

國立交通大學

生物資訊及系統生物研究所

博士論文

A 型 H3N2 流行性感冒病毒之
基因演化與抗原性演化之關聯性研究

A Study of Relationships between Genetic and Antigenic
Evolution of Influenza A (H3N2) Viruses

研究生：黃章維

指導教授：楊進木 教授

中華民國九十九年九月

A 型 H3N2 流行性感冒病毒之基因演化與抗原性演化之關聯性研究

A Study of Relationships between Genetic and Antigenic Evolution of
Influenza A (H3N2) Viruses


研 究 生：黃章維

Student : Jhang-Wei Huang

指導教授：楊進木

Advisor : Jinn-Moon Yang

國 立 交 通 大 學
生 物 資 訊 及 系 統 生 物 研 究 所
博 士 論 文

The logo of National Chiao Tung University is a circular emblem with a gear-like border. Inside the circle, there is a stylized building and the year '1896'. The text 'A Thesis' is written across the center of the logo.

A Thesis
Submitted to Institute of Bioinformatics and Systems Biology
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
PhD
in

Bioinformatics and Systems Biology

September 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年九月

A 型 H3N2 流行性感冒病毒之 基因演化與抗原性演化之關聯性研究

研究生:黃章維

指導教授: 楊進木博士

國立交通大學 生物資訊與系統生物研究所 博士班

摘 要

流行性感冒病毒經常對人類造成大規模的感染與死亡。發生在病毒表面紅血球凝集素(hemagglutinin 簡稱為 HA)上的胺基酸突變在逐漸累積之後會產生不同抗原特性的病毒株(稱為抗原性變異株),並且造成抗原性漂變(antigenic drift),此時疫苗常常需要在下一波疫情來臨前重新設計以提供足夠的保護力。目前人們對於流感病毒的基因演化(genetic evolution)與抗原性演化(antigenic evolution)間之關聯性尚未十分清楚,探究它們的關聯性對於公共衛生與疫苗發展是一個重要且有高度急迫性的議題。

在流感病毒中,A 型(H3N2)對人有高的致死率,且演化快速。本論文提供三個構面來研究 A 型(H3N2)流感病毒之基因演化與抗原性演化之關聯性。在第一構面中,針對 HA 的抗原性變異株提出一個以決策規則為主的方法用以挑選關鍵的胺基酸位置、建立規則並研究共同改變的胺基酸位置。做法是使用資訊獲得量(information gain, IG)與亂度(entropy)來量度一個胺基酸位置用於區分抗原性變異株與相似株的鑑別力高低。該規則根據紅血球凝集素的胺基酸突變來描述一株流行病毒株是否能被疫苗株產生之抗體所抑制,而共同改變的胺基酸位置常與逃脫抗體辨識以及抗原性漂變相關。

在第二構面中,本研究加入抗原與抗體交互作用的概念,並且發展了一套抗原決定位(epitope)為主的方法以鑑別抗原性漂變。首先定義一個「變異的抗原決定位」是一個具有累積構形改變且逃脫抗體辨識的抗原決定位。實驗結果顯示,兩個關鍵胺基酸位置的改變可以引起一個抗原決定位的構形改變。除此之外,鄰近受體結合區(receptor-binding site)的兩個抗原決定位(A 與 B)在逃脫抗體辨識上扮演重要的角色。通常兩個改變的抗原決定位可以造成抗原性漂變。

在第三構面中,本研究探討胺基酸位置是否具有相同抗原性影響力,並且發展了一套以貝式理論為基礎的方法用以鑑別抗原性漂變。做法是利用概率比(likelihood ratio, LR)量度每個胺基酸位置所造成的抗原性變化大小。根據單純貝式網路與概率比,此方法定義 AD_{LR} 用於量度一對紅血球凝集素序列間的抗原性距離(antigenic distance)。實驗結果顯示,位於抗原決定位與空間上鄰近受體結合區的位置對於抗原性漂變有決定性的影響。除此之外, AD_{LR} 與血球凝集抑制試驗(hemagglutination inhibition, HI)之血清測試值有高度的相關性,且可以解釋自西元 1968 年至 2008 的 A 型(H3N2)疫苗株選擇。

整體而言,此論文顯示了上述模型對於描述基因演化與抗原性演化之關聯性具有可行性與穩健度。根據 HI 之血清測試值、紅血球凝集素與抗體之結晶結構,此研究發現 A 型(H3N2)流感病毒抗原性變異株的關鍵胺基酸位置、共同演化位置、胺基酸位置規則與抗原決定位規則;更重要的是這些模型可以有效反映流感疫苗株的選擇、預測抗原性變異株及

對於抗原性漂變提供具有生物意義的新洞察角度。我們相信此研究有助於未來流感疫苗的發展與了解流感病毒的演化，並能指引如何快速研發更有效的流感疫苗。未來可能的研究方向包括研究季節性 H1N1 流行性感冒病毒以及抗原與抗體間的交互作用。



A Study of Relationships between Genetic and Antigenic Evolution of Influenza A (H3N2) Viruses

Student: Jhang-Wei Huang

Advisor: Dr. Jinn-Moon Yang

Institute of Bioinformatics and Systems Biology
National Chiao Tung University

ABSTRACT

Influenza viruses often cause significant human morbidity and mortality. Gradually accumulated mutations on the glycoprotein hemagglutinin (HA) occur immunologically distinct strains (named as antigenic variants), which lead to the antigenic drift. The emergence and spread of antigenic variants often require a new vaccine strain to be formulated before each annual epidemic. The relationship between the genetic and antigenic evolution remains unclear and to understand the relationship is an emergent issue to public health and vaccine development.

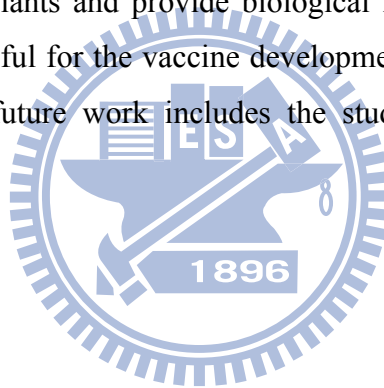
Among the influenza viruses, the influenza A (H3N2) subtype causes high mortality rates and evolves rapidly. In this thesis, we study the relationship between the genetic and antigenic evolution of influenza A (H3N2) viruses focusing on the following three dimensions. In the first dimension, we proposed a rule-based method for identifying critical amino acid positions, rules, and co-mutated positions for antigenic variants. The information gain (IG) and the entropy are used to measure the score of an amino acid position on HA for discriminating between antigenic variants and similar viruses. Based on the IG, we identified the rules describing when one (e.g. circulating) strain will not be recognized by antibodies against another (e.g. vaccine) strain. In addition, our experimental results reveal that the co-mutated positions are often related to antibody recognition and the antigenic drift.

In the second dimension, we incorporated the concept of antigen-antibody interactions and developed an epitope-based method to identify the antigenic drift of influenza A utilizing the conformation changes on antigenic sites (epitopes). A changed epitope, an antigenic site on HA with accumulated conformation changes to escape from neutralizing antibody, can be considered as a "key feature" for representing the antigenic drift. Our experimental results show that two critical position mutations can induce the conformation change of an epitope. The epitopes (A and B), which are near the receptor-binding site of HA, play key role for neutralizing antibodies. Two changed epitopes often drive the antigenic drift.

In the third dimension, we addressed the issue of whether the amino acid positions are

antigenically equivalent and developed a Bayesian method to identify the antigenic drift of influenza A by quantifying the antigenic effect of each amino acid position on HA. We utilized the likelihood ratio (LR) to quantify the antigenic distance of an amino acid position. Based on naïve Bayesian network and LR, we developed an index, AD_{LR} , to quantify the antigenic distance of a given pair of HA sequences. Our experimental results show that the positions locating on the epitopes and near the receptor-binding site are crucial to the antigenic drift. In addition, the AD_{LR} values are highly correlated to the hemagglutination inhibition (HI) assays and can explain WHO vaccine strain selection from 1968 to 2008.

In summary, this thesis demonstrates that our models are feasible and robust to describe the relationship between the genetic and antigenic evolution. According to the HI assays and HA/antibody complex structures, we statistically derived the critical amino acid positions, co-evolution positions, residue-based rules and epitope-based rules of the antigenic variants for influenza A (H3N2) viruses. More importantly, our models can reflect the WHO vaccine strain selection, predict antigenic variants and provide biological insights for the antigenic drift. We believe that our models are useful for the vaccine development and understanding the evolution of influenza A viruses. The future work includes the study of seasonal H1N1 viruses and antigen-antibody interactions.



Acknowledgements

誌 謝

交大十年，在我的求學過程中扮演關鍵的角色。很榮幸地，在這段期間我遇見了許多充滿研究熱情的師長與懷抱夢想的伙伴，在大家的陪伴下，走過了大學、碩士以及博士。此論文得以完成，要歸功於許多人的熱心幫助與建議，在此衷心地感謝大家。

首先要感謝我的指導教授楊進木老師，在這段期間的用心指導與耐心栽培，並且花了許多時間與我討論研究。此外也要感謝老師引導我進入生物資訊這個領域，更進一步展示了許多有趣的研究議題。除了在研究上提供許多指引，在做人處世上老師也給了我很多啟發，積極進取，樂於助人，每當我遇到困難或者心情低潮時，老師總是樂於伸出援手以及適時給予鼓勵，能在老師的帶領下做研究是一件很幸福的事。此外也要感謝老師在助學金上以及實驗室器材上提供豐富的資源，讓我們沒有後顧之憂地專心在研究上。

再來我要感謝台大流行病學研究所的金傳春教授啟發了我們團隊對於流感的研究，並且分享內心對於研究產生熱情的燭火，以及對於此論文提供許多建設性意見。除此之外，金教授懷抱著以公共衛生為己任的胸懷，亦是值得學生效法學習的精神之一。我也感謝高成炎教授、金傳春教授、楊進木教授、盧錦隆教授、黃憲達教授以及蔡懷寬教授在百忙中抽空擔任我的學位口試委員，並且在論文與報告內容上提出了有幫助的意見與指教。

接著我要感謝高成炎教授、盧錦隆教授、黃憲達教授以及蔡懷寬教授無償地分享個人的寶貴研究經驗，以及給予我在研究上與人生上的鼓勵。此外在就讀研究所期間，我也感謝盧錦隆教授在序列分析上的教導、黃憲達教授在資料庫上的教導以及黃鎮剛教授在分子模擬上的教導，使我獲益良多。

BioXGEM 實驗室的伙伴們，也感謝你們提供地一切幫助。俊辰學長在文稿修改與日常生活上提供了許多幫忙、其樺學長在同組的合作與討論中付出了許多心力、彥甫學長在繪圖上的巧思與研究上的討論、凱程在研究上的討論以及這九年來的無數的無償協助與聊天談心、怡馨在各方面的鼓勵與心情分享、宇書在研究上替我指出不足之處、一原在機器管理的部分為大家提供方便的計算資源、丹尼爾在英文對話與文稿修改的協助，還有已經畢業的永強、振寧、登凱分享在職場以及人生的許多心得。另外，也感謝學弟妹，彥修、敬立、峻宇、彥超、韋帆、力仁、御哲、怡瑋、伸融，在實驗室生活上提供的歡樂與幫助。

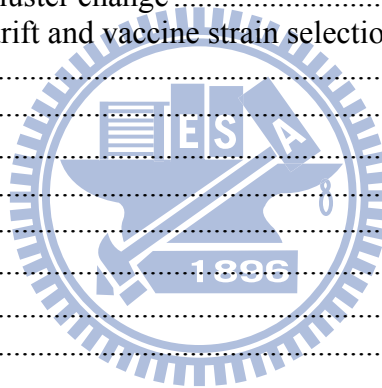
最後要感謝我的父母親黃成湖先生與廖曉翎女士的多年養育之恩，以及妹妹郁芬的不時鼓勵，因為有了你們的支持與照顧，此論文才得完成，你們是我最愛的家人，感謝你們。

除了上述的師長、伙伴與家人，還有許多貴人曾經提供幫助，請原諒我無法一一列出，我仍誠摯感謝。

Contents

摘要	i
ABSTRACT	iii
Acknowledgements	v
Contents	vi
List of figures	viii
List of tables	ix
Chapter 1 Introduction	1
1.1. Background	1
1.1.1. Influenza and its impact	1
1.1.2. Genetic and antigenic evolution of influenza viruses	1
1.1.3. Annually reviewed vaccine and vaccine strain selection	2
1.2. Previous works	3
1.3. Challenges	5
1.4. Thesis organization	5
Chapter 2 Co-evolution Positions and Rules for Antigenic Variants of Influenza A (H3N2)	
Viruses	7
2.1. Introduction	7
2.2. Motivation and aim	8
2.3. Materials and Methods	9
2.3.1. Data sets	9
2.3.2. Identifying critical positions on HA	10
2.3.3. Discovering the rules of antigenic variants	11
2.3.4. Predicting antigenic variants	11
2.3.5. Identifying co-mutated positions for antigenic variants	12
2.4. Results	12
2.4.1. Critical positions on HA	12
2.4.2. The rules of antigenic variants and predicting accuracies	17
2.4.3. Co-mutated positions for antigenic variants	19
2.5. Discussion	22
2.6. Summary	23
Chapter 3 Changed Epitopes Drive the Antigenic Drift for Influenza A (H3N2) Viruses	24
3.1. Introduction	24
3.2. Motivation and aim	24
3.3. Materials and Methods	25
3.3.1. Changed epitopes	26
3.3.2. Data sets	26
3.3.3. Identifying critical positions on HA	28
3.3.4. Models for antigenic variants based on changed epitopes	28
3.3.5. Variant ratio for measuring the antigenic drift	29
3.4. Results	30
3.4.1. Antigenic critical positions	30
3.4.2. Changed epitopes for antigenic variants	30
3.5. Discussion	39

