). Effect of the Neuraminidase Mutation H274Y Conferring Resistance to Oseltamivir on the Replicative Capacity and Virulence of Old and Recent Human Influenza A(H1N1) Viruses. *The Journal of Infectious Diseases*, **201**(5), 740–745.

Bedford, T. *et al.* (2011). Strength and tempo of selection revealed in viral gene genealogies. *BMC Evolutionary Biology*, **11**(1), 220.

Bedford, T. *et al.* (2014). Integrating influenza antigenic dynamics with molecular evolution. *eLife*.

Belongia, E. A. *et al.* (2016). Variable influenza vaccine effectiveness by subtype: a systematic review and meta-analysis of test-negative design studies. *The Lancet Infectious Diseases*, **16**(8), 942–951.

Benkert, P. *et al.* (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, **27**(3), 343–350.

Berendsen, H. J. *et al.* (1984). Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, **81**(8), 3684–3690.

Berman, H. M. *et al.* (2000). The protein data bank. *Nucleic acids research*, **28**(1), 235–242.

Bhadra, P. *et al.* (2018). Ampep: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Scientific reports*, **8**(1), 1697.

Bhatt, S. *et al.* (2011). The genomic rate of molecular adaptation of the human influenza a virus. *Molecular biology and evolution*, **28**(9), 2443–2451.

Biasini, M. *et al.* (2013). Openstructure: an integrated software framework for computational structural biology. *Acta Crystallogr D Biol Crystallogr*, **69**(Pt 5), 701–9.

Biasini, M. *et al.* (2014). Swiss-model: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic acids research*, page gku340.

Biere, B. *et al.* (2010). Differentiation of influenza b virus lineages yamagata and victoria by real-time pcr. *Journal of Clinical Microbiology*, **48**(4), 1425–1427.

Bloom, J. D. *et al.* (2010). Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science*, **328**(5983), 1272–1275.

Bodewes, R. *et al.* (2013). Recurring influenza b virus infections in seals. *Emerging infectious diseases*, **19**(3), 511.

Bornholdt, Z. A. and Prasad, B. V. V. (2006). X-ray structure of influenza virus NS1 effector domain. *Nature Structural &Amp; Molecular Biology*, **13**, 559.

Bouckaert, R. *et al.* (2014). Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, **10**(4), e1003537.

Bouvier, N. M. and Lowen, A. C. (2010). Animal models for influenza virus pathogenesis and transmission. *Viruses*, **2**(8), 1530–1563.

Bouvier, N. M. and Palese, P. (2008). The biology of influenza viruses. *Vaccine*, **26**, D49–D53.

Bragstad, K. *et al.* (2008). The evolution of human influenza a viruses from 1999 to 2006: a complete genome study. *Virology journal*, **5**(1), 40.

Brownlee, G. and Fodor, E. (2001). The predicted antigenicity of the haemagglutinin of the 1918 spanish influenza pandemic suggests an avian origin. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **356**(1416), 1871–1876.

Burke, D. F. and Smith, D. J. (2014). A recommended numbering scheme for influenza a ha subtypes. *PloS one*, **9**(11), e112302.

Burley, S. K. *et al.* (2018). Rcsb protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic acids research*, **47**(D1), D464–D474.

Bush, R. M. *et al.* (1999). Predicting the evolution of human influenza a. *Science*, **286**(5446), 1921–1925.

Cai, Z. *et al.* (2010). A computational framework for influenza antigenic cartography. *PLoS Computational Biology*, **6**(10), e1000949.

Cai, Z. *et al.* (2012). Antigenic distance measurements for seasonal influenza vaccine selection. *Vaccine*, **30**(2), 448–453.

Caton, A. J. *et al.* (1982). The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell*, **31**(2), 417–427.

CDC, C. f. D. C. *et al.* (2015). Selecting viruses for the seasonal influenza vaccine. *Centers for Disease Control and Prevention, Atlanta, GA*.

CDC, C. f. D. C. *et al.* (2018). Vaccine effectiveness - how well does the flu vaccine work?

Centers for Disease Control and Prevention (CDC) (2018). Influenza antiviral medications: Summary for clinicians.

Chaudhary, K. *et al.* (2016). A web server and mobile app for computing hemolytic potency of peptides. *Scientific reports*, **6**, 22843.

Chen, H. *et al.* (2016). Rules of co-occurring mutations characterize the antigenic evolution of human influenza a/h3n2, a/h1n1 and b viruses. *BMC medical genomics*, **9**(3), 69.

Chen, W. *et al.* (2001). A novel influenza a virus mitochondrial protein that induces cell death. *Nat Med*, **7**(12), 1306–12.

Chen, W. *et al.* (2012). The evolutionary pattern of glycosylation sites in influenza virus (h5n1) hemagglutinin and neuraminidase. *PLOS ONE*, **7**(11), 1–11.

Cheng, Y.-Y. *et al.* (2017). Amino acid residues 68–71 contribute to influenza a virus pb1-f2 protein stability and functions. *Frontiers in microbiology*, **8**, 692.

Chi, X. S. *et al.* (2005). Detection and characterization of new influenza b virus variants in 2002. *J Clin Microbiol*, **43**(5), 2345–9.

Chien, C.-y. *et al.* (1997). A novel RNA-binding motif in influenza A virus non-structural protein 1. *Nature Structural Biology*, **4**(11), 891–895.

Choi, W.-S. *et al.* (2017). Rapid acquisition of polymorphic virulence markers during adaptation of highly pathogenic avian influenza h5n8 virus in the mouse. *Scientific Reports*, **7**.

Chong, Y. and Ikematsu, H. (2018). Spread of predominant neuraminidase and hemagglutinin co-mutations in the influenza a/h3n2 virus genome. *Journal of infection and chemotherapy*, **24**(3), 193–198.

Chou, Y.-y. *et al.* (2011). The m segment of the 2009 new pandemic h1n1 influenza virus is critical for its high transmission efficiency in the guinea pig model. *Journal of virology*, **85**(21), 11235–11241.

Coburn, B. J. *et al.* (2009). Modeling influenza epidemics and pandemics: insights into the future of swine flu (h1n1). *BMC medicine*, **7**(1), 30.

Codoñer, F. M. and Fares, M. A. (2008). Why should we care about molecular coevolution? *Evolutionary Bioinformatics*, **4**, 117693430800400003.

Conenello, G. M. *et al.* (2007). A single mutation in the pb1-f2 of h5n1 (hk/97) and 1918 influenza a viruses contributes to increased virulence. *PLoS Pathog*, **3**(10), 1414–21.

Consortium, U. *et al.* (2018). Uniprot: the universal protein knowledgebase. *Nucleic acids research*, **46**(5), 2699.

Cox, N. J. *et al.* (2014). Pandemic preparedness and the influenza risk assessment tool (irat). In *Influenza Pathogenesis and Control-Volume I*, pages 119–136. Springer.

Cutter, J. L. *et al.* (2010). Outbreak of pandemic influenza a (h1n1-2009) in singapore, may to september 2009. *Ann Acad Med Singapore*, **39**(4), 273–10.

Daniels, P. *et al.* (1987). The receptor-binding and membrane-fusion properties of influenza virus variants selected using anti-haemagglutinin monoclonal antibodies. *The EMBO journal*, **6**(5), 1459.

Daniels, R. S. *et al.* (1984). Antigenic analyses of influenza virus haemagglutinins with different receptor-binding specificities. *Virology*, **138**(1), 174–7.

Davies, S. E. (2019). Artificial intelligence in global health. *Ethics & International Affairs*, **33**(2), 181–192.

Dawood, F. S. *et al.* (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza a h1n1 virus circulation: a modelling study. *The Lancet infectious diseases*, **12**(9), 687–695.

de Jong, J. C. *et al.* (2000). Mismatch between the 1997/1998 influenza vaccine and the major epidemic a(h3n2) virus strain as the cause of an inadequate vaccine-induced antibody response to this strain in the elderly. *Journal of Medical Virology*, **61**(1), 94–99.

Deng, Y. M. *et al.* (2015). A simplified sanger sequencing method for full genome sequencing of multiple subtypes of human influenza a viruses. *J Clin Virol*, **68**, 43–8.

Desmet, E. A. *et al.* (2013). Identification of the n-terminal domain of the influenza virus pa responsible for the suppression of host protein synthesis. *Journal of Virology*, **87**(6), 3108–3118.

Dias, A. *et al.* (2009). The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature*, **458**, 914.

Ding, H. *et al.* (2014). Identification of bacteriophage virion proteins by the anova feature selection and analysis. *Molecular BioSystems*, **10**(8), 2229–2235.