3.6.	Summary	41
Chapter 4 A Bayesian Approach for Quantifying the Antigenic Distance of Influenza A (H3N2)		
	Viruses	42
4.1.	Introduction	42
4.2.	Motivation and aim	43
4.3.	Materials and methods	44
4.3.1.	Data sets	44
4.3.2.	Quantifying the antigenic distance of amino acid positions	49
4.3.3.	Quantifying the antigenic distance of a pair of HA sequences	49
4.3.4.	Variant ratio for studying the antigenic drift	50
4.3.5.	Shannon entropy	50
4.3.6.	Contact-pair distance on antigen-antibody interaction	51
4.3.7.	Amino acid distance to sialic acid	52
4.4.	Results	52
4.4.1.	Antigenic distance of amino acid positions	52
4.4.2.	Antigen-antibody interaction	54
4.4.3.	Antigenic distance for a pair of HA sequences	58
4.4.4.	Predicting antigenic variants	58
4.4.5.	Vaccine-vaccine transitions	58
4.4.6.	Antigenic cluster change	60
4.4.7.	Antigenic drift and vaccine strain selection	62
4.5.	Discussion	64
4.6.	Summary	65
Chapter 5 Conclusion		
5.1.	Summary	66
5.2.	Future work	67
References		
Appendix A		
	List of Publications	73

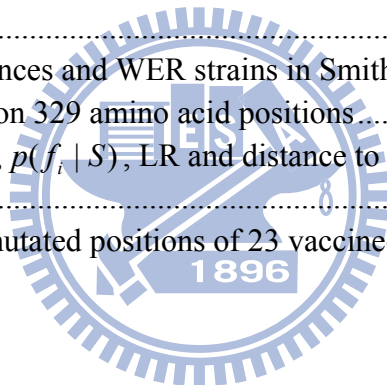


List of figures

Figure 1.1 Overview of this thesis for studying the relationships between genetic and antigenic evolution.	6
Figure 2.1 Overview of our method for predicting the antigenic variants of influenza A (H3N2) viruses.	8
Figure 2.2 The relationship between entropies and information gains of 329 amino acids on HA.	14
Figure 2.3 The distribution of IG values and co-mutation scores on HA structure.	16
Figure 2.4 The decision tree and rules for predicting antigenic variants.	17
Figure 2.5 Comparison of our method with other two methods on predicting antigenic variants on two data sets.	18
Figure 2.6 The co-mutation z-score distributions of six positions on the HA sequence.	21
Figure 3.1 Overview of our method for the antigenic drift.	25
Figure 3.2 The relationships between number of changed epitopes and antigenic variants based on four proposed models.	31
Figure 3.3 The changed-epitope composition and antigenic variants on 4 models.	32
Figure 3.4 The three HA-antibody complex structures.	34
Figure 3.5 Comparison of our method with the other two methods on predicting antigenic variants on two data sets.	35
Figure 3.6 The HA/antibody structure and interface.	36
Figure 3.7 The epitope evolution and the antigenic drift from 1982-1983 to 2008 influenza season.	37
Figure 3.8 The comparison between our method and Wilson & Cox's model in the antigenic drift from 1982-1983 to 2008 influenza season.	38
Figure 4.1 Overview of our method for quantifying the antigenic distance for amino acid positions and a pair of HA sequences.	43
Figure 4.2 The statistics of 54 antigen-antibody complex structures.	51
Figure 4.3 The frequency diagram of 10 amino acid positions on HA.	54
Figure 4.4 The relationships between LR and HA-antibody complexes.	56
Figure 4.5 The LR values distribution of Smith's 43 positions.	57
Figure 4.6 The relationships between AD_{LR} and HI assays.	57
Figure 4.7 The distribution of AD_{LR} and the antigenic drift from 1968 to 2003.	61
Figure 4.8 The comparison of vaccine strain and other strains in BK79 and SY97 cluster.	62
Figure 4.9 The distribution of AD_{LR} and the antigenic drift from 1982-1983 to 2008 influenza season.	63

List of tables

Table 2.1 The entropy, information gain, and co-mutated positions of 15 amino acid positions on HA sequences.....	15
Table 2.2 Comparison of our method with other methods for predicting the antigenic variants on 31,878 pairs.....	19
Table 2.3 The number of co-mutation positions of five epitopes and the other area on HA.....	20
Table 3.1 The number of HI assays, number of sequences in Smith's dataset and WER strains from 1968 to 2007.....	27
Table 3.2 Summary of 4 models.....	28
Table 3.3 The changed epitopes and mutations of 11 virus-pairs under 4 models.....	29
Table 3.4 The list of 64 critical positions in the five different epitopes [9, 31].....	30
Table 3.5 Example of 13 antigenic variants without changed epitopes.....	39
Table 4.1 The data sources and composition of HI assay dataset from 1968 to 2007.....	46
Table 4.2 The number of sequences and WER strains from 1982-1983 to 2008 influenza season.....	47
Table 4.3 The number of sequences and WER strains in Smith's dataset from 1968 to 2003...	48
Table 4.4 The summary of LR on 329 amino acid positions.....	53
Table 4.5 The TP, FP, $p(f_i V)$, $p(f_i S)$, LR and distance to sialic acid of 10 amino acid positions on HA.....	53
Table 4.6 The HD, AD_{LR} and mutated positions of 23 vaccine-vaccine pairs.....	59



Chapter 1

Introduction

1.1. Background

1.1.1. Influenza and its impact

Influenza is one of the most important infectious diseases occurring in humans. The annual epidemics cause an estimated 500,000 deaths in the world every year [1]. Moreover, the global pandemics can cause high mortality in humans with four pandemics occurring during the last 100 years. Among the four pandemics, the 1918 H1N1 pandemic caused about 20-50 million deaths [2]. Recently, the 2009 H1N1 pandemic that originated from swine influenza virus presented new threats to public health worldwide [3].

Influenza is a single-stranded, negative-sense RNA virus that infects humans and other animals including pigs, ferrets and many avian species. Three types (A, B and C) of influenza viruses circulate in human population. Type A virus has high genetic diversity and causes the highest rates of morbidity in humans [4]. There are eight genome segments that encode eleven proteins in the influenza A virus [5]. Among these eleven proteins, the two surface proteins hemagglutinin (HA) and neuraminidase (NA) are the main targets for the human immune system. In addition, the influenza A viruses are divided into subtypes based on major differences in HA and NA. Currently, 16 HA subtypes and 9 NA subtypes have been identified [6] and most of them are carried by the wild waterfowls [7].

1.1.2. Genetic and antigenic evolution of influenza viruses

The high genetic diversity of influenza viruses comes from error-prone RNA polymerase, high replication rates and gene segments reassortment [8]. The mutations (substitutions, deletions and insertions) are one of the most important mechanisms for generating genetic variation in influenza viruses.

1.1.2.1. Antigenic drift.

Although both HA and NA are surface proteins that are targeted by the antibodies, the HA contains the highest proportion of antigenic sites that can be recognized by the immune system [9-11]. Frequent and accumulated mutations on the influenza genome can cause conformational changes in the HA. Since the human immune system is not fully cross-protected against viral infection [12], the new mutations of HA may cause antibodies to no longer recognize the variant viruses and let the viruses to escape recognition by the immune system. This gradual change in antigenic structure with time is called antigenic drift [13]. In addition, the global influenza surveillance network regularly screens the emerging antigenic variants by hemagglutination inhibition (HI) assay [14-15], which is a binding assay representing the binding affinity between one (e.g. circulating) strain and animal antisera against another (e.g. vaccine) strain. Moreover, the HA is the primary component of current influenza vaccines [15].

1.1.2.2. Antigenic shift

When a host is co-infected by two or more different subtypes of influenza viruses, the segmented genomes between them may reassort and generate a new subtype of virus having novel mixtures of the HA and NA. This event is called antigenic shift [16-17] and the new subtype of virus often causes significant damage to humans because the human population is immunologically naive to the new virus. In the last 100 years, there were four pandemics, which originated from a reassortment among HA and NA: 1918 (H1N1 subtype) [18], 1957 (H2N2 subtype) [19], 1968 (H3N2 subtype) [20] and recent 2009 pandemic (H1N1 subtype) [7] in which the genome of virus reassorted from swine [3, 21]. Moreover, the novel influenza virus is often the ancestor of circulating influenza viruses in the following years [22].

1.1.3. Annually reviewed vaccine and vaccine strain selection

Currently, vaccination is the primary preventive measure against influenza [23-24]. The vaccines can provide effective protection when the HA between vaccine strain and circulating strains share highly similar antigenic properties [25]. The human immune system can provide lifelong immunity for the invading influenza strain with one single infection [26]; however, the variant influenza viruses undergoing antigenic drift may infect people in the coming years. To ensure

that efficacy of a vaccine is sufficient against the circulating strains. WHO established a global surveillance network to detect the emergence of novel influenza viruses [27]. Each influenza season, a panel of experts meets together to select a suitable strain from recent isolates as the vaccine strain to be used in the coming winter [14]. This method raises the problem of which of today's strain is judged to cause epidemic in the following winter [28].

Since the production of flu vaccine requires 6 or more months [29], the recommendation of vaccine strain is made about 9-12 months before the season in which the vaccine is used [22]. Because of long production time for vaccines, the mismatch between vaccine strain and circulating strains may arise when the emerging variants are not identified early enough. A good example is the mismatch between vaccine strain and pandemic strains (H1N1 subtype), which occurred in year 2009.

1.2. Previous works

One of the emergent issues of influenza viruses is the vaccine strains selection. A suitable vaccine strain can provide sufficient vaccine efficacy against the circulating strains [30]. To address this issue, many methods have been proposed to study the evolution of HA and vaccine development [15, 25, 31-32]. We divided them into several types according to the materials that they analyze.

1.2.1. Phylogenic methods

Many works focus on the genetic evolution of influenza viruses because a huge amount of sequence data is available in the public databases. Bush *et al.* proposed the first method to predict the evolution of influenza virus based on HA sequences [31]. They collected 357 HA sequences from 1983 to 1997 and constructed a phylogenetic tree. Based on the phylogenetic tree, they identified 18 codons under positive selection [33]. According to the retrospective tests, their study showed that, "Viral lineages undergoing the greatest number of mutations in the positively selected codons on phylogenic tree were the progenitors of future H3 lineages in 9 of 11 recent influenza seasons." [31] Their study demonstrated that understanding the genetic evolution of HA is helpful for the vaccine strain selection.

1.2.2. Clustering methods based on genetic data

Plotkin *et al.* proposed a clustering method to predict the future dominant HA sequences and discussed its potential relevance to vaccine strain selection [34]. Based on the HA sequence clusters, their method select the most recent sequence in the current season's most dominant cluster as the future vaccine strain. Furthermore, they studied the spatio-temporal distribution of viral swarms and compared it to the influenza vaccines recommended by the WHO. Their study demonstrated that cluster structure analysis of HA sequences is helpful for the vaccine strain selection.

1.2.3. Clustering methods based on antigenic data

The global influenza surveillance network regularly characterizes antigenic properties of circulating strains by HI assay [14-15]. Although antigenic data is one of the key criteria for vaccine strain selection, the antigenic data are largely unexplored due to difficulties in quantitative interpretation. Smith *et al.* proposed an antigenic map of influenza A virus and showed that how antigenic evolution is mapped to genetic evolution [15]. The punctuated nature of the antigenic evolution of HA has been visualized in the antigenic map. Their approach quantifies the antigenic distances among vaccine strains and circulating strains from 1968 to 2003 and therefore helps with selection of vaccine strain. One of the most important discoveries in Smith *et al.*'s work is that, "Antigenic evolution was more punctuated than genetic evolution, and genetic change sometimes had a disproportionately large antigenic effect." [15] Their study demonstrated that both genetic and antigenic data provides valuable insights for the evolution of influenza viruses.

1.2.4. Hybrid method considering genetic and antigenic data

Currently, the HI assay is the primary method to characterize the antigenic properties for circulating strains. However, the quantity of HI assay data in public databases is far less than sequence data [35-37]. Lee *et al.* proposed the first method to predict antigenic variants based on HA sequences [25]. This data set contained 181 pairs of HA sequences pairs and the results showed that the model based on 5 antigenic sites had the best accuracy for predicting variants.