Doraisingham, S. *et al.* (1988). Influenza surveillance in singapore: 1972-86. *Bull World Health Organ*, **66**(1), 57–63.

Drummond, A. J. and Rambaut, A. (2007). Beast: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, **7**(1), 214.

Du, X. *et al.* (2008). Networks of genomic co-occurrence capture characteristics of human influenza a (h3n2) evolution. *Genome research*, **18**(1), 178–187.

Du, X. *et al.* (2012). Mapping of h3n2 influenza antigenic evolution in china reveals a strategy for vaccine strain recommendation. *Nature communications*, **3**, 709.

Dubois, J. *et al.* (2014). Influenza viruses and mrna splicing: doing more with less. *MBio*, **5**(3), e00070–14.

Dudek, S. E. *et al.* (2011). The influenza virus pb1-f2 protein has interferon antagonistic activity. *Biological chemistry*, **392**(12), 1135–1144.

Duncan, I. G. *et al.* (2012). Planning influenza vaccination programs: a cost benefit model. *Cost Eff Resour Alloc*, **10**(1), 10.

Duprex, W. P. *et al.* (2015). Gain-of-function experiments: time for a real debate. *Nature Reviews Microbiology*, **13**(1), 58.

Durrant, M. G. *et al.* (2015). Investigation of a recent rise of dual amantadine-resistance mutations in the influenza a m2 sequence. *BMC genetics*, **16**(2), S3.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**(5), 1792–1797.

Edinger, T. O. *et al.* (2014). Entry of influenza a virus: host factors and antiviral targets. *Journal of General Virology*, **95**(2), 263–277.

Egorov, A. *et al.* (1998). Transfectant Influenza A Viruses with Long Deletions in the NS1 Protein Grow Efficiently in Vero Cells. *Journal of Virology*, **72**(8), 6437–6441.

Eisfeld, A. J. *et al.* (2014). Influenza a virus isolation, culture and identification. *Nature Protocols*, **9**, 2663–2681.

Eng, C. L. *et al.* (2014). Predicting host tropism of influenza a virus proteins using random forest. *BMC Medical Genomics*, **7**(3), S1.

Engelhardt, O. G. and Fodor, E. (2006). Functional association between viral and cellular transcription during influenza virus infection. *Reviews in Medical Virology*, **16**(5), 329–345.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, **17**(6), 368–376.

Ferguson, L. *et al.* (2016). Pathogenesis of influenza d virus in cattle. *Journal of Virology*, **90**(12), 5636–5642.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, **20**(4), 406–416.

Fournier-Viger, P. and Tseng, V. S. (2013). Tns: mining top-k non-redundant sequential rules. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 164–166. ACM.

Fournier-Viger, P. *et al.* (2008). A knowledge discovery framework for learning task models from user interactions in intelligent tutoring systems. In *Mexican International Conference on Artificial Intelligence*, pages 765–778. Springer.

Fournier-Viger, P. *et al.* (2012). Mining top-k association rules. In *Canadian Conference on Artificial Intelligence*, pages 61–73. Springer.

Frickey, T. and Lupas, A. (2004). Clans: a java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**(18), 3702–3704.

Gabisonia, K. *et al.* (2019). Archive for may, 2019.

Gabriel, G. *et al.* (2008). Interaction of Polymerase Subunit PB2 and NP with Importin α_1 Is a Determinant of Host Range of Influenza A Virus. *PLOS Pathogens*, **4**(2), 1–10.

Gambaryan, A. S. *et al.* (1998). Effects of Host-Dependent Glycosylation of Hemagglutinin on Receptor-Binding Properties of H1n1 Human Influenza A Virus Grown in MDCK Cells and in Embryonated Eggs. *Virology*, **247**(2), 170 – 177.

Gamblin, S. J. and Skehel, J. J. (2010). Influenza hemagglutinin and neuraminidase membrane glycoproteins. *Journal of Biological Chemistry*, **285**(37), 28403–28409.

Gamblin, S. J. *et al.* (2004). The Structure and Receptor Binding Properties of the 1918 Influenza Hemagglutinin. *Science*, **303**(5665), 1838–1842.

Gao, H. *et al.* (2011). New methods to measure residues coevolution in proteins. *BMC bioinformatics*, **12**(1), 206.

García-Sastre, A. (2001). Inhibition of Interferon-Mediated Antiviral Responses by Influenza A Viruses and Other Negative-Strand RNA Viruses. *Virology*, **279**(2), 375 – 384.

Garten, R. J. *et al.* (2009). Antigenic and genetic characteristics of swine-origin 2009 a (h1n1) influenza viruses circulating in humans. *science*, **325**(5937), 197–201.

Gerdil, C. (2003). The annual production cycle for influenza vaccine. *Vaccine*, **21**(16), 1776–1779.

Ghedin, E. *et al.* (2005). Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, **437**(7062), 1162–1166.

Ginting, T. E. *et al.* (2012). Amino acid changes in hemagglutinin contribute to the replication of oseltamivir-resistant h1n1 influenza viruses. *Journal of virology*, **86**(1), 121–127.

Girard, M. P. *et al.* (2010). The 2009 a (h1n1) influenza virus pandemic: A review. *Vaccine*, **28**(31), 4895–4902.

Glaser, L. *et al.* (2005). A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. *Journal of virology*, **79**(17), 11533–11536.

Gong, L. I. *et al.* (2013). Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife*, **2**, e00631.

Grabenstein, J. D. *et al.* (2006). Immunization to Protect the US Armed Forces: Heritage, Current Practice, and Prospects. *Epidemiologic Reviews*, **28**(1), 3–26.

Graef, K. M. *et al.* (2010). The pb2 subunit of the influenza virus rna polymerase affects virulence by interacting with the mitochondrial antiviral signaling protein and inhibiting expression of beta interferon. *Journal of virology*, **84**(17), 8433–8445.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, **185**(4154), 862–864.

Grund, S. *et al.* (2011). Comparison of hemagglutination inhibition assay, an elisa-based microneutralization assay and colorimetric microneutralization assay to detect antibody responses to vaccination against influenza a h1n1 2009 virus. *Journal of virological methods*, **171**(2), 369–373.

Guex, N. *et al.* (2009). Automated comparative protein structure modeling with swiss-model and swiss-pdbviewer: A historical perspective. *Electrophoresis*, **30**(S1), S162–S173.

Gupta, R. and Brunak, S. (2001). Prediction of glycosylation across the human proteome and the correlation to protein function. In *Pac. Symp. Biocomput*, volume 20022002, pages 310–322.

Gupta, V. *et al.* (2006). Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine*, **24**(18), 3881–3888.

Hale, B. G. *et al.* (2008a). The multifunctional ns1 protein of influenza a viruses. *Journal of General Virology*, **89**(10), 2359–2376.

Hale, B. G. *et al.* (2008b). Structure of an avian influenza a virus ns1 protein effector domain. *Virology*, **378**(1), 1–5.

Hale, B. G. *et al.* (2010). Inefficient Control of Host Gene Expression by the 2009 Pandemic H1n1 Influenza A Virus NS1 Protein. *Journal of Virology*, **84**(14), 6909–6922.

Harvey, W. T. (2016). *Quantifying the genetic basis of antigenic variation among human influenza A viruses*. Ph.D. thesis, University of Glasgow.

Hasegawa, M. *et al.* (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution*, **22**(2), 160–174.

Hatta, M. *et al.* (2001). Molecular Basis for High Virulence of Hong Kong H5n1 Influenza A Viruses. *Science*, **293**(5536), 1840–1842.

Hatta, M. *et al.* (2007). Growth of H5n1 Influenza A Viruses in the Upper Respiratory Tracts of Mice. *PLOS Pathogens*, **3**(10), 1–6.

Hayashi, T. *et al.* (2015). Influenza a virus protein pa-x contributes to viral growth and suppression of the host antiviral and immune responses. *Journal of virology*, **89**(12), 6442–6452.

Hensley, S. E. *et al.* (2009). Hemagglutinin Receptor Binding Avidity Drives Influenza A Virus Antigenic Drift. *Science*, **326**(5953), 734–736.

Herfst, S. *et al.* (2010). Introduction of virulence markers in pb2 of pandemic swine-origin influenza virus does not result in enhanced virulence or transmission. *Journal of Virology*, **84**(8), 3752–3758.

Herfst, S. *et al.* (2012). Airborne transmission of influenza a/h5n1 virus between ferrets. *science*, **336**(6088), 1534–1541.

Hernandez, M. *et al.* (2009). Sitehound-web: a server for ligand binding site identification in protein structures. *Nucleic acids research*, **37**(suppl 2), W413–W416.

Heui Seo, S. *et al.* (2002). Lethal H5n1 influenza viruses escape host anti-viral cytokine responses. *Nature Medicine*, **8**, 950.

Ho, H. P. *et al.* (2014). Effectiveness of seasonal influenza vaccinations against laboratory-confirmed influenza-associated infections among singapore military personnel in 2010-2013. *Influenza Other Respir Viruses*, **8**(5), 557–66.

Hobson, D. *et al.* (1972). The role of serum haemagglutination-inhibiting antibody in protection against challenge infection with influenza a and b viruses. *Epidemiology & Infection*, **70**(4), 767–777.

Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, **11**, 63–91.

Hopf, T. A. *et al.* (2017). Mutation effects predicted from sequence co-variation. *Nature biotechnology*, **35**(2), 128.

Horimoto, T. and Kawaoka, Y. (2005). Influenza: lessons from past pandemics, warnings from current incidents. *Nature Reviews Microbiology*, **3**(8), 591.

Horton, D. L. *et al.* (2010). Quantifying antigenic relationships among the lyssaviruses. *Journal of virology*, **84**(22), 11841–11848.

Hu, J. *et al.* (2015). Pa-x decreases the pathogenicity of highly pathogenic h5n1 influenza a virus in avian species by inhibiting virus replication and host response. *Journal of virology*, **89**(8), 4126–4142.

Huang, J.-W. *et al.* (2009a). Co-evolution positions and rules for antigenic variants of human influenza a/h3n2 viruses. *BMC bioinformatics*, **10**(1), S41.

Huang, S. W. *et al.* (2009b). Reemergence of enterovirus 71 in 2008 in taiwan: dynamics of genetic and antigenic evolution from 1998 to 2008. *J Clin Microbiol*, **47**(11), 3653–62.

Ibricevic, A. *et al.* (2006). Influenza virus receptor specificity and cell tropism in mouse and human airway epithelial cells. *Journal of virology*, **80**(15), 7469–7480.

Imai, M. *et al.* (2012). Experimental adaptation of an influenza h5 ha confers respiratory droplet transmission to a reassortant h5 ha/h1n1 virus in ferrets. *Nature*, **486**(7403), 420.

Islam, S. A. *et al.* (2017). Protein classification using modified n-grams and skip-grams. *Bioinformatics*, **34**(9), 1481–1487.

Ivan, F. X. *et al.* (2017). 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*, pages 21–26. ACM.

Jackson, D. *et al.* (2008). A new influenza virus virulence determinant: the ns1 protein four c-terminal residues modulate pathogenicity. *Proc Natl Acad Sci U S A*, **105**(11), 4381–6.

Jagger, B. *et al.* (2012). An overlapping protein-coding region in influenza a virus segment 3 modulates the host response. *Science*, **337**(6091), 199–204.

Jagger, B. W. *et al.* (2010). The pb2-e627k mutation attenuates viruses containing the 2009 h1n1 influenza pandemic polymerase. *MBio*, **1**(1), e00067–10.

Jiao, P. *et al.* (2008). A Single-Amino-Acid Substitution in the NS1 Protein Changes the Pathogenicity of H5n1 Avian Influenza Viruses in Mice. *Journal of Virology*, **82**(3), 1146–1154.

Jiménez-Alberto, A. *et al.* (2013). Analysis of adaptation mutants in the hemagglutinin of the influenza a (h1n1) pdm09 virus. *PloS one*, **8**(7), e70005.

Jorgensen, W. L. *et al.* (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, **79**(2), 926–935.

Joseph, U. *et al.* (2017). The ecology and adaptive evolution of influenza a interspecies transmission. *Influenza and other respiratory viruses*, **11**(1), 74–84.

Jukes, T. H. *et al.* (1969). Evolution of protein molecules. *Mammalian protein metabolism*, **3**(21), 132.

Kaczanowski, S. and Zielenkiewicz, P. (2010). Why similar protein sequences encode similar three-dimensional structures? *Theoretical Chemistry Accounts*, **125**(3-6), 643–650.

Kannan, S. and Kolandaivel, P. (2016). Computational studies of pandemic 1918 and 2009 h1n1 hemagglutinins bound to avian and human receptor analogs. *Journal of Biomolecular Structure and Dynamics*, **34**(2), 272–289.

Kash, J. C. *et al.* (2010). Prior infection with classical swine h1n1 influenza viruses is associated with protective immunity to the 2009 pandemic h1n1 virus. *Influenza and other respiratory viruses*, **4**(3), 121–127.

Katoh, K. and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, **30**(4), 772–780.

Kawaoka, Y. and Webster, R. G. (1988). Sequence requirements for cleavage activation of influenza virus hemagglutinin expressed in mammalian cells. *Proceedings of the National Academy of Sciences*, **85**(2), 324–328.

Kawashima, S. *et al.* (2007). Aaindex: amino acid index database, progress report 2008. *Nucleic acids research*, **36**(suppl_1), D202–D205.

Khanna, M. *et al.* (2008). Emerging influenza virus: a global threat. *Journal of biosciences*, **33**(4), 475–482.

Kiefer, F. *et al.* (2009). The swiss-model repository and associated resources. *Nucleic acids research*, **37**(suppl 1), D387–D392.

Kilbourne, E. D. (2006). Influenza pandemics of the 20th century. *Emerging Infectious Disease*, **12**.

Kilbourne, E. D. *et al.* (2002). The total influenza vaccine failure of 1947 revisited: Major intrasubtypic antigenic change can explain failure of vaccine in a post-world war ii epidemic. *Proceedings of the National Academy of Sciences*, **99**(16), 10748–10752.

Killian, M. L. (2008). Hemagglutination assay for the avian influenza virus. In *Avian influenza virus*, pages 47–52. Springer.

Kim, P. *et al.* (2018). Glycosylation of Hemagglutinin and Neuraminidase of Influenza A Virus as Signature for Ecological Spillover and Adaptation among Influenza Reservoirs. *Viruses*, **10**(4).

Kimura, H. *et al.* (1997). Interspecies transmission of influenza c virus between humans and pigs. *Virus Research*, **48**(1), 71 – 79.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, **16**(2), 111–120.

Klemm, C. *et al.* (2018). Immunomodulatory Nonstructural Proteins of Influenza A Viruses. *Trends in Microbiology*, **26**(7), 624–636.

Koday, M. T. *et al.* (2016). A computationally designed hemagglutinin stem-binding protein provides in vivo protection from influenza independent of a host immune response. *PLoS Pathog*, **12**(2), e1005409.

Koel, B. F. *et al.* (2014). Antigenic variation of clade 2.1 h5n1 virus is determined by a few amino acid substitutions immediately adjacent to the receptor binding site. *MBio*, **5**(3), e01070–14.

Košík, I. *et al.* (2015). The ubiquitination of the influenza a virus pb1-f2 protein is crucial for its biological function. *PLoS One*, **10**(4), e0118477.

Krammer, F. and Grabherr, R. (2010). Alternative influenza vaccines made by insect cells. *Trends in molecular medicine*, **16**(7), 313–320.

Krammer, F. *et al.* (2018). Influenza. *Nature Reviews Disease Primers*, **4**(1), 3.

Krieger, E. *et al.* (2003). Homology modeling. *Methods of biochemical analysis*, **44**, 509–524.

Kryazhimskiy, S. *et al.* (2011). Prevalence of epistasis in the evolution of influenza a surface proteins. *PLoS genetics*, **7**(2), e1001301.

Kumari, R. *et al.* (2014). g_mmpbsa: A gromacs tool for high-throughput mm-pbsa calculations. *Journal of chemical information and modeling*, **54**(7), 1951–1962.

Lazniewski, M. *et al.* (2017). The structural variability of the influenza a hemagglutinin receptor-binding site. *Briefings in functional genomics*, **17**(6), 415–427.

Le Goffic, R. *et al.* (2011). Transcriptomic analysis of host immune and cell death responses associated with the influenza a virus pb1-f2 protein. *PLOS Pathogens*, **7**(8), 1–12.

Lee, H. K. *et al.* (2013). Simplified large-scale sanger genome sequencing for influenza a/h3n2 virus. *PLoS One*, **8**(5), e64785.

Lee, H. K. *et al.* (2015a). Molecular surveillance of antiviral drug resistance of influenza a/h3n2 virus in singapore, 2009-2013. *PLoS One*, **10**(1), e0117822.

Lee, H. K. *et al.* (2015b). Predicting clinical severity based on substitutions near epitope a of influenza a/h3n2. *Infect Genet Evol*, **34**, 292–7.