1.3. Challenges

One of the key issues in the development of influenza vaccine is to improve the accuracy of vaccine strain selection: that is, to select which of today's strain is likely to be dominant in the coming year's epidemic [28]. Furthermore, a more comprehensive understanding the relationship between genetic and antigenic evolution is useful to predict the evolution of influenza virus in advance while the surveillance system is not able to detect the variants in early stages. Moreover, there are several thousands HA sequences in public database that lack antigenic information. If a method could link the genetic evolution (sequence data) to antigenic evolution (antigenic data), it could provide valuable insights for the understanding of antigenic drift and vaccine development.

1.4. Thesis organization

In this thesis, we study the relationships between genetic and antigenic evolution focusing on three dimensions and the thesis is organized as follows (Fig. 1.1). In Chapter 2, we developed a method for identifying antigenic critical amino acid positions, rules, and co-mutated positions for antigenic variants. The rules describe when one (e.g. circulating) strain will not be recognized by antibodies against another (e.g. vaccine) strain based on HA sequences. The co-mutated positions are two positions that mutate simultaneously on HA. We first identified the co-mutated positions and discussed its relatedness to the antigenic drift.

The critical positions are widely distributed on HA structure; however the antibody recognition of HA is highly correlated to the conformation changes on the antigenic sites (epitopes). In Chapter 3, we developed an antigenic site based method to identify the antigenic drift of influenza A utilizing the conformation changes on epitopes. We address two issues in this chapter: first, how to quantify the degree of conformational change in a changed epitope; second, what are the relationships between changed epitopes and antigenic drift.

From the previous two dimensions, we observed that some amino acid mutations can cause antigenic variants while other mutations have few effects for antigenic variants. In addition, we also noticed that mutations on epitope A and B seem more likely to cause antigenic variants. The above observations raise the question of whether the amino acid positions are antigenically equivalent or not. In Chapter 4, we developed a Bayesian method to identify the antigenic drift of influenza A by quantifying the antigenic effect of each amino acid position on HA. Based on

the accumulated HI assay during last 40 years, we utilized the likelihood ratio (LR) to quantify the antigenic distance of an amino acid position. We discuss the relationships between LR values of positions and antigenic drift. Moreover, we developed an index, AD_{LR} , to quantify the antigenic distance of a given pair of HA sequences based on naïve Bayesian network and LR. We evaluated AD_{LR} for predicting antigenic variants, explaining the vaccine-vaccine transitions and selection the WHO vaccines on 2,789 circulating strains. Finally, Chapter 5 presents the conclusions and the future work.

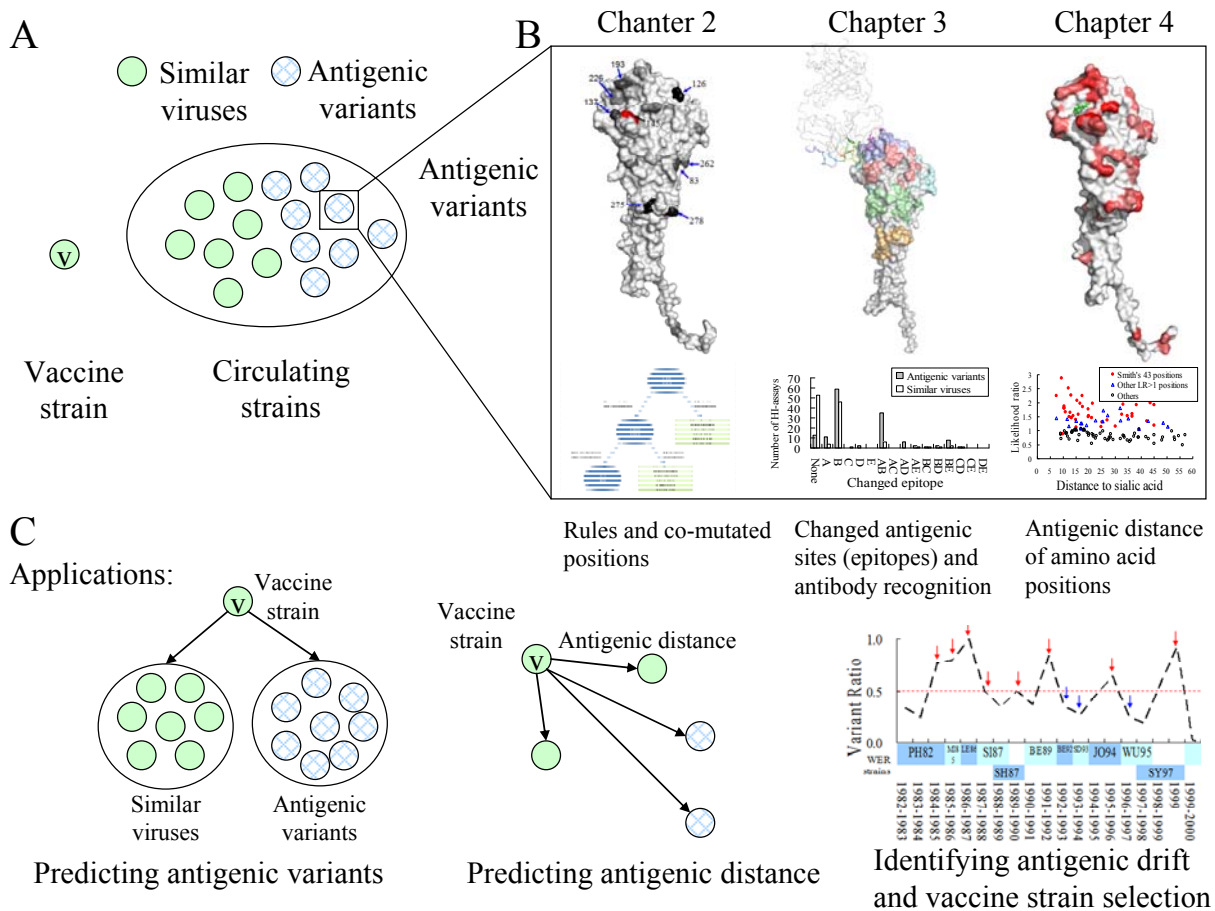


Figure 1.1 Overview of this thesis for studying the relationships between genetic and antigenic evolution. (A) The vaccine strain and circulating strains. (B) The Chapter 2, 3 and 4 in this thesis. (C) The applications for our methods.

Chapter 2

Co-evolution Positions and Rules for Antigenic Variants of Influenza A (H3N2)

Viruses

2.1. Introduction

Pathogenic avian and influenza viruses often cause significant damage to human society and economics [23]. The influenza viruses are divided into subtypes based on differences in the surface proteins HA and NA, which are the main targets for the human immune system. In circulating influenza viruses, gradually accumulated mutations on HA occur immunologically distinct strains (named as antigenic variants), which lead to antigenic drift. The antigenic drift often implies that vaccines should be updated to correspond with the dominant epidemic strains [23]. Mapping the genetic evolution to the antigenic drift of influenza viruses is one of key issues to public health. Many methods have been proposed to study the antigenic drift and vaccine development [15, 33, 38-40].

Retrospective quantitative analyses of the genetic data have revealed important insights into the evolution of influenza viruses [31, 33, 41]. In the current global influenza surveillance system, the ferret serum HI assay is the primary method to define the antigenic variants. Several studies used statistical models to predict the antigenic variant of a given pair of HA sequences based on these known HI assays and their respective HA sequences [15, 40]. Furthermore, Smith *et al.* demonstrated that the antigenic evolution was more punctuated than the genetic evolution [15], and the genetic change sometimes has a disproportionately large antigenic effect. Recently, few studies discuss the relationship between evolution and co-mutated positions on influenza virus [39, 42].

2.2. Motivation and aim

The current trivalent vaccine contains seasonal H1N1, H3N2 and influenza B virus strains [23]. Among the influenza viruses, the H3N2 subtype causes higher mortality [43] and evolves more rapidly [44]. In addition to all of the above, the large amount of genetic and antigenic data for H3N2 virus provides valuable opportunity for us to understand the relationships between genetic and antigenic evolution of influenza A viruses.

Here, we proposed a method to predict the antigenic variants of A (H3N2) viruses by identifying critical positions and rules which describe when one (e.g. circulating) strain will not be recognized by antibodies against another (e.g. vaccine) strain. Our method is also able to detect the co-mutated positions for predicting the antigenic variants. These critical positions and rules were evaluated on two datasets which consist of 181 and 31,878 pairs, respectively. The results demonstrate that our model is able to reflect the biological meanings and achieve high prediction accuracy.

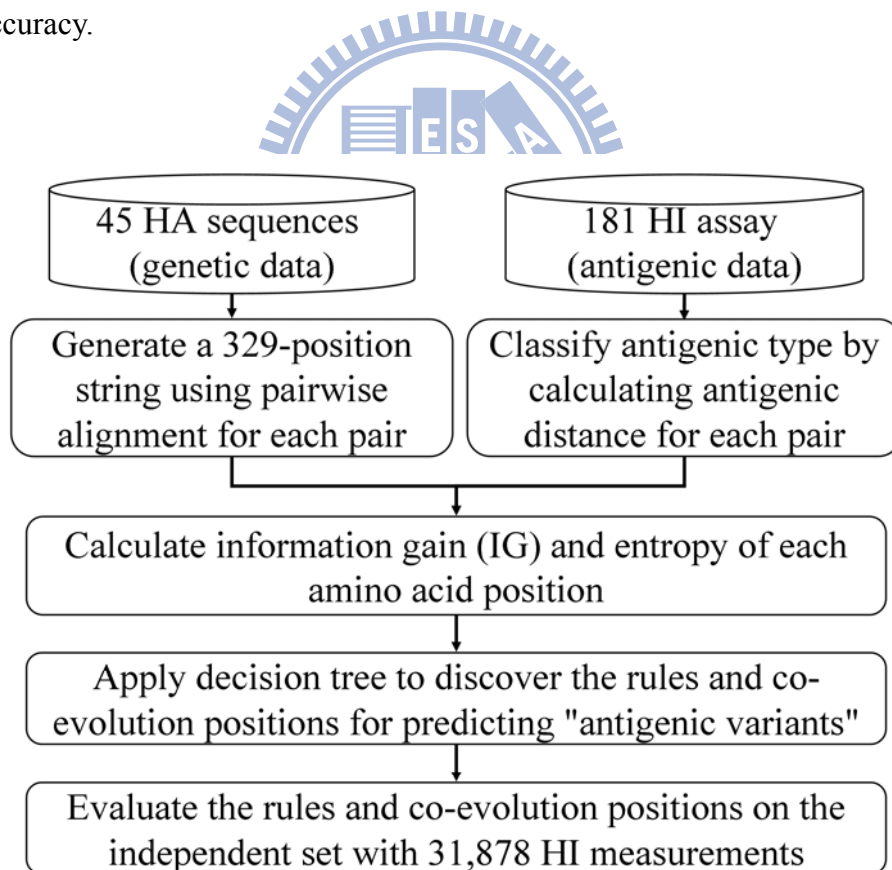


Figure 2.1 Overview of our method for predicting the antigenic variants of influenza A (H3N2) viruses.

2.3. Materials and Methods

Figure 2.1 shows the overview of our method for predicting the antigenic variants of influenza A (H3N2) viruses by identifying critical positions, rules and their co-evolution on the HA.

2.3.1. Data sets

We collected an HI assay data set, which contains 181 pairs of HA sequences with 45 HA (H3N2 viruses) sequences having 329 amino acids collected during the period, 1971 to 2002, from related work [40]. According to this data set, we applied the decision tree C4.5 [45] to predict the antigenic variants by identifying critical positions as well as discovering the rules and co-mutated positions. In this data set, the main samples (65%, 122 pairs among 181 pairs) consist of pairs of vaccine-circulating strains, and for each pair it is known whether there is inhibition of the circulating strain by antibodies against the vaccine strain ("antigenic variants" and "similar viruses"). Vaccine strains are selected by World Health Organization (WHO) and are often the dominant strains of influenza seasons. Each pair includes the HI assay value (i.e. antigenic distance) and a bit string of 329 binary bits by aligning a pair of HA sequences (329 amino acids). For a specific position on a pair of HA sequences, the binary value is "1 (named as mutation)" if the residue types of the two sequences on this position are different; conversely, its binary value is "0 (named as no mutation)". In general, an influenza vaccine should be updated if an HI assay value is more than 4.0 between the current vaccine strain and the strains expected to circulate in next season [15]. The antigenic distance is defined as the reciprocal of the geometric mean of two ratios between the heterologous and homologous antibody titers [40]. Among 181 pairs of HA sequences, 125 pairs with antigenic distance ≥ 4 are considered as "antigenic variants" and 56 pairs with antigenic distance < 4 are classified as "similar viruses". For example, the antigenic distance of the pair of HA sequences, A/Port_Chalmers/1/73 and A/Victoria/3/75, is 16 and this pair is considered as "antigenic variants". Conversely, the antigenic distance of the pair of HA sequences, A/Wuhan/359/95 and A/Nanchang/933/95, is 1 and this pair is considered as "similar viruses".

Furthermore, we prepared another HI assay data set proposed by Smith *et al.* to independently evaluate our model and compare with other methods for predicting the antigenic variants [15]. This data set consists of 253 H3N2 viruses which are clustered into 11 antigenic groups. We assume that a virus-pair in the same antigenic group is considered as a "similar

viruses" pair and a virus-pair in different groups is considered as a "antigenic variants" pair. Finally, we obtained 31,878 HI measurements and these sequences were extracted from supporting materials of publication [15].