Lee, H. K. *et al.* (2016). Contamination-controlled high-throughput whole genome sequencing for influenza a viruses using the miseq sequencer. *Sci Rep*, **6**, 33318.

Lee, M.-S. and Chen, J. S.-E. (2004). Predicting antigenic variants of influenza a/h3n2 viruses. *Emerging infectious diseases*, **10**(8), 1385.

Lee, V. J. *et al.* (2008). Twentieth century influenza pandemics in singapore. *Ann Acad Med Singapore*, **37**(6), 470–6.

Lee, V. J. *et al.* (2011). A clinical diagnostic model for predicting influenza among young adult military personnel with febrile respiratory illness in singapore. *PLoS One*, **6**(3), e17468.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13), 1658–9.

Li, Y. *et al.* (2013). Single Hemagglutinin Mutations That Alter both Antigenicity and Receptor Binding Avidity Influence Influenza Virus Antigenic Clustering. *Journal of Virology*, **87**(17), 9904–9910.

Liao, Y.-C. *et al.* (2008). Bioinformatics models for predicting antigenic variants of influenza a/h3n2 virus. *Bioinformatics*, **24**(4), 505–512.

Lin, Y. P. *et al.* (2012). Evolution of the receptor binding properties of the influenza a (h3n2) hemagglutinin. *Proceedings of the National Academy of Sciences*, **109**(52), 21474–21479.

Lindorff-Larsen, K. *et al.* (2010). Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, **78**(8), 1950–1958.

Lindstrom, S. E. *et al.* (1998). Phylogenetic Analysis of the Entire Genome of Influenza A (H3n2) Viruses from Japan: Evidence for Genetic Reassortment of the Six Internal Genes. *Journal of Virology*, **72**(10), 8021–8031.

Liu, Q. *et al.* (2013). Characteristics of human infection with avian influenza viruses and development of new antiviral agents. *Acta Pharmacologica Sinica*, **34**(10), 1257.

Long, J. S. *et al.* (2013). The effect of the pb2 mutation 627k on highly pathogenic h5n1 avian influenza virus is dependent on the virus lineage. *Journal of virology*, **87**(18), 9983–9996.

Ludi, A. B. *et al.* (2014). Antigenic variation of foot-and-mouth disease virus serotype a. *Journal of General Virology*, **95**(2), 384–392.

Lyons, D. and Lauring, A. (2018). Mutation and epistasis in influenza virus evolution. *Viruses*, **10**(8), 407.

Maines, T. R. *et al.* (2009). Transmission and pathogenesis of swine-origin 2009 a (h1n1) influenza viruses in ferrets and mice. *Science*, **325**(5939), 484–487.

Mair, C. M. *et al.* (2014). Receptor binding and ph stability — how influenza a virus hemagglutinin affects host-specific virus infection. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, **1838**(4), 1153 – 1168. Viral Membrane Proteins - Channels for Cellular Networking.

Mak, K.-K. and Pichika, M. R. (2018). Artificial intelligence in drug development: present status and future prospects. *Drug discovery today*.

Malathi, K. and Ramaiah, S. (2018). Bioinformatics approaches for new drug discovery: a review. *Biotechnology and Genetic Engineering Reviews*, **34**(2), 243–260.

Marc, D. (2014). Influenza virus non-structural protein ns1: interferon antagonism and beyond. *Journal of General Virology*, **95**(12), 2594–2611.

Mase, M. *et al.* (2006). Recent H5n1 avian Influenza A virus increases rapidly in virulence to mice after a single passage in mice. *Journal of General Virology*, **87**(12), 3655–3659.

Massin, P. *et al.* (2001). Residue 627 of pb2 is a determinant of cold sensitivity in rna replication of avian influenza viruses. *Journal of virology*, **75**(11), 5398–5404.

Mazur, I. *et al.* (2008). The proapoptotic influenza a virus protein pb1-f2 regulates viral polymerase activity by interaction with the pb1 protein. *Cellular microbiology*, **10**(5), 1140–1152.

McAuley, J. L. *et al.* (2007a). "expression of the 1918 influenza a virus pb1-f2 enhances the pathogenesis of viral and secondary bacterial pneumonia". *"Cell Host & Microbe"*, **2**(4), 240 – 249.

McAuley, J. L. *et al.* (2007b). Expression of the 1918 Influenza A Virus PB1-F2 Enhances the Pathogenesis of Viral and Secondary Bacterial Pneumonia. *Cell Host & Microbe*, **2**(4), 240–249.

McAuley, J. L. *et al.* (2010). PB1-F2 Proteins from H5n1 and 20th Century Pandemic Influenza Viruses Cause Immunopathology. *PLOS Pathogens*, **6**(7), 1–12.

Meher, P. K. *et al.* (2017). Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into chou's general pseaac. *Scientific reports*, **7**, 42362.

Mehle, A. and Doudna, J. A. (2009a). Adaptive strategies of the influenza virus polymerase for replication in humans. *Proceedings of the National Academy of Sciences*, **106**(50), 21312–21316.

Mehle, A. and Doudna, J. A. (2009b). Adaptive strategies of the influenza virus polymerase for replication in humans. *Proceedings of the National Academy of Sciences*, **106**(50), 21312–21316.

Merler, S. and Ajelli, M. (2009). The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings of the Royal Society B: Biological Sciences*, **277**(1681), 557–565.

Miotto, O. *et al.* (2008). Identification of human-to-human transmissibility factors in pb2 proteins of influenza a by large-scale mutual information analysis. *BMC Bioinformatics*, **9**(1), S18.

Mitzner, D. *et al.* (2009). Phosphorylation of the influenza a virus protein pb1-f2 by pkc is crucial for apoptosis promoting functions in monocytes. *Cellular microbiology*, **11**(10), 1502–1516.

Morens, D. M. *et al.* (2004). The challenge of emerging and re-emerging infectious diseases. *Nature*, **430**(6996), 242–249.

Muramoto, Y. *et al.* (2013). Identification of novel influenza a virus proteins translated from pa mrna. *Journal of virology*, **87**(5), 2455–2462.

Nair, H. *et al.* (2011). Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. *The Lancet*, **378**(9807), 1917–1930.

Narasaraju, T. *et al.* (2009). Adaptation of human influenza h3n2 virus in a mouse pneumonitis model: insights into viral virulence, tissue tropism and host pathogenesis. *Microbes and infection*, **11**(1), 2–11.

NCBI, R. C. (2014). Database resources of the national center for biotechnology information. *Nucleic acids research*, **42**(Database issue), D7.

Ndifon, W. *et al.* (2009). On the use of hemagglutination-inhibition for influenza surveillance: surveillance data are predictive of influenza vaccine effectiveness. *Vaccine*, **27**(18), 2447–2452.

Nedland, H. *et al.* (2018). Serological evidence for the co-circulation of two lineages of influenza D viruses in equine populations of the Midwest United States. *Zoonoses and public health*, **65**(1), e148–e154.

Neher, R. A. and Bedford, T. (2015). nextflu: Real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, **31**(21), 3546–3548.

Neher, R. A. *et al.* (2016). Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proceedings of the National Academy of Sciences*, **113**(12), E1701–E1709.

Ng, T. P. *et al.* (2002). Influenza in singapore: assessing the burden of illness in the community. *Ann Acad Med Singapore*, **31**(2), 182–8.

Ni, F. *et al.* (2014). The roles of hemagglutinin phe-95 in receptor binding and pathogenicity of influenza b virus. *Virology*, **450**, 71–83.

Nicholls, J. M. *et al.* (2008). Evolving complexities of influenza virus and its receptors. *Trends in microbiology*, **16**(4), 149–157.

Niefind, K. and Schomburg, D. (1991). Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles. *Journal of molecular biology*, **219**(3), 481–497.

Obenauer, J. C. *et al.* (2006). Large-Scale Sequence Analysis of Avian Influenza Isolates. *Science*, **311**(5767), 1576–1580.

Osterhaus, A. *et al.* (2000). Influenza b virus in seals. *Science*, **288**(5468), 1051–1053.

Osterholm, M. T. *et al.* (2012). Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, **12**(1), 36–44.

Palese, P. (2006). Making better influenza virus vaccines? *Emerging infectious diseases*, **12**(1), 61.

Pan, D. *et al.* (2012). Molecular mechanism of the enhanced virulence of 2009 pandemic influenza a (h1n1) virus from d222g mutation in the hemagglutinin: a molecular modeling study. *Journal of molecular modeling*, **18**(9), 4355–4366.

Parrinello, M. and Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, **52**(12), 7182–7190.

Pauly, M. D. *et al.* (2017). Epistatic interactions within the influenza a virus polymerase complex mediate mutagen resistance and replication fidelity. *MSphere*, **2**(4), e00323–17.

Peng, Y. *et al.* (2017). A universal computational model for predicting antigenic variants of influenza A virus based on conserved antigenic structures. *Scientific Reports*, **7**, 42051.

Philpott, M. *et al.* (1990). Hemagglutinin mutations related to attenuation and altered cell tropism of a virulent avian influenza a virus. *Journal of virology*, **64**(6), 2941–2947.

Ping, J. *et al.* (2011). Genomic and protein structural maps of adaptive evolution of human influenza a virus to increased virulence in the mouse. *PloS one*, **6**(6), e21740.

Plotch, S. J. *et al.* (1981). A unique cap(m7gpppxm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription. *Cell*, **23**(3), 847 – 858.

Plotkin, J. B. *et al.* (2002). Hemagglutinin sequence clusters and the antigenic evolution of influenza a virus. *Proceedings of the National Academy of Sciences*, **99**(9), 6263–6268.

Plotkin, S. (2014). History of vaccination. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(34), 12283–12287.

Poole, E. *et al.* (2004). Functional domains of the influenza a virus pb2 protein: identification of np-and pb1-binding sites. *Virology*, **321**(1), 120–133.

Pu, J. *et al.* (2010). Synergism of co-mutation of two amino acid residues in ns1 protein increases the pathogenicity of influenza virus in mice. *Virus research*, **151**(2), 200–204.

Qiang, X. and Kou, Z. (2010). Prediction of interspecies transmission for avian influenza a virus based on a back-propagation neural network. *Mathematical and Computer Modelling*, **52**(11), 2060 – 2065. The BIC-TA 2009 Special Issue.

Rambaut, A. *et al.* (2014). Tracer v1. 6 http://beast. bio. ed. ac. uk. *Tracer (Online 2015, May 29)*.

Ramírez, D. and Caballero, J. (2016). Is it reliable to use common molecular docking methods for comparing the binding affinities of enantiomer pairs for their protein target? *International journal of molecular sciences*, **17**(4), 525.

Rapaport, D. C. and Rapaport, D. C. R. (2004). *The art of molecular dynamics simulation*. Cambridge university press.

Rédei, G. P. (2008). *Encyclopedia of genetics, genomics, proteomics, and informatics*. Springer Science & Business Media.

Reich, S. *et al.* (2014). Structural insight into cap-snatching and RNA synthesis by influenza polymerase. *Nature*, **516**, 361.

Remmert, M. *et al.* (2012). Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, **9**(2), 173.

Rimmelzwaan, G. *et al.* (2004). Functional compensation of a detrimental amino acid substitution in a cytotoxic-t-lymphocyte epitope of influenza a viruses by comutations. *Journal of virology*, **78**(16), 8946–8949.

Rimmelzwaan, G. *et al.* (2005). Full restoration of viral fitness by multiple compensatory co-mutations in the nucleoprotein of influenza a virus cytotoxic t-lymphocyte escape mutants. *Journal of general virology*, **86**(6), 1801–1805.

Rogers, G. *et al.* (1983). Single amino acid substitutions in influenza haemagglutinin change receptor binding specificity. *Nature*, **304**(5921), 76–78.

Rondy, M. *et al.* (2018). Interim 2017/18 influenza seasonal vaccine effectiveness: combined results from five european studies. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, **23**(9), 18–00086.

Rose, A. S. *et al.* (2016). Web-based molecular graphics for large complexes. In *Proceedings of the 21st International Conference on Web3D Technology*, pages 185–186. ACM.

Rose, A. S. *et al.* (2018). Ngl viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**(21), 3755–3758.

Šali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, **234**(3), 779–815.

Salmaso, V. (2018). Exploring protein flexibility during docking to investigate ligand-target recognition.

Salmaso, V. and Moro, S. (2018). Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Frontiers in pharmacology*, **9**.

Samji, T. (2009). Influenza a: understanding the viral life cycle. *The Yale journal of biology and medicine*, **82**(4), 153.

Sauter, N. K. *et al.* (1989). Hemagglutinins from two influenza virus variants bind to sialic acid derivatives with millimolar dissociation constants: a 500-mhz proton nuclear magnetic resonance study. *Biochemistry*, **28**(21), 8388–8396.

Sauter, N. K. *et al.* (1992). Binding of influenza virus hemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and x-ray crystallography. *Biochemistry*, **31**(40), 9609–21.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in r. *Bioinformatics*, **27**(4), 592–593.

Seah, S. G. *et al.* (2010). Viral agents responsible for febrile respiratory illnesses among military recruits training in tropical singapore. *J Clin Virol*, **47**(3), 289–92.

Selman, M. *et al.* (2012). Adaptive mutation in influenza a virus non-structural gene is linked to host switching and induces a novel protein by alternative splicing. *Emerging microbes & infections*, **1**(1), 1–10.

Severin, C. *et al.* (2016). The cap-binding site of influenza virus protein pb2 as a drug target. *Acta Crystallographica Section D: Structural Biology*, **72**(2), 245–253.

Sharma, A. *et al.* (2013). Computational approach for designing tumor homing peptides. *Scientific reports*, **3**, 1607.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 683–690.

Shen, Y. and Lu, H. (2017). Global concern regarding the fifth epidemic of human infection with avian influenza a (h7n9) virus in china. *Bioscience trends*, **11**(1), 120–121.

Shih, A. C.-C. *et al.* (2007). Simultaneous amino acid substitutions at antigenic sites drive influenza a hemagglutinin evolution. *Proceedings of the National Academy of Sciences*, **104**(15), 6283–6288.

Shtyrya, Y. A. *et al.* (2009). Influenza virus neuraminidase: structure and function. *Acta naturae*, **1**(2), 26–32.

Shu, Y. and McCauley, J. (2017). Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, **22**(13).

Simonetti, F. L. *et al.* (2013). Mistic: mutual information server to infer coevolution. *Nucleic acids research*, **41**(W1), W8–W14.

Skehel, J. J. and Wiley, D. C. (2000). Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annual review of biochemistry*, **69**(1), 531–569.

Skehel, J. J. *et al.* (1984). A carbohydrate side chain on hemagglutinins of hong kong influenza viruses inhibits recognition by a monoclonal antibody. *Proceedings of the National Academy of Sciences*, **81**(6), 1779–1783.

Smith, D. J. *et al.* (1999). Variable efficacy of repeated annual influenza vaccination. *Proceedings of the National Academy of Sciences*, **96**(24), 14001–14006.

Smith, D. J. *et al.* (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science*, **305**(5682), 371–376.

Smith, W. *et al.* (1933). A virus obtained from influenza patients. *Lancet*, pages 66–8.

Soundararajan, V. *et al.* (2011). Networks link antigenic and receptor-binding sites of influenza hemagglutinin: mechanistic insight into fitter strain propagation. *Scientific reports*, **1**, 200.

Squires, R. B. *et al.* (2012). Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and other respiratory viruses*, **6**(6), 404–416.

Starr, T. N. and Thornton, J. W. (2016). Epistasis in protein evolution. *Protein Science*, **25**(7), 1204–1218.

Steel, J. *et al.* (2009). Transmission of Influenza Virus in a Mammalian Host Is Increased by PB2 Amino Acids 627k or 627e/701n. *PLOS Pathogens*, **5**(1), 1–11.

Stein, R. A. (2009). Lessons from outbreaks of h1n1 influenza. *Annals of internal medicine*, **151**(1), 59–62.

Stevens, J. *et al.* (2004). Structure of the uncleaved human h1 hemagglutinin from the extinct 1918 influenza virus. *Science*, **303**(5665), 1866–1870.

Stray, S. J. *et al.* (2000). Influenza virus infection of desialylated cells. *Glycobiology*, **10**(7), 649–658.

Su, C. T.-T. *et al.* (2013). 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*, **14**(16), S7.

Su, S. *et al.* (2017). Novel influenza d virus: Epidemiology, pathology, evolution and biological characteristics. *Virulence*, **8**(8), 1580–1591.

Subbarao, E. K. *et al.* (1993). A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. *Journal of Virology*, **67**(4), 1761–1764.

Subbarao, K. and Joseph, T. (2007). Scientific barriers to developing vaccines against avian influenza viruses. *Nature Reviews Immunology*, **7**, 267.

Sun, H. *et al.* (2013). Using sequence data to infer the antigenicity of influenza virus. *mBio*, **4**(4), e00230–13.

Taft, A. S. *et al.* (2015). Identification of mammalian-adapting mutations in the polymerase complex of an avian h5n1 influenza virus. *Nature communications*, **6**.