2.3.2. Identifying critical positions on HA

In this study, positions with a both highly antigenic discriminating score and highly genetic diversity are considered as critical positions. We first evaluate the genetic diversity, which commonly believed, relates to immune selection [33], of each amino acid position on HA. Here, Shannon entropy was used to measure the genetic diversity of an amino acid position i ($i=1\sim 329$) with 20 amino acid types and is defined as

$$H(i) = -\sum_{T=1}^{20} P(A_i = T) \log(P(A_i = T)) \quad (1)$$

where $P(A_i=T)$ is the probability of the position i with amino acid type T . The information gain [45] measures the score of an amino acid position on HA for discriminating between antigenic variants and similar viruses. An amino acid with high IG at a specific position implies that a mutation on this position is highly correlated to antigenic variants. The IG of the position i associates to antigenic type Y (i.e. antigenic variants (V) and similar viruses (S)) is defined as

$$IG(i, Y) = H(Y) - H(Y | i) \quad (2)$$

$H(Y)$ is the entropy of antigenic type Y and is defined as

$$H(Y) = -\sum_{T \in \{V, S\}} P(Y = T) \log(P(Y = T)) \quad (3)$$

$H(Y | i)$ is the conditional entropy of Y when given the position i . Two states of the position i are mutation (M) and non-mutation (N). $H(Y | i)$ is defined as

$$H(Y | i) = -\sum_{K \in \{M, N\}} P(A_i = K) H(Y | A_i = K) \quad (4)$$

$P(A_i=K)$ is the probability of the position i in state K . $H(Y|A_i=K)$ is the entropy of antigenic type Y when given the position i in state K . $H(Y|A_i=K)$ is given as

$$H(Y | A_i = K) = - \sum_{T \in \{V, S\}} P(Y = T | A_i = K) \log(P(Y = T | A_i = K)) \quad (5)$$

For example, for the position 145, the numbers of the "mutation" and "non-mutation" are 62 and 119, respectively, among 181 pair-wise HA sequences in the training data set. For 62 mutation pairs, the numbers of "antigenic variants" and "similar viruses" are 61 and 1, respectively. The numbers of "antigenic variants" and "similar viruses" are 55 and 64, respectively, for 119 non-mutation pairs. According to these data, we can calculate that $P(A_{145}=M)$ is 0.34 and $H(Y|A_{145}=M)$ is 0.12 for the mutation state; $P(A_{145}=N)$ is 0.66 and $H(Y|A_{145}=N)$ is 1.0 for the non-mutation state. Finally, we obtained $H(Y | i)=0.70$. The values of information gain and entropy of 329 HA positions are normalized in the range from 0 to 1.

2.3.3. Discovering the rules of antigenic variants

After identifying critical positions, we discovered the rules for predicting antigenic variants by applying the decision tree C4.5 [45]. These antigenic amino acid positions are considered as the attributors (features). An amino acid position with high IG was selected as an internal node in the tree to discriminate "antigenic variants" and "similar viruses". According to the selected positions and constructed tree, we can easily identify the rules according to the paths from the root to the leaves of the tree.

2.3.4. Predicting antigenic variants

In order to evaluate and compare our model with other methods [9, 40] for predicting antigenic variants, we collected two data sets. The first data set consists of 181 pair-wise HI measurements and the second independent data set contains 31,878 HI measurements proposed by Smith *et al.* [15]. Wilson & Cox [9] suggested that a drift viral variant of epidemiologic importance usually contains more than 4 residues changes located on at least 2 of the five epitopes on the HA. Lee & Chen [40] proposed a model based on the hamming distance (HD) of 131 positions on all the five epitopes of HA to predict antigenic variants. Their model predicted a pair of HA sequences as the antigenic variants if there are more than 6 amino acid mutations between this pair of HA sequences.

2.3.5. Identifying co-mutated positions for antigenic variants

Here, we used the decision tree hierarchy to identify co-mutation of two amino acid positions. In order to identify all potential co-mutated pairs on HA, the positions (i.e. 101 positions among 329 positions), which occur mutations in 181 pairs of HA sequences, are sequentially selected to identify its co-mutated positions. Based on these 101 positions, the total number of two-position combinations is 10,100. For each amino acid positions (i), the co-mutation score ($S(i,j)$) between the position i and its partner position j is defined as

$$S(i, j) = IG_W(j, Y) - IG_{R_i}(j, Y) \quad (6)$$

where $IG_W(j, Y)$ is the IG value, which is derived from the whole data set (i.e. 181 pairs of HA sequences in the training set) using Equation (2), of the position j ; $IG_{R_i}(j, Y)$ is the IG value of the position j derived from the data set R by removing the pairs, in which the position i is mutated, from the whole data set. The z-score of the $S(i,j)$ of a pair of co-mutated positions is derived from 10,100 pairs and it is defined as

$$Z(i, j) = \frac{S(i, j) - \mu}{\sigma} \quad (7)$$

μ and σ are the mean and standard deviation of all 10,100 position pairs. For example, position 145 (IG is 1.0) is selected as the first node in the tree. Among 181 pairs, 62 pairs are mutated on the position 145. The amino acid positions are considered as co-mutated positions of the position 145 if their IG values significantly decrease after these 62 pairs are removed from the data set. For example, the z-score of the $S(145, 137)$ of the pair-positions 145 and 137 is 3.58

2.4. Results

2.4.1. Critical positions on HA

In this study, we used the information gain (IG) and Shannon entropy to measure the scores of an amino acid, which is located at a specific position on HA, for discriminating antigenic variants and similar viruses. The highest and lowest values of both IG and entropy are 1 and 0, respectively. An amino acid with high IG at a specific position implied that this position is highly correlated to the antigenic variants. An amino acid with high entropy means that this position is

often mutated in the data set. [Figure 2.2](#) shows the relationship of IG values and entropies of HA positions. The summary of some amino acid positions are listed in [Table 2.1](#). Of the 329 amino acids of HA, 131 positions are considered to lie in or near the five antibody combining sites (named as epitopes) which are labeled A through E [9]. The first rank (i.e. position 145-A) locates at the epitope A of HA. Its IG and entropy are 1.0 and 0.87, respectively. Among 181 pairs of HA sequences in the training set, the position 145-A mutates on 62 pairs and 61 pairs are the antigenic variants. This result implies that a mutation on this position highly induces an antigenic drift. This observation is consistent to previous results [15], that is, the single amino acid substitution N145K can be responsible for antigenic cluster transition. We observed that the other positions with high IG values obtained the similar behaviors.

The relationship between IG values and entropies of 101 positions in HA is shown in [Fig. 2.2](#) by excluding 228 positions which have zero for both IG and entropy. All positions can be classified into four groups according to the values of IG (antigenic degree) and entropy (i.e. genetic diversity). Those 19 positions with high IG and high entropy (i.e. Area I) are considered as critical positions in this work. According to the HA structure obtained from protein data bank (PDB code 1HGF [46]), 18 of the positions locate at all the five epitopes and 15 of them are on the surface ([Fig. 2.3](#)) by using PyMOL [47]. The positions in Area II (i.e. high entropy and low antigenic degree) imply that high genetic diversity may infer low antigenic discriminating score. For example, the positions (e.g. 226-D, 135-A, 121-D, 142-A and 186-B) have high entropies and low IG values ([Table 2.1 and Fig. 2.2](#)). Among 181 pairs of HA sequences, the position 226-D mutates on 61 pairs and 34 of these pairs are the antigenic variant. A low IG position indicates that a mutation on this position less preferred to be an antigenic variant. Our method can avoid the disadvantage of considering only the genetic data, which was widely used in previous works.

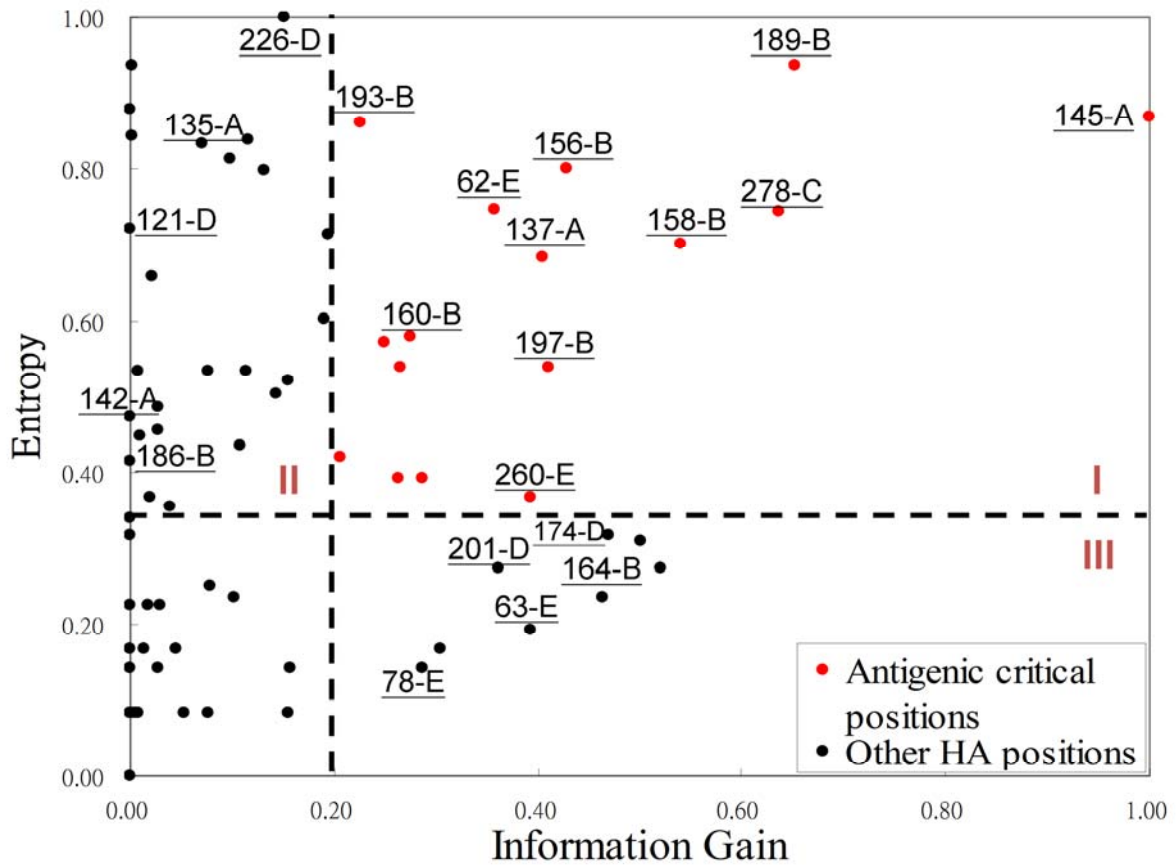


Figure 2.2 The relationship between entropies and information gains of 329 amino acids on HA. The positions in area I (e.g. 145-A, 189-B and 278-C) with both high entropy and high IG values are highly correlated to the antigenic variants. 145-A denotes the amino acid position 145 located at the epitope A.

Table 2.1 The entropy, information gain, and co-mutated positions of 15 amino acid positions on HA sequences

Position-epitope	Entropy	IG	Number of co-mutate positions	Co-mutated positions	Positive selection	Cluster Transition
145-A ¹	0.87	1.00	12	9,31,63,78,83,126,137,160,193,197,242,278	+	+
137-A	0.68	0.41	23	9,31,53,54,62,63,83,126,143,145,146,158,160,164,174,189,193,201,213,217,244,260,278		+
193-B	0.86	0.23	17	9,31,63,78,83,126,137,145,158,160,164,174,201,217,242,260,278	+	+
160-B	0.58	0.28	16	2,31,54,62,126,137,143,146,156,158,164,197,217,244,260,278		+
156-B	0.80	0.43	8	54,62,143,146,160,197,244,260	+	+
226-D	1.00	0.15	2	145,189	+	
135-A	0.83	0.07	1	165	+	
121-D	0.72	0.00	0		+	
142-A	0.47	0.00	0		+	
186-B	0.41	0.00	0		+	
164-B	0.24	0.46	6	126,137,158,174,201,217,		+
201-D	0.27	0.36	4	137,164,174,217		+
78-E	0.14	0.29	4	31,63,126,242		
174-D	0.32	0.47	4	137,164,201,217		+
63-E	0.19	0.39	6	78,83,126,137,242,278		

¹ The epitope of the position on HA sequence.

² the position is under positive selection defined by Bush *et al.* [33].

³ the position is a cluster-difference substitution defined by Smith *et al.* [15].

The relationship between IG values and structural locations of 329 positions is shown in Fig. 2.3A. The positions with four highest IG values (i.e. 145-A, 189-B, 278-C, and 158-B) are blue and other positions are near to gray based on the IG values. The positions with high IG values are located on the protein surface. Three (145-A 189-B and 158-B) of top four IG-value positions are located around the receptor-binding site, which is the key for neutralizing influenza virus. In addition, the high IG positions also prefer to locate on the top head, which are more exposed and preferable recognized by antibodies, of HA and on the interface between HA monomers.