Tan, A. L. *et al.* (2015). Surveillance and clinical characterization of influenza in a university cohort in singapore. *PLoS One*, **10**(3), e0119485.

Taubenberger, J. K. and Morens, D. M. (2006). 1918 influenza: the mother of all pandemics. *Emerging Infectious Disease*, **12**.

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, **17**, 57–86.

Te Velthuis, A. J. W. and Fodor, E. (2016). Influenza virus RNA polymerase: insights into the mechanisms of viral RNA synthesis. *Nature reviews. Microbiology*, **14**(8), 479–493.

Tewawong, N. *et al.* (2015). Molecular epidemiology and phylogenetic analyses of influenza b virus in thailand during 2010 to 2014. *PLoS One*, **10**(1), e0116302.

Thompson, W. W. *et al.* (2009). Estimates of us influenza-associated deaths made using four different methods. *Influenza and Other Respiratory Viruses*, **3**(1), 37–49.

Tomii, K. and Kanehisa, M. (1996). Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Engineering, Design and Selection*, **9**(1), 27–36.

Tong, S. *et al.* (2013). New world bats harbor diverse influenza a viruses. *PLOS Pathogens*, **9**(10), 1–12.

Treanor, J. (2004). Influenza vaccine—outmaneuvering antigenic shift and drift. *New England Journal of Medicine*, **350**(3), 218–220.

Trifonov, V. *et al.* (2009a). The contribution of the pb1-f2 protein to the fitness of influenza a viruses and its recent evolution in the 2009 influenza a (h1n1) pandemic virus. *PLoS currents*, **1**.

Trifonov, V. *et al.* (2009b). Geographic dependence, surveillance, and origins of the 2009 influenza a (h1n1) virus. *New England Journal of Medicine*, **361**(2), 115–119. PMID: 19474418.

Trock, S. C. *et al.* (2012). Development of an influenza virologic risk assessment tool. *Avian Diseases*, **56**(4s1), 1058–1061.

Tscherne, D. M. and García-Sastre, A. (2011). Virulence determinants of pandemic influenza viruses. *The Journal of clinical investigation*, **121**(1), 6–13.

Tumpey, T. M. *et al.* (2005). Characterization of the reconstructed 1918 spanish influenza pandemic virus. *science*, **310**(5745), 77–80.

Tusche, C. *et al.* (2012). Detecting patches of protein sites of influenza a viruses under positive selection. *Molecular biology and evolution*, **29**(8), 2063–2071.

Twu, K. Y. *et al.* (2007). The h5n1 influenza virus ns genes selected after 1998 enhance virus replication in mammalian cells. *J Virol*, **81**(15), 8112–21.

Tzeng, J. *et al.* (2008). Multidimensional scaling for large genomic data sets. *BMC bioinformatics*, **9**(1), 179.

Uno, T. *et al.* (2004). Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In *Fimi*, volume 126.

Varga, Z. T. *et al.* (2011). The influenza virus protein pb1-f2 inhibits the induction of type i interferon at the level of the mavs adaptor protein. *PLoS pathogens*, **7**(6), e1002067.

Vidy, A. *et al.* (2016). The influenza virus protein pb1-f2 increases viral pathogenesis through neutrophil recruitment and nk cells inhibition. *PloS one*, **11**(10), e0165361.

Vijaykrishna, D. *et al.* (2010). Reassortment of pandemic h1n1/2009 influenza a virus in swine. *Science*, **328**(5985), 1529–1529.

Virk, R. K. *et al.* (2014). Prospective surveillance and molecular characterization of seasonal influenza in a university cohort in singapore. *PLoS One*, **9**(2), e88345.

Virk, R. K. *et al.* (2017). Molecular evidence of transmission of influenza a/h1n1 2009 on a university campus. *PLoS One*, **12**(1), e0168596.

Wahlgren, J. (2011). Influenza a viruses: an ecology review. *Infection ecology & epidemiology*, **1**(1), 6004.

Wang, Q. *et al.* (2007). Structural basis for receptor specificity of influenza b virus hemagglutinin. *Proc Natl Acad Sci U S A*, **104**(43), 16874–9.

Wang, Q. *et al.* (2018). Host interaction analysis of pa-n155 and pa-n182 in chicken cells reveals an essential role of uba52 for replication of h5n1 avian influenza virus. *Frontiers in Microbiology*, **9**, 936.

Wang, X. *et al.* (2016). Computational approach for predicting the conserved b-cell epitopes of hemagglutinin h7 subtype influenza virus. *Experimental and therapeutic medicine*, **12**(4), 2439–2446.

Watanabe, T. *et al.* (2011). Avian-type receptor-binding ability can increase influenza virus pathogenicity in macaques. *Journal of Virology*, **85**(24), 13195–13203.

Waterhouse, A. *et al.* (2018). Swiss-model: homology modelling of protein structures and complexes. *Nucleic acids research*, **46**(W1), W296–W303.

Webster, R. *et al.* (1982). Molecular mechanisms of variation in influenza viruses. *Nature*, **296**(5853), 115.

Webster, R. G. (1999). Antigenic variation in influenza viruses. In *Origin and evolution of viruses*, pages 377–390. Elsevier.

Webster, R. G. and Laver, W. G. (1980). Determination of the number of nonoverlapping antigenic areas on hong kong (h3n2) influenza virus hemagglutinin with monoclonal antibodies and the selection of variants with potential epidemiological significance. *Virology*, **104**(1), 139–48.

Wei, L. *et al.* (1997). Using the radial distributions of physical features to compare amino acid environments and align amino acid sequences. In *Pac Symp Biocomput*, volume 5, pages 465–476. Citeseer.

Weis, W. *et al.* (1988). Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature*, **333**(6172), 426.

Wen, J. and Zhang, Y. (2009). A 2d graphical representation of protein sequence and its numerical characterization. *Chemical Physics Letters*, **476**(4-6), 281–286.

Wiley, D. *et al.* (1981). Structural identification of the antibody-binding sites of hong kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*, **289**(5796), 373.

Wilschut, J. *et al.* (2005). The influenza virus: structure and replication. *Influenza. Elsevier, London*, pages 23–44.

Wilson, I. A. and Cox, N. J. (1990). Structural basis of immune recognition of influenza virus hemagglutinin. *Annual review of immunology*, **8**(1), 737–787.

Wise, H. M. *et al.* (2009). A complicated message: Identification of a novel pb1-related protein translated from influenza a virus segment 2 mrna. *Journal of virology*, **83**(16), 8021–8031.

Wise, H. M. *et al.* (2012). Identification of a novel splice variant form of the influenza a virus m2 ion channel with an antigenically distinct ectodomain. *PLoS pathogens*, **8**(11), e1002998.

World Health Organization (2014). Who information for molecular diagnosis of influenza virus in humans - update. march 2014. Report, WHO.

World Health Organization *et al.* (2018a). Influenza - surveillance and monitoring.

World Health Organization *et al.* (2018b). Influenza - vaccines.

World Health Organization (WHO) (2018a). Global influenza surveillance and response system (gisrs).

World Health Organization (WHO) (2018b). Influenza (seasonal).

Wu, C. (2014). *Phenotype Inference from Genotype in RNA Viruses*. Ph.D. thesis, Carnegie Mellon University.

Wu, C.-Y. *et al.* (2017a). Influenza a surface glycosylation and vaccine design. *Proceedings of the National Academy of Sciences*, **114**(2), 280–285.

Wu, N. C. *et al.* (2017b). Diversity of functionally permissive sequences in the receptor-binding site of influenza hemagglutinin. *Cell host & microbe*, **21**(6), 742–753.

Xia, Z. *et al.* (2009). Using a mutual information-based site transition network to map the genetic evolution of influenza a/h3n2 virus. *Bioinformatics*, **25**(18), 2309–2317.

Xiang, N. (2016). Assessing change in avian influenza a (h7n9) virus infections during the fourth epidemic—china, september 2015–august 2016. *MMWR. Morbidity and Mortality Weekly Report*, **65**.

Xiao, X. *et al.* (2013). iamp-2l: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical biochemistry*, **436**(2), 168–177.

Xiong, X. *et al.* (2014). Receptor binding properties of the influenza virus hemagglutinin as a determinant of host range. *Current topics in microbiology and immunology*, **385**, 63–91.

Xu, R. *et al.* (2012a). Functional balance of the hemagglutinin and neuraminidase activities accompanies the emergence of the 2009 h1n1 influenza pandemic. *Journal of virology*, **86**(17), 9221–9232.

Xu, R. *et al.* (2012b). Structural characterization of the hemagglutinin receptor specificity from the 2009 h1n1 influenza pandemic. *J Virol*, **86**(2), 982–90.

Xu, R. *et al.* (2013). Preferential recognition of avian-like receptors in human influenza a h7n9 viruses. *Science*, **342**(6163), 1230–1235.

Yamada, S. *et al.* (2006). Haemagglutinin mutations responsible for the binding of h5n1 influenza a viruses to human-type receptors. *Nature*, **444**(7117), 378.

Yamayoshi, S. *et al.* (2016). Identification of a novel viral protein expressed from the pb2 segment of influenza a virus. *Journal of virology*, **90**(1), 444–456.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *J Mol Evol*, **39**(1), 105–11.

Yang, Z. (2000). Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus a. *Journal of molecular evolution*, **51**(5), 423–432.

Yap, J. *et al.* (2012). Differing clinical characteristics between influenza strains among young healthy adults in the tropics. *BMC Infect Dis*, **12**, 12.

Yen, H.-L. *et al.* (2011). Hemagglutinin–neuraminidase balance confers respiratory-droplet transmissibility of the pandemic h1n1 influenza virus in ferrets. *Proceedings of the National Academy of Sciences*, **108**(34), 14264–14269.

Yip, K. Y. *et al.* (2007). An integrated system for studying residue coevolution in proteins. *Bioinformatics*, **24**(2), 290–292.

Yu, G. *et al.* (2017). Ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, **8**(1), 28–36.

Yuan, P. *et al.* (2009). Crystal structure of an avian influenza polymerase PAN reveals an endonuclease active site. *Nature*, **458**, 909.

Yuanji, G. *et al.* (1983). Isolation of influenza c virus from pigs and experimental infection of pigs with influenza c virus. *Journal of General Virology*, **64**(1), 177–182.

Zamarin, D. *et al.* (2005). Influenza virus pb1-f2 protein induces cell death through mitochondrial ant3 and vdac1. *PLOS Pathogens*, **1**(1).

Zhang, W. *et al.* (2013). Molecular basis of the receptor binding specificity switch of the hemagglutinins from both the 1918 and 2009 pandemic influenza a viruses by a d225g substitution. *J Virol*, **87**(10), 5949–58.

Zhou, X. *et al.* (2018a). Computational analysis of the receptor binding specificity of novel influenza a/h7n9 viruses. *BMC genomics*, **19**(2), 88.

Zhou, X. *et al.* (2018b). A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza a viruses. *BMC Genomics*, **19**(10), 936.

Zhou, X. *et al.* (2018c). An encoding scheme capturing generic priors and properties of amino acids improves protein classification. *IEEE Access*.

Zhou, X. *et al.* (2019). Detecting co-occurring and sequential mutations of influenza a/h1n1, a/h3n2 and b viruses. *Viruses*, **10**(8), 407.

Zhu, H. *et al.* (2010). Substitution of lysine at 627 position in pb2 protein does not change virulence of the 2009 pandemic h1n1 virus in mice. *Virology*, **401**(1), 1 – 5.

Zhu, W. *et al.* (2012a). Mutations in polymerase genes enhanced the virulence of 2009 pandemic h1n1 influenza virus in mice. *PloS one*, **7**(3), e33383.

Zhu, X. *et al.* (2012b). Influenza Virus Neuraminidases with Reduced Enzymatic Activity That Avidly Bind Sialic Acid Receptors. *Journal of Virology*, **86**(24), 13371–13383.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary Divergence and Convergence in Proteins. In V. Bryson and H. J. Vogel, editors, *Evolving Genes and Proteins*, pages 97 – 166. Academic Press.

Appendix A

CFreeEnS improves protein classification

A dilemma in feature engineering is that domain-specific knowledge can benefit the design of an effective data representation for a specific dataset, but it can be tedious, time-consuming and limited by human subjectivity. A representation with more generic priors can help automate the design of features and facilitate large-scale analyses of different datasets. The explosion of protein sequence data and the increasing availability of computing power make the latter more urgent and promising. When it comes to the protein classification problem, existing methods of protein representation have achieved moderate performance, and can compete with the design incorporated with domain-specific knowledge especially classifying proteins into families (Frickey and Lupas, 2004).

The effectiveness of CFreeEnS on casting the protein sequences to numeric representations, the encoding scheme was applied to different tasks of protein classification. The protein classification problems are about detecting peptides with different functions, namely antimicrobial peptides, tumor homing peptides, hemolytic peptides and phage virion proteins. Sequences in the four selected datasets not only have different biological backgrounds, but also vary in the data size, sample distribution and sequence length distribution. The number of training samples range from hundreds to thousands, and the lengths of peptides range from several residues to a few hundred. The predicting accuracy of CFreeEnS exceeds 88% on each dataset, regardless of being balanced or not for the positive and negative classes, with short or long peptides. The results show the robustness of CFreeEnS, suggesting that CFreeEnS encodes the generic priors of protein sequences into features representative enough for distinguishing them at the functional level.

This appendix is mainly based on the publication Zhou *et al.* (2018c).

A.1 Datasets for protein classification

An overview of the datasets for protein classification is presented in Table A.1. The abbreviations of datasets are identical with the methods proposed explicitly for them.

Table A.1: An overview of datasets for protein classification

Datasets (Methods)	**Description**	**Sequence Lengths**	**# Sequences**
iAMP	Antimicrobial peptides data involves anti-bacteria, anti-cancer, anti-fungal and anti-viral sequences. The task is to classify antimicrobial peptides from non-antimicrobial peptides.	10–255; Median: 26	6214
TumorHPD	TumorHPD classifies tumor homing peptides, helping to design analogs of tumor homing ability	4–31; Median: 10	2240
HemoPI	Identifying the hemolytic peptides from non-hemolytic peptides.	4–98; Median: 18	1104
PVPred	Phage virion proteins are classified from other non-phage virion proteins	23–1825; Median: 213	337

A.1.1 iAMP

The iAMP dataset, abbreviated for identifying antimicrobial peptides, includes antibacterial peptides, antiviral peptides, and antifungal peptides. The antimicrobial peptides are important host defense molecules in the innate immune system against pathogens. Computational identification of AMPs saves the researchers from expensive *in vitro* wet-lab experiments. Previous analyses tried incorporating several designed features, including the distribution patterns of amino acids (Bhadra *et al.*, 2018), pseudo amino acid composition and some selected physiochemical features (Xiao *et al.*, 2013; Meher *et al.*, 2017). The benchmark dataset of iAMP includes 3107 positive samples and an equal number of negative samples generated from UniProt. Sequence lengths of antimicrobial peptides in the dataset vary from 10 to 255, with a median of 26.

A.1.2 TumorHPD

The TumorHPD is a web sever for recognizing tumor homing peptides, which are able to recognize tumor cells (Sharma *et al.*, 2013). Amino acid composition profile, dipeptide composition, and binary profile are generated in the TumorHPD to capture the features of input sequences. The benchmark dataset includes 651 and 469 positive samples obtained from the TumorHoPe database as the training set and validation set respectively. An equal number of negative samples are generated from Swiss-Prot database. The median of lengths of the tumor homing peptides is 10.

A.1.3 HemoPI

The HemoPI, short for hemolytic peptide identification, is to screen hemolytic peptides from the non-hemolytic, where quantitative matrices are developed for measuring the hemotoxicity (Chaudhary *et al.*, 2016). Motifs observed in hemolytic peptides are utilized as features to differentiate them from the non-hemolytic ones. There are 552 positive samples which are

experimentally validated highly hemolytic peptides from the Hemolytik Database. The same amount of negative samples are generated from SwissProt.

A.1.4 PVPred

PVPred predicts the phage virion proteins by analyzing the variance and optimizing the g-gap dipeptide (Ding *et al.*, 2014). Most phage virion proteins in the dataset have long primary sequences with several hundred residues. The sequence lengths vary from 23 to 1825, but with a median of 213. There are 99 positive samples and 208 negative samples in the training set; 11 positive samples and 19 negative samples in the validation set.

A.2 Results of protein classification

The four datasets are encoded by CFreeEnS, using all the available 566 amino acid indexes in the AAindex database. Sequences with varying length are represented by vectors with length 566. Columns with high correlation are dropped before inputting into a downstream learning method. To compare the effectiveness of data representation, we keep the same downstream learning procedure as those traditional methods using designed features for each dataset. Besides, the *m-NGSG*, a state-of-the-art method treating protein sequences as normal text and generating features from a text mining perspective, has been applied to the four datasets (Islam *et al.*, 2017).