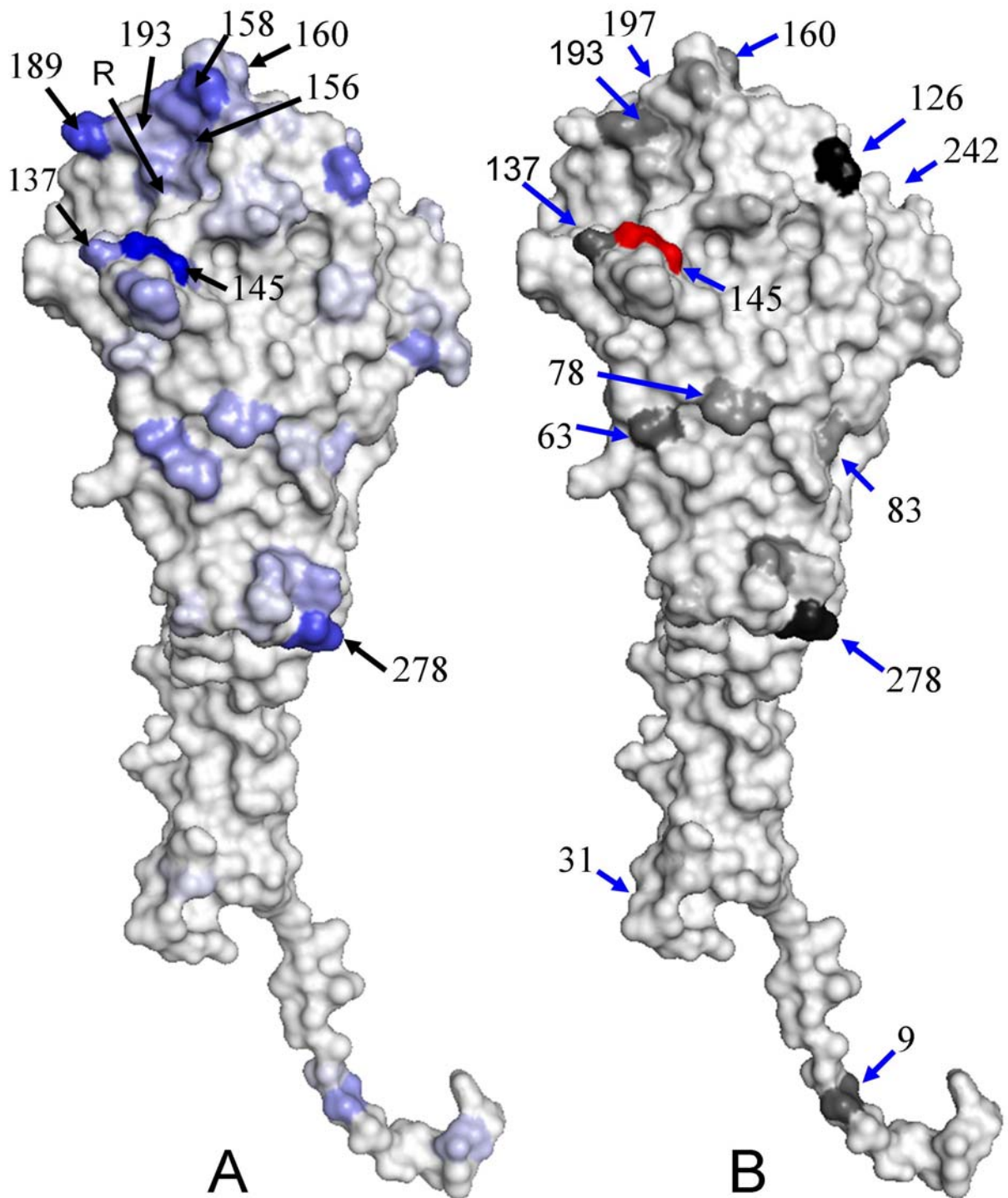


Figure 2.3 The distribution of IG values and co-mutation scores on HA structure. (A) The distribution of IG values of 329 amino acids on HA structure (PDB code 1HGF [46]) and the R indicates the receptor-binding site. The blue and gray indicate the highest IG value and the lowest IG value, respectively. (B) The structural locations and scores of 12 co-mutation positions of the position 145. These structures are presented by using PyMOL [47].

2.4.2. The rules of antigenic variants and predicting accuracies

We used the decision tree (Fig. 2.4) to build a model for predicting antigenic variants of influenza A (H3N2) virus. Based on the IG values of 329 amino acid positions derived from 181 pairs in training data set, six amino acid positions are selected as internal nodes in this tree. The first rule of this tree is that the antigenic type is predicted as the antigenic variant if the position 145 is mutated, that is, the residue types of a pair of sequences on the position 145 are different. Among 181 pairs of sequences in the training set, 62 pairs can apply this rule and 61 pairs can be predicted correctly. The last rule of this tree is that the antigenic type is predicted as the similar viruses if six positions (i.e. 145, 189, 62, 155, 213, and 214) are not mutated.

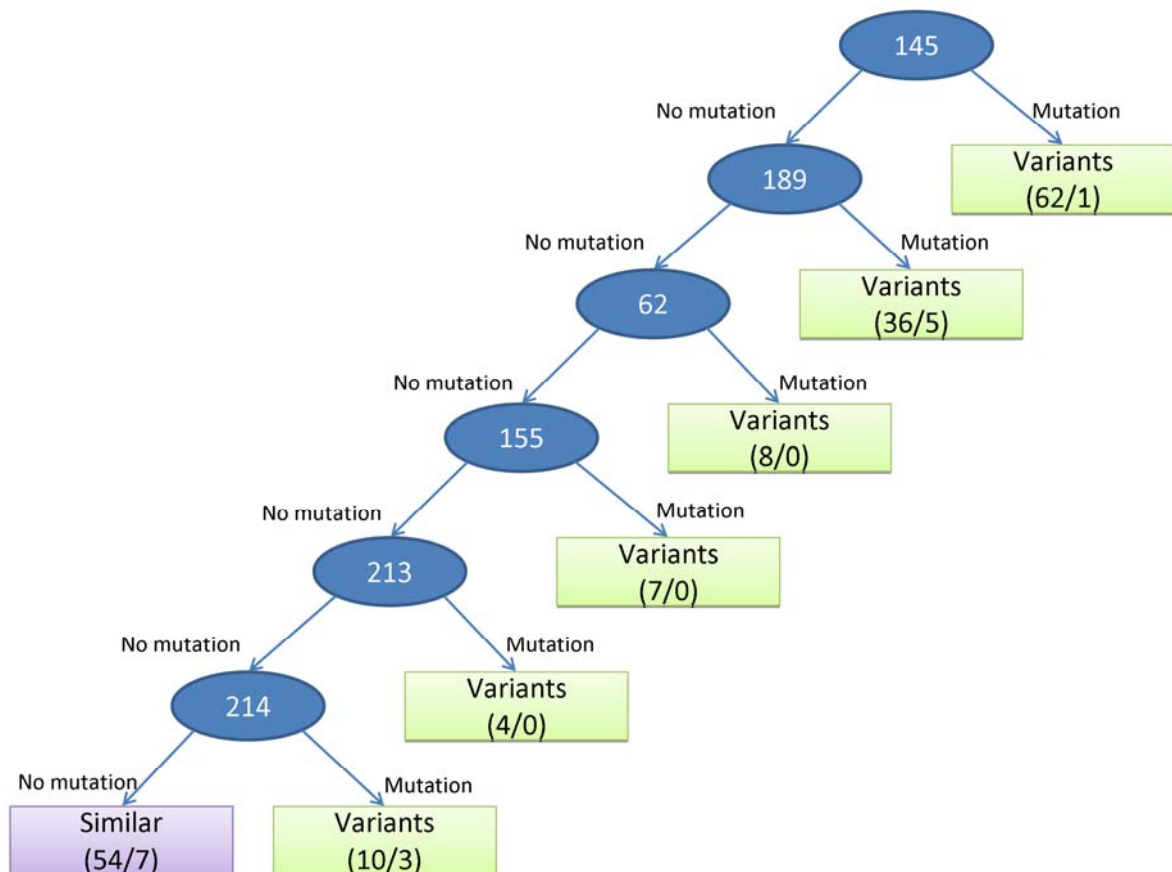


Figure 2.4 The decision tree and rules for predicting antigenic variants. Each internal node (circle) is represented as an amino acid position. The leaf node (square) includes the predicted antigenic type (i.e. "antigenic variants" and "similar viruses"), the numbers of total pairs (the first value) and predicted error pairs (the second value) by applying this rule in this node.

Based on this model, we can derive seven rules and the predicted accuracies are 91.2% (165/181) for training data set and 96.2% (30,675/31,878) for independent data set, respectively. As shown in Fig. 2.5 and Table 2.2, our method outperformed two comparative methods, i.e. Wilson & Cox (89.7%) [9] and Lee & Chen (92.4%) [40], on the independent data set. For the independent data set, the accuracies of Wilson & Cox method on predicting the antigenic variants and the similar viruses are 99.71% and 32.74%, respectively. Conversely, our model performed well for predicting the antigenic variants (99.73%) and the similar viruses (76.34%).

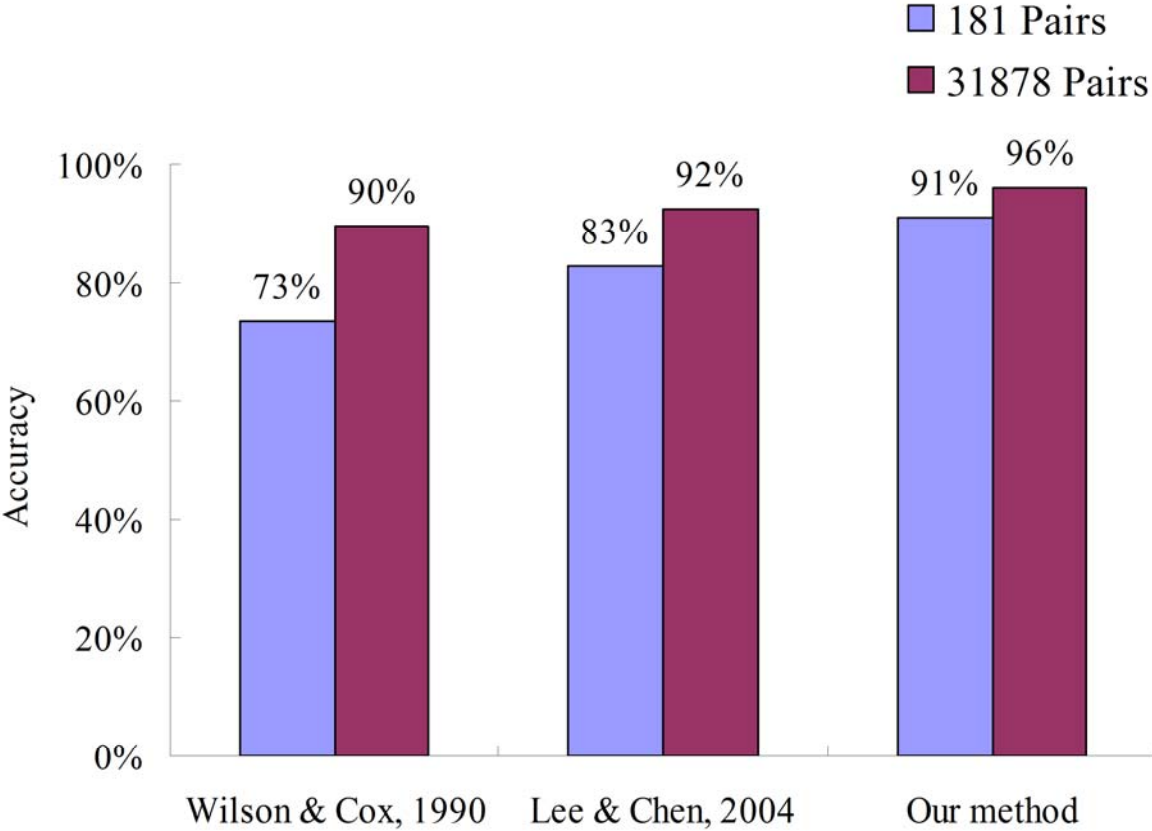


Figure 2.5 Comparison of our method with other two methods (Wilson & Cox [9]; Lee and Chen [25]) on predicting antigenic variants on two data sets

Table 2.2 Comparison of our method with other methods for predicting the antigenic variants on 31,878 pairs

Antigenic variants	Wilson & Cox, 1990 [9]	Lee & Chen, 2004 [40]	Our method	Similar viruses	Wilson & Cox, 1990 [9]	Lee & Chen, 2004 [40]	Our method
HK68-EN72 (210 ¹)	210	206	210	HK68 (91 ¹)	24	52	37
EN72-VI75 (135)	135	135	135	EN72 (105)	36	79	48
VI75-TX77 (27)	27	27	27	VI75 (36)	30	36	21
TX77-BA79 (48)	48	48	45	TX77 (3)	1	2	1
BA79-SI87 (400)	400	381	400	BA79 (120)	13	46	58
SI87-BE89 (1600)	1577	863	1600	SI87 (300)	125	233	276
BE89-BE92 (3648)	3648	3648	3648	BE89 (2016)	872	1725	2016
BE92-WU95 (1596)	1542	1391	1562	BE92 (1596)	372	928	732
WU95-SY97 (448)	448	448	448	WU95 (378)	53	156	325
SY97-FU02 (96)	96	96	96	SY97 (120)	24	65	120
Other inter clusters (18890)	18889	18870	18855	FU02 (15)	15	15	15
Number of predicted pairs	27020	26113	27026	Number of predicted pairs	1565	3337	3649
Accuracy	99.71%	96.37%	99.73%	Accuracy	32.74%	69.81%	76.34%

¹ the number of the pairs in the cluster.