Table A.2 shows the classification results of CFreeEnS, taking 0.95 as the dropout threshold. There are 190, 146, 170 and 211 features dropped for *iAMP*, *TumorHPD*, *HemoPI* and *PVPred* respectively. The performance of CFreeEnS on each dataset is evaluated with accuracy, precision, recall, F-score, AUC, geometric-mean (g-mean) and Matthews correlation coefficient (MCC). The CFreeEnS works best on the *HemoPI* database with the highest AUC (0.936) and MCC (0.874), while worst on the *TumorHPD* database with a moderate AUC of 0.881.

When comparing with other methods, as presented in Figure A.1, we can observe that CFreeEnS outperforms the state-of-the-art method *m-NGSG* and traditional methods using designed features. The predicting accuracy scores of the four datasets are improved. The increases of accuracy range from 5.54% to 14.14% when compared with the traditional method, from 0.82% to 13.44% when compared with *m-NGSG*. Although the accuracy of predicting tumor homing peptides seems not high enough, it has been improved by 5.54% compared with the traditional method using several designed profiles. Even compared with the *m-NGSG*, the accuracy has been increased by 4.66%. The fact that tumor homing peptides are generally short with a median of 10 may partially contribute to the difficulty in accurate prediction. Both *m-NGSG* and CFreeEnS work well on the *iAMP* dataset, which may benefit from the large amount and balanced training samples. It is worth noting that the CFreeEnS also works well on the *PVPred* dataset with a small number of training samples.

Table A.2: The classification results of CFreeEnS applied to iAMP, TumorHPD, HemoPI and PVPred datasets.

Datasets	**Accuracy**	**Precision**	**Recall**	**F-score**	**AUC**	**G-mean**	**MCC**
iAMP	0.9207	0.9261	0.9203	0.9201	0.9203	0.9185	0.8464
TumorHPD	0.8806	0.8987	0.8806	0.8792	0.8806	0.8741	0.7791
HemoPI	0.9364	0.9377	0.9364	0.9363	0.9364	0.9360	0.8740
PVPred	0.9333	0.9397	0.9333	0.9317	0.9091	0.9045	0.8604

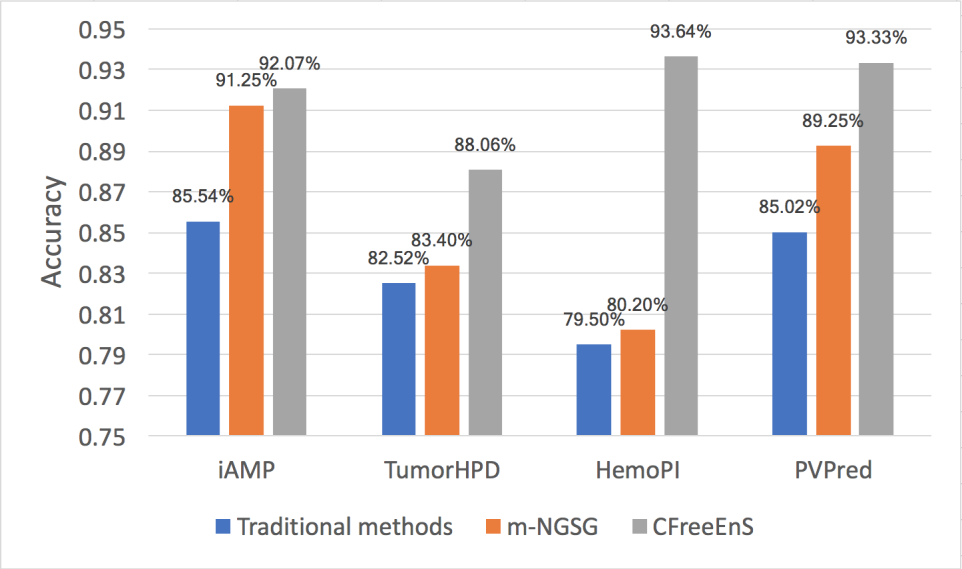


Figure A.1: Predicting accuracy of CFreeEnS on protein classification compared with traditional methods and *m-NGSG*

Appendix B

Supplements for MD simulation

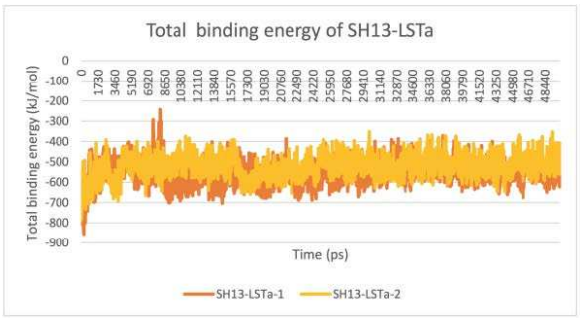
In Section 5.3, a 50 ns MD simulation for SH13-LSTa, SH13-LSTc, TW17-LSTa and TW17-LSTc were performed. Each trajectory fluctuated within a small range after 30 ns. To ensure convergence and to assess the reliability of the simulation data, a simulation replica was performed per system. The results for a replica 50 ns MD simulation are presented in this appendix.

B.1 Total binding energy

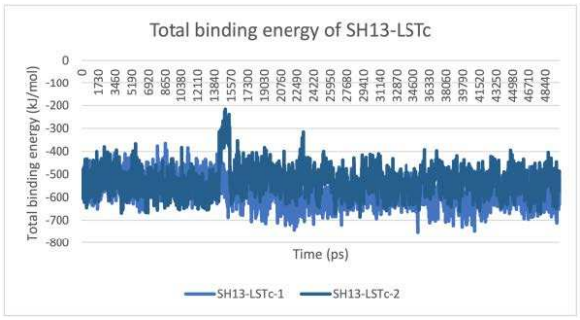
Figure B.1 presents the comparison of total binding energy for each protein-ligand simulation. For SH13-LSTa, SH13-LSTc and TW17-LSTa, the two rounds 50 ns MD simulation share a similar pattern of total binding energy. The total energy of the second round of TW17-LSTc MD simulation has smaller deviation than the first round over the whole process. But the TW17-LSTc systems in the two rounds of simulation converged after 30 ns.

B.2 Average binding energy

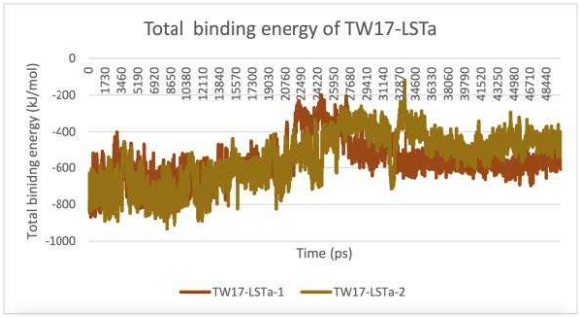
Table B.1 lists the averages of total binding energy of the second round MD simulation for the HA-LSTa/LSTc complexes. The results of a replica simulation are consistent with the first 50 ns MD simulation. The mutant TW17 HA obtained the largest binding energy, and enhanced the binding with two types receptors, especially LSTc. Both SH13 and TW17 strains had binding preferences for LSTc.



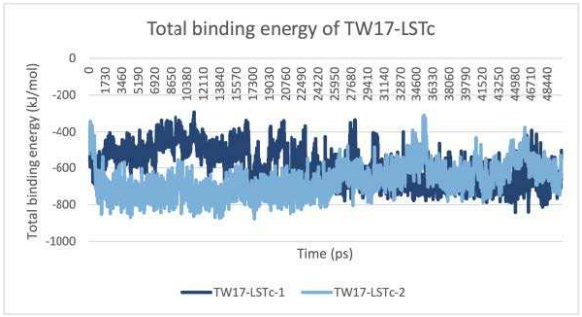
(a) Total binding energy of SH13-LSTa.



(b) Total binding energy of SH13-LSTc.



(c) Total binding energy of TW17-LSTa.



(d) Total binding energy of TW17-LSTc.

Figure B.1: Comparison of total binding energy between two rounds 50 ns MD simulation for SH13-LSTa, SH13-LSTc, TW17-LSTa and TW17-LSTc.

Table B.1: Average total binding energy (kJ/mol) of the HA-LSTa/LSTc complexes.

	LSTa	**LSTc**	**$^{1}\Delta E_1$**
SH13	-539.168	-554.499	+15.331
TW17	-557.506	-633.218	+75.712
$^{2}\Delta E_2$	+18.338	+78.719	

1. Binding preference of HA protein: $\Delta E_1 = \Delta E_{HA,LSTa} - \Delta E_{HA,LSTc}$
2. Difference of HAs binding to receptors: $\Delta E_2 = \Delta E_{SH13,receptor} - \Delta E_{TW17,receptor}$