2.4.3. Co-mutated positions for antigenic variants

Two amino acid positions may mutate simultaneously to cause antigenic drift or highly co-evolution in H3N2 virus. Understanding the co-mutation of amino acid position-pairs is one of the key steps to recognizing the antigen-antibody interactions. Here, we used the co-mutation score, $S(i,j)$, between the position i and its co-mutated position j to measure the co-mutated pair (i,j) for predicting the antigenic variants. We calculated all of the co-mutated combinations (i.e. 10,100 pairs) of 101 amino acid positions which mutated more than once on 181 pairs of HA sequences in the training data set.

Table 1 show the co-mutated positions of some HA positions. In this work, the position (j) is considered as the co-mutation position of the position (i) when its co-mutation z-score (i.e. $Z(i,j)$ defined as Equation (7)) is more than 2.3 because the score of the position i and j is significant (p-value is 0.01) derived from 10,100 pairs. Among 329 positions of HA sequences, 40 positions have co-mutated positions. The number of co-mutated positions for a position ranges from 0 to 23 and the total number of the significant pairs are 308 among 10,100 pairs.

In the tree model (Fig. 2.4), the position 145-A is selected as first node and has 12 significant co-mutated positions (Table 2.1 and Fig. 2.3B). The top three significant co-mutated positions of

145-A are (145-A, 126-A), (145-A, 278-C) and (145-A, 137-A). The 145-A, 278-C, and 137-A are the residues to cause the transition from cluster EN72 into cluster VI75 [15]. In addition to position 145-A, the residue 156-B has 8 significant co-mutated positions (Table 2.1). Seven (except position 260-E) of these 8 positions co-mutate with 156-B to cause the transition from the cluster TX77 into the cluster BK79 [15].

Table 2.3 The number of co-mutation positions of five epitopes and the other area on HA

	Epitope A	Epitope B	Epitope C	Epitope D	Epitope E	Other area	sum
Epitope A	15	24	8	11	16	8	82
Epitope B	19	15	6	13	13	5	71
Epitope C	15	11	3	5	9	4	47
Epitope D	12	13	3	8	6	4	46
Epitope E	13	11	4	6	7	3	44
Other area	4	2	1	3	4	4	18

Table 2.3 shows the numbers of significant co-mutation positions on six blocks, including five epitopes and the other area on the HA. The numbers (24 and 19 pairs, respectively) of the co-mutation pairs, which located at epitopes A and B, are significantly higher than other block. This result implies that the mutation on epitopes A and B could yield a high probability to cause the antigenic drift. Moreover, residues in epitopes A and B form 82 and 71, respectively, significant co-mutation pairs which are much higher than other blocks. On the other hand, the number (i.e. 18 pairs) of significant co-mutation pairs formed by the residues in non-epitope block is the smallest among 36 combinations of six blocks (Table 2.3). These observations demonstrate that epitopes A and B are more important than other blocks and the five epitopes are more important than the other area. Previous works shows that epitopes A and B are more antigenic important since they are around the receptor-binding site [9].

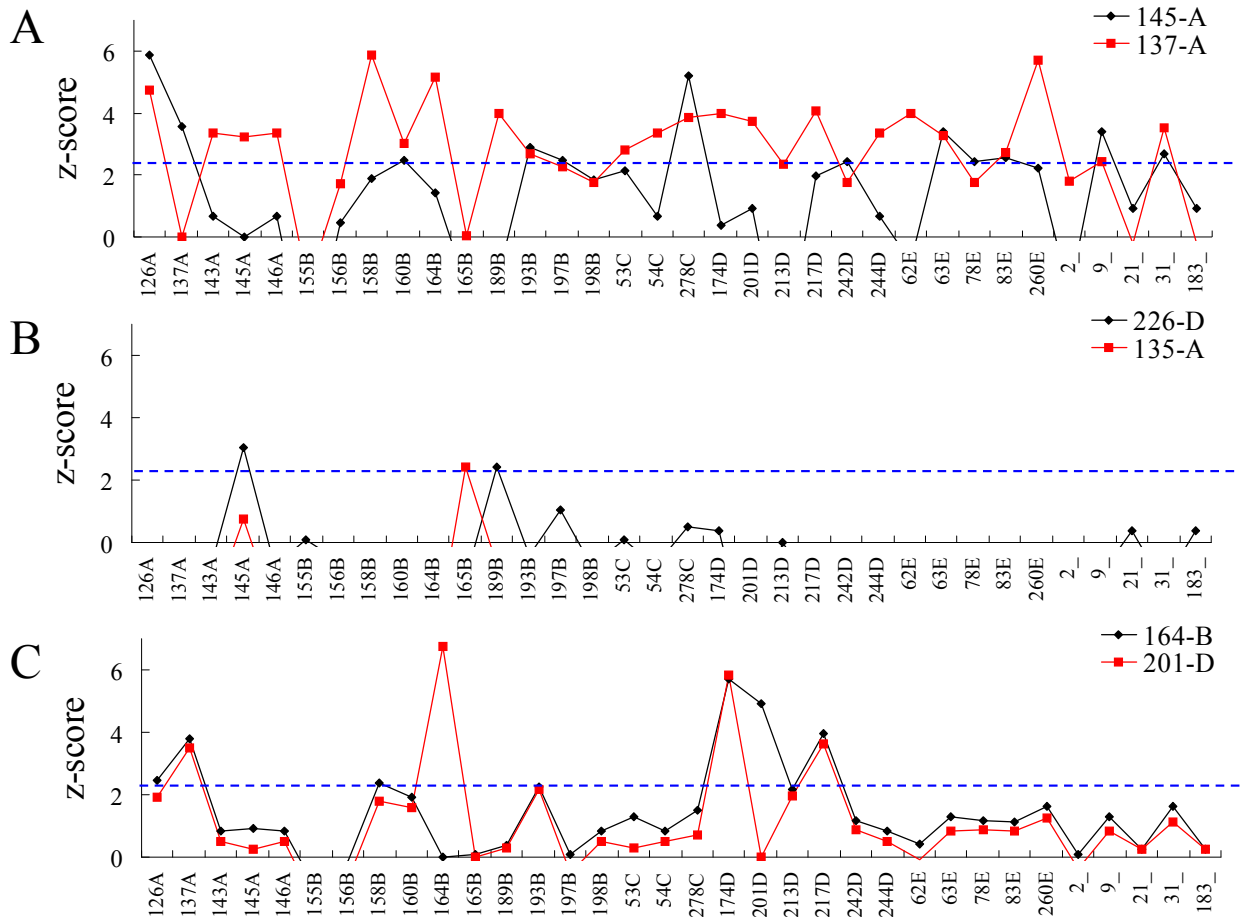


Figure 2.6 The co-mutation z-score distributions of six positions on the HA sequence. A position is considered as a co-evolution residue if its z-score is more than 2.3 (i.e. the blue line).

Figure 2.6 shows the distributions of co-mutation z-scores of six HA positions. The positions (i.e. 145-A and 137-A located in Area I in Fig. 2.2) which have high IG values and high entropies, own 12 and 23 co-mutated positions (Fig. 2.6A and Table 2.1), respectively. On the other hand, Figure 2.6B shows two positions (i.e. 226-D and 135-A located in Area II in Fig. 2.2), which have low IG values and high entropies, own 2 and 1 co-mutated positions (Table 2.1), respectively. Finally, the positions 164-B and 201-D have similar distributions (Fig. 2.6C) and their correlation coefficient is 0.73. To consider both IG values and entropies provide insight to the antigenic drift and co-evolution positions on influenza virus. These observations show our method is able to identify co-mutated positions that participate in the antigenic drift for influenza seasons. These significant co-mutated positions show biological meaning.

2.5. Discussion

Previous works using genetic data for identifying highly diverse positions which are exposed to immune selection have shown prospective [33]. However, Smith *et al.* demonstrated that antigenic evolution is more punctuated than genetic evolution [15], which implies that only genetic data may not be enough to detect critical positions. For example, the antigenic discriminating score (i.e. $IG=0.15$) of position 226-D is low, while its genetic diversity (i.e. entropy is 1.0) is largest. The position 226-D is also selected as a positive selection codon [33]. According to 181 pairs in the training set, position 226-D has 61 mutations, but 27 of them are "similar viruses" pairs. Therefore, its antigenic discriminating score is low and a mutation on this position does not cause the antigenic drift. The position 121-D, which is under positive selection, has similar behaviors.

Although the HI assay can successfully detect antigenic drift, this assay is labour-intensive and time-consuming. Therefore, the quantity of HI data is far less than sequence data and sometimes the problem of bias sampling is encountered [34]. The position 164-B, which was identified by Smith *et al.* as a cluster-difference substitution from 253 sequences [15], has 28 mutations in 181 pairs and all of them happen in "antigenic variants" pairs. Mutations on this position (i.e. IG is 0.46) have high preference to antigenic variant. But our method didn't select this position because the genetic diversity (i.e. entropy is 0.24) of this position is not high enough.

In the independent data set (31,878 pairs), the accuracies of three methods are more than 96% for the "antigenic variants", but their accuracies on the "similar viruses" pairs are significantly different (Table 2.2). The method proposed by Wilson and Cox [9] falsely predicts 67% of "similar viruses" pairs, which implies this method is very sensitive in the same antigenic group. Comparing our model with the hamming distance (HD) model which is based on epitope positions proposed by Lee & Chen [40], our model has higher accuracies in three groups, i.e. BE89, WU95 and SY97 (Table 2.2). For example, for 2016 "similar viruses" pairs in the BE89 group, the HD model falsely predicted 291 pairs, which are correctly predicted by our model, and the average HD of these 291 pairs is 7.3. Most of these 291 pairs mutate on seven positions (i.e. 50-D, 80-E, 137-A, 159-B, 167-D, 173-D and 197-B). Except positions 137-A and 197-B, the other five positions have low antigenic discriminating scores based on our model.

For each of the position-pairs (i,j) and (j,i) in Fig. 2.6, their z-scores are different because the

position i and j have different antigenic discriminating scores. For example, the z-scores of position-pairs (133-A, 156-B) and (156-B, 133-A) are 5.03 and -1.13, respectively. Furthermore, the IG values of positions 156-B and 133-A are 0.43 and 0.11, respectively. The antigenic effect of the only mutation on position 133-A is not significant. On the other hand, the antigenic discriminating score is significant when position 133-A co-mutates with position 156-B. Among 181 pairs in the training set, position 133-A has 38 mutations; 32 of them are "antigenic variants" pairs, and 31 pairs of them co-mutate with the position 156-B. This position pair is observed to cause the transitions from cluster TX77 into cluster BK79 and from the cluster BE89 into the cluster BE92 [15].

Among 329 positions of HA sequences, 137-A, 193-B, and 160-B are top three positions with the highest numbers of co-mutated positions. The position 137-A has 23 co-mutation positions and the top three pairs are (137-A, 158-B), (137-A, 260-E) and (137A, 164-B). These four positions are observed to cause the transitions from cluster EN72 into cluster VI75 and from cluster VI75 into cluster TX77 [15].

There are total 308 significant position-pairs but only 142 pairs of them are observed in cluster-difference substitutions [15]. For example, 15 pairs with top 50 z-scores not identified as cluster-difference substitution are: (83-E, 126-A), (145-A,126-A), (193-B, 126-A), (126-A, 63-E), (278-C, 126-A), (63-E, 126-A), (137-A, 126-A) (83-E, 278-C), (193-B, 63-E), (31,9), (83-E, 63-E), (126-A, 278-C), (9,31), (275-C, 145-A) and (126-A, 145-A). Nine pairs of them could be observed in the 1976 fixation [41] in which they analyzed large amount of HA protein sequences (2248 sequences from 1968 to 2005). These observations imply our method is able to detect potential co-mutated positions related to antigenic drift from limited HI-data.

2.6. Summary

This study demonstrates our model is robust and feasible by considering both genetic and antigenic data. Based on decision tree, our method is able to identify critical amino acid positions of HA and the rules of antigenic variants for influenza H3N2 viruses. The accuracies of our method are 91.2% and 96.2% for the training set and independent data set, respectively, and our method is significantly better than the other two methods being compared on these two sets. The identified critical amino acid positions are similar to related works and the co-mutated positions are able to reflect the biological meanings. We believe that our method is useful for vaccine development and understanding the evolution of influenza A viruses.

Chapter 3

Changed Epitopes Drive the Antigenic Drift for Influenza A (H3N2) Viruses

3.1. Introduction

Influenza spreads around the world and causes significant morbidity and mortality [14]. The surface proteins HA and NA are the primary targets of the protective immune system. In circulating influenza viruses, gradually accumulated mutations on the HA, which interacts with infectivity-neutralizing antibodies, lead to the escape of immune system.

Most of methods measuring the antigenic variances on HA focused on amino acid position mutations, such as hamming distance [34] or phylogenic distance [31]. Recently, few studies discuss the relationships between the antigenic sites (epitopes) and vaccine efficiency [48].

3.2. Motivation and aim

We have identified critical positions and rules for antigenic variants in previous chapter. However, the critical positions are widely distributed on HA structure and the antibody recognition is highly correlated to the conformation change on the epitopes, which locate on HA surface. Moreover, an antibody often utilizes complementarily-determining regions (CDRs) to bind two epitopes on the antigen (HA) [49]. To quantify a changed epitope for escaping from neutralizing antibodies is the basis for the antigenic drift and vaccine development.

Here, we have proposed a method to identify the antigenic drift of influenza A by quantifying the conformation change of an epitope. Our method is able to predict antigenic variants of a given pair of HA sequences which are often a vaccine strain and a circulating strain. Our model was evaluated to measure the antigenic drifts and vaccine updates on 2,789 circulating strains

(from year 1983 to 2008) and to predict the antigenic variants on two data sets (i.e. 343 and 31,878 HI assays). These observations demonstrate that our model is able to reflect the biological meanings and can explain the WHO vaccine strain selection.

3.3. Materials and Methods

Figure 3.1 presents the overview of our method for the antigenic drift of influenza A (H3N2) viruses by quantifying changed epitopes. We first identified the critical amino acid positions based on both the antigenic variant and genetic diversity. We then measured a changed epitope by calculating the accumulated conformation change based on critical amino acid mutations on an epitope. Finally, we evaluated our model for predicting antigenic variants and selecting the WHO vaccines.

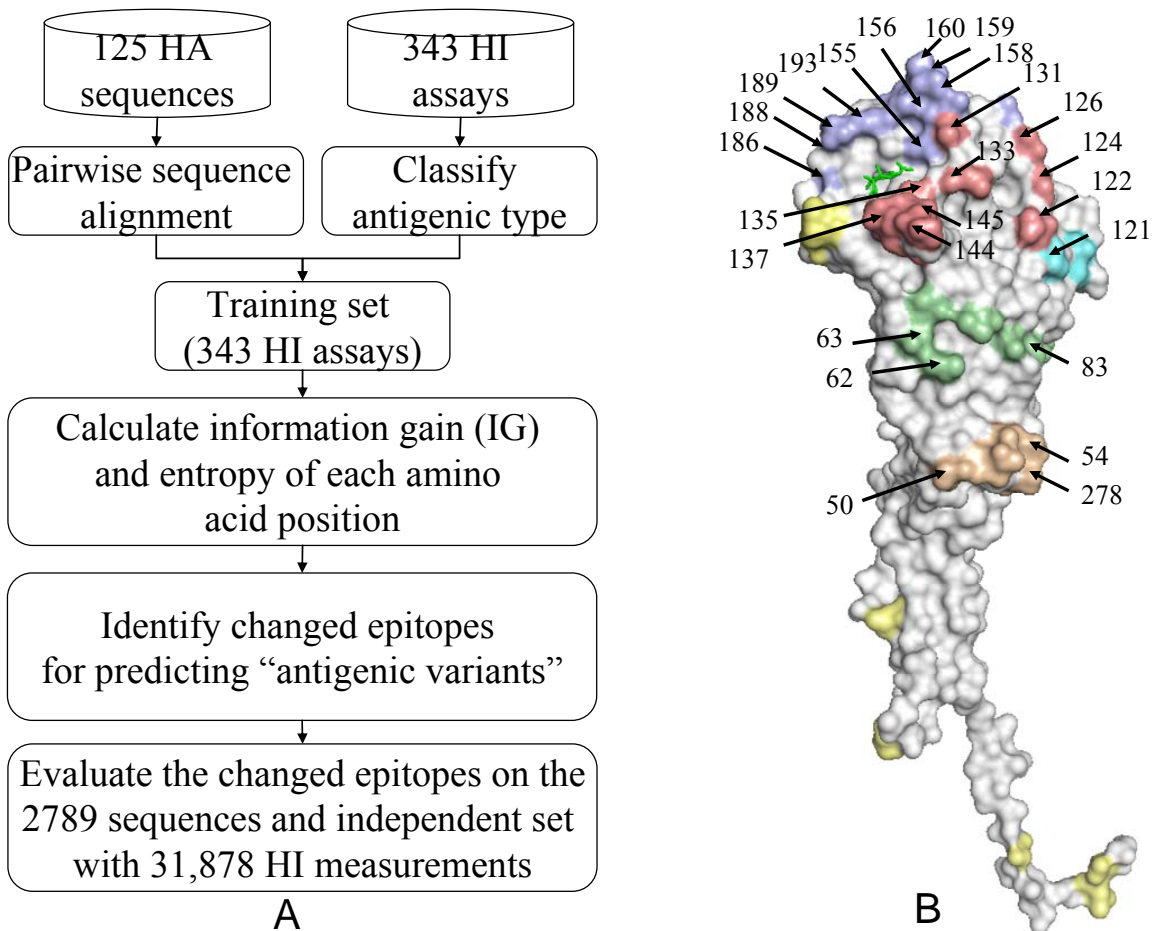


Figure 3.1 Overview of our method for the antigenic drift. (A) The overview of our method. (B) The structural locations of selected 64 critical amino acid positions on all the five epitopes (Epitope A in red; B in purple; C in orange; D in cyan; E in green). The sialic acid is in green. All structures are presented by using PyMOL [47].

3.3.1. Changed epitopes

The changed epitope is the core of our method. Here, we defined a changed epitope as follows: an antigenic site (epitope) on HA with accumulated amino acid mutations induces the conformation change to escape from the neutralizing antibody. The conformation change of a mutation depends on its position on HA structure and the mutation rate during 40 years. A changed epitope can be considered as a "key feature" for measuring antigenic variants of a pair of HA sequences. Here, a changed epitope can be used to predict antigenic variants and antigenic drifts for the selections of vaccine strains.

3.3.2. Data sets

To describe and evaluate the ability of the changed epitopes for predicting antigenic variants, we collected HI assays, describing the antigenic variants and similar viruses of the current global influenza surveillance system. The HI assay describes whether one (e.g. circulating) strain will be recognized by an antibody against the vaccine strain. We collected 343 H3N2 virus HI assays with 125 HA sequences from Weekly Epidemiological Record (WER) [50] (Table 3.1), World Health Organization (WHO) collaborating center [51] and related publications [52-54]. Each pair includes an HI assay value (i.e. antigenic distance) and a pair of HA sequences (329 amino acids). In general, an influenza vaccine should be updated if an antigenic distance is more than 4.0 between the current vaccine strain and the circulating strain in next season [15] [55]. Among 343 pairs of HA sequences, 225 pairs with antigenic distance ≥ 4 are considered as "antigenic variants" and 118 pairs are considered as "similar viruses". For example, the antigenic distance of the pair of HA sequences, A/England/42/72 and A/PortChalmers/1/73, is 12 and this pair is considered as "antigenic variants". Conversely, the antigenic distance of the pair of HA sequences, A/Wuhan/359/95 and A/Nanchang/933/95, is 1 and this pair is considered as "similar viruses". In addition to the training set, we prepared another HI assay data set to independently evaluate our model for predicting antigenic variants proposed by Smith *et al.* [15]. We assume that a virus-pair in the same antigenic group is considered as a "similar viruses" pair and a virus-pair in different groups is considered as "antigenic variants" pair. Finally, we obtained 31,878 HI measurements from the supporting materials [15].

To study the antigenic drifts and WHO vaccine updates, we collected 2789 HA sequences with influenza season assignment from influenza virus resource [36] and influenza sequence

database [35].

Table 3.1 The number of HI assays, number of sequences in Smith's dataset and WER strains from 1968 to 2007

Year	HI assay data set	Number of strains	WER strains ²
1968	0	4	A/Hong Kong/1/68
1969	0	3	A/Hong Kong/1/68
1970	1 ¹ [50]	2	A/Hong Kong/1/68
1971	2 [50]	4	A/Hong Kong/1/68
1972	2 [50]	5	A/Hong Kong/1/68
1973	2 [50]	4	A/England/42/72
1974	3 [50]	5	A/Port Chalmers/1/73
1975	6 [50]	3	A/Port Chalmers/1/73
1976	7 [50]	6	A/Victoria/3/75
1977	4 [50]	5	A/Victoria/3/75
1978	0	0	A/Texas/1/77
1979	0	0	A/Texas/1/77
1980	6 [50]	2	A/Bangkok/1/79
1981	0	1	A/Bangkok/1/79
1982	3 [50]	4	A/Bangkok/1/79
1983	5 [50], 23[53]	1	A/Phillipines/2/82
1984	0	1	A/Phillipines/2/82
1985	3 [50]	4	A/Phillipines/2/82
1986	9 [50]	2	A/Christchurch/4/85, A/Mississippi/1/85 ³
1987	3 [50]	3	A/Leningrad/360/86
1988	17 [50]	4	A/Sichuan/2/87
1989	0	16	Si/87; Sh/87 ³
1990	10 [50]	5	A/Shanghai/11/87
1991	5 [50]	17	A/Beijing/353/89
1992	5 [50]	45	A/Beijing/353/89
1993	21 [50]	43	A/Beijing/32/92
1994	6 [50]	10	A/Shangdong/9/93
1995	3 [50], 20 [52]	15	A/Johannesburg/33/94
1996	12 [50]	10	A/Johannesburg/33/94
1997	28 [56]	9	A/Wuhan/359/95
1998	3 [50]	4	Wu/95; Sy/97 ³
1999	2 [50]	3	A/Sydney/5/97
2000	0	1	A/Moscow/10/99 ⁴
2001	11 [54]	3	A/Moscow/10/99
2002	0	3	A/Moscow/10/99
2003	11[50], 20 [51]	6	A/Moscow/10/99
2004	4 [50], 17 [51]	0	A/Fujian/411/2002
2005	14 [50], 1 [51]	0	A/California/7/2004
2006	11 [50], 9 [51]	0	A/California/7/2004; A/Wisconsin/67/2005
2007	3 [50], 31 [57]	0	A/Wisconsin/67/2005; A/Brisbane/10/2007
Total	343 pairs	253 sequences	

¹ the number of HI assays collected in the document.

² we followed Plotkin's definition [34], WER strains were the dominant recommended virus based on HI assays in influenza season, as reported by the WHO in Weekly Epidemiological Record (WER).

³ for the purpose of detecting emerging variants, the later strain was selected to comparing with circulating strains.

⁴ the widely used vaccine strain A/Panama/2007/99 was used instead in following years.

3.3.3. Identifying critical positions on HA

Recently, we proposed a method to identify critical positions [32] by utilizing both antigenic variants and genetic diversity. The Shannon entropy and information gain (IG) were used to measure genetic diversity and antigenic discriminating score for amino acid positions on HA, respectively. Here, we based on these rules to select 64 amino acid positions as critical positions.

Table 3.2 Summary of 4 models

Model	Regarding HA positions	Changed epitope	Antigenic variants
Model one	329 positions	≥ 1 mutation	≥ 2 changed epitopes
Model two	329 positions	≥ 2 mutations	≥ 2 changed epitopes
Model three	64 selected positions	≥ 2 mutations	≥ 2 changed epitopes
Model four	64 selected positions	≥ 3 mutations (epitope B) ≥ 2 mutations (others)	≥ 1 (epitopes A or B) ≥ 2 (others)

3.3.4. Models for antigenic variants based on changed epitopes

To address the issue of measuring accumulated mutations on an epitope to escape from neutralizing antibody, we proposed 4 models considering the number of amino acid mutations on 329 amino acids and 64 selected critical positions of HA (Table 3.2). Models one and two regarded an epitope as "changed" if there are more than 1 and 2 mutations within an epitope, respectively, based on 329 amino acids. A changed epitope of Model three is defined as two amino acid mutations on 64 critical positions. Models one, two, and three regarded a pair of HA sequences as "antigenic variants" if there are more than two changed epitopes. Conversely, one changed epitope is viewed as "similar viruses".

Model four treated one changed epitope (A or B) as "antigenic variants". Epitopes A and B, which are near the receptor-binding site, often play the key role for escaping from neutralizing antibody. Here, the epitopes A and B (denoted as "B+") were regarded as "changed" if there are more than 2 and 3 mutations, respectively. For the pair A/Mississippi/1/85 and A/Leningrad/360/86 (Table 3.3), the numbers of mutations were 1, 3, 0, 1, and 1 on epitopes A, B, C, D and E, respectively. The numbers of changed epitopes for Models one and two are 4 (epitopes A, B, D, and E) and 1 (epitope B), respectively. Models three and four regarded the epitope B as a changed epitope because these three mutations (i.e. positions 156, 159 and 188) were the selected critical positions.

Finally, we compared our models with two related methods [9, 25] for predicting antigenic variants. Wilson & Cox [9] suggested that a viral variant usually contains more than 4 residue mutations located on \geq two of the five epitopes. Lee & Chen [25] proposed a model based on the hamming distance (HD) of 131 positions on all the five epitopes to predict antigenic variants. Their models predicted a pair of HA sequences as "antigenic variants" if the number of mutation is more than 6.

Table 3.3 The changed epitopes and mutations of 11 virus-pairs under 4 models

Virus A	Virus B	Type ¹	Changed epitopes				HD ²	Mutation positions				
			Model one	Model two	Model three	Model four		Epitope A	Epitope B	Epitope C	Epitope D	Epitope E
A/PortChalmers/1/73	A/Singapore/4/75	S	ABCDE	B	B	B	9	126 ³	160, 189	278	242	83
A/Nanchang/933/95	A/NewYork/43/96	S	ABCE	E	none	none	6	122	190	275		57, 92, 262
A/Alaska/10/95	A/France/75/97	S	ABCDE	BC	none	none	12	135	128, 165	275, 312	226	262
A/Sydney/5/97	A/Ireland/10586/99	S	ABDE	ABD	none	none	7	137, 142	192, 194		172, 226	57
A/Mississippi/1/85	A/Leningrad/360/86	V	ABDE	B	B	B+	6	138	156, 159, 188		226	88
A/Guizhou/54/89	A/Beijing/353/89	V	ABC	A	A	A	5	135, 144, 145	159	44,		
A/Wellington/1/2004	A/Victoria/505/2004	S	ABDE	AD	none	none	10	138, 145	189		219, 226, 94, 227	
A/Shangdong/9/93	A/Pennsylvania/9/93	S	ABCD	CD	C	C	12	135	164	53, 276	214, 219, 226, 229, 238	
A/England/42/72	A/PortChalmers/1/73	V	BDE	B	B	B+	6		160, 188, 193		208	63
A/NewYork/55/2004	A/Anhui/1239/2005	V	ABD	B	B	B+	7	138,	156, 160, 193		219,	138,
A/Shanghai/16/89	A/Beijing/353/89	V	AB	A	A	A	3	135, 145	159			

¹V represents antigenic variant and S represents similar virus.

² denote hamming distance of a pair sequences

³ Bold is the antigenic critical position.

3.3.5. Variant ratio for measuring the antigenic drift

We used the variant ratio (VR) to measure the vaccine efficiency on year y . The VR is defined as

$$VR(y) = \frac{V_y}{N_y},$$

where N_y is total number of circulating strains in the year y and V_y is the number of

circulating strains which are "antigenic variants" against the vaccine strain in the year. Here, we considered an influenza vaccine should be updated and the circulating strains are emerging if the VR value is more or equal than 0.5.

3.4. Results

3.4.1. Antigenic critical positions

In this study, we continued our previous work to select the critical positions [32] having high IGs, statistically derived from 343 HI assays, and high entropies, which were calculated using 2789 HA sequences. 64 positions on HA were selected as critical positions (Table 3.4). Among these 64 critical positions, 54 positions locate on the epitopes (54/64) and 53 positions locate on the HA surface (Fig. 3.1B). Additionally, 13 and 42 of these 64 critical positions were the positive selections [31] and cluster substitutions [15], respectively.

Table 3.4 The list of 64 critical positions in the five different epitopes [9, 31]

List of critical positions	
Epitope A	122,124,126,131,133,135,137,140,143,144,145,146
Epitope B	128,155,156,157,158,159,160,164,186,188,189,193,196,197,198
Epitope C	50,53,54,275,276,278,307
Epitope D	121,172,174,201,207,213,216,217,230,242,244,248
Epitope E	62,63,75,78,82,83,260,262
Other area	2,3,9,25,31,199,202,222,225,326

3.4.2. Changed epitopes for antigenic variants

Currently, several methods measured a changed epitope to escape from neutralizing antibody [9]. Here, we utilized the degree of accumulated mutations within an epitope to evaluate a changed epitope according to 329 positions and 64 selected positions. Figures 3.2 and 3.3 show the relationships between changed epitopes and antigenic variants on 4 models.

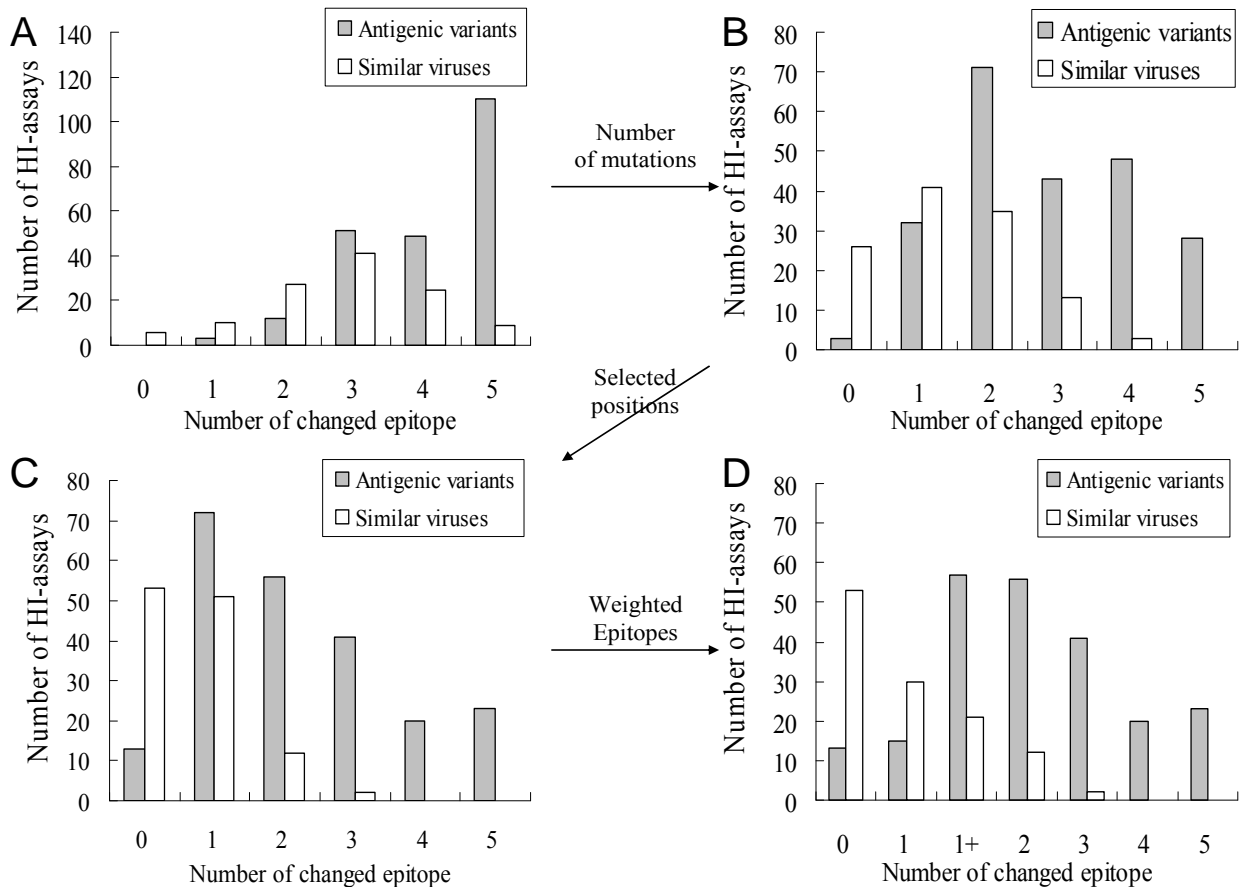


Figure 3.2 The relationships between number of changed epitopes and antigenic variants based on four proposed models. (A) The first model considered an epitope as changed if there is at least one mutation within it. (B) The second model considered an epitope as changed if there are at least two mutations within it. (C) The third model considered an epitope as changed if there are at least two critical mutations within it. (D) The fourth model was derived from model three and further defined "1+" type if there are at least 2 and 3 critical mutations in epitope A and B, respectively.

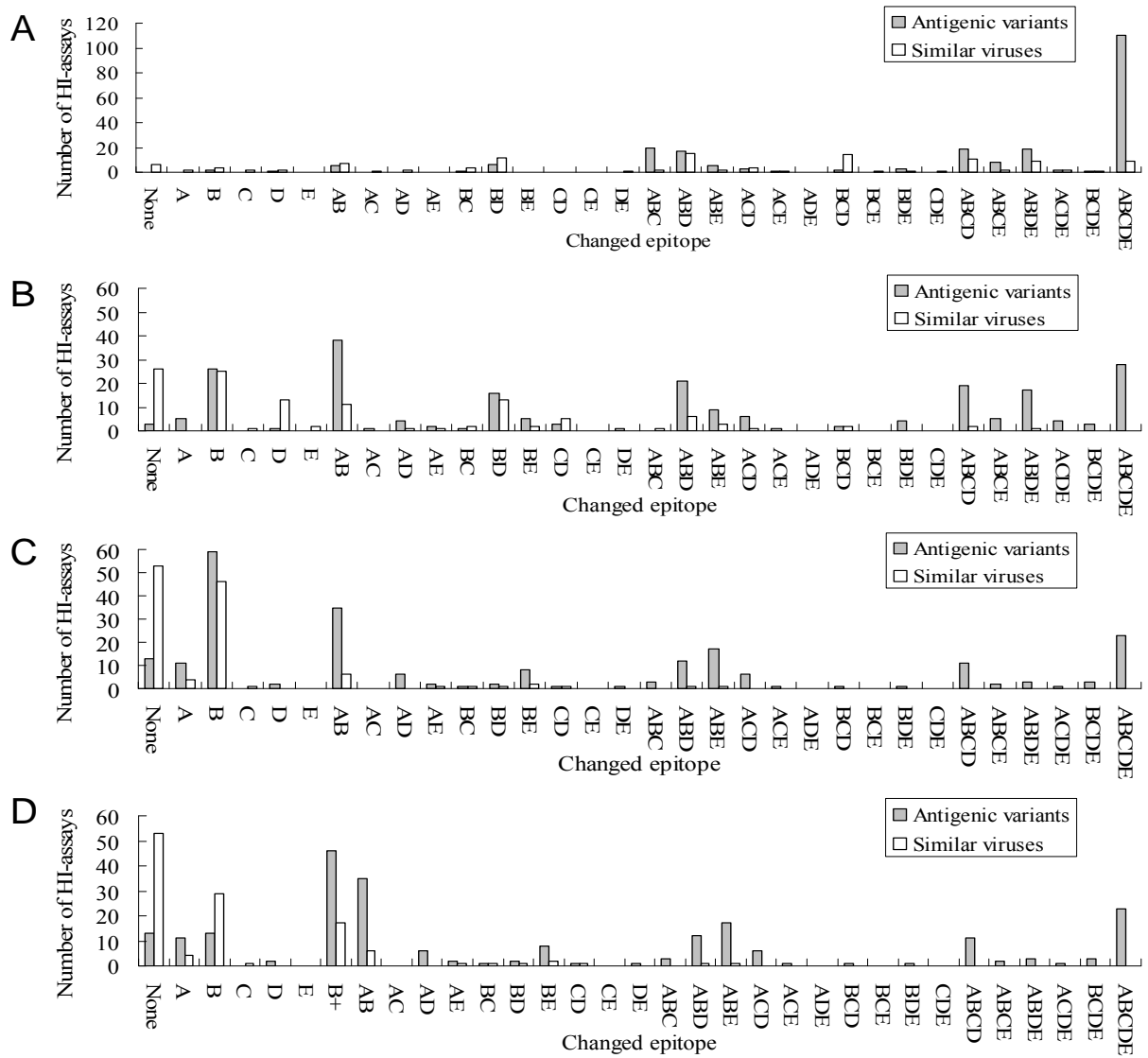


Figure 3.3 The changed-epitope composition and antigenic variants on 4 models. (A) Model one. (B) Model two (C) Model three and (D) Model four.

Models one and two: Changed epitopes on 329 positions

Figures 3.2A (Model one) and 3.2B (Model two) show the relationships between number of changed epitopes and "antigenic variants" on 343 pair of HA sequences with HI assays. Among these 343 pairs for Model one, the changed epitopes of 225 "antigenic variants" pairs range from 1 to 5 and the changed epitopes of 118 "similar viruses" pairs range from 0 to 5. Among 34 similar viruses with more than 4 changed epitopes for Model one, we observed the following results: (1) the average number of changed epitopes was 4.2; (2) the average number of changed epitopes with only one mutation was 2.02 and 33 pairs have more than one changed epitope with only one mutation. For example, the virus pair, A/PortChalmers/1/73 and A/Singapore/4/75, has four changed epitopes with one mutation (i.e. Epitopes A, C, D, and E) (Table 3.3). In general, these 34 similar viruses should be regarded as "antigenic variants" because there are more than four changed epitopes. This result shows that the Model one is not reasonable.

For Model two, the average number of changed epitopes was 2.2 for these 34 similar viruses. According to the distribution (Fig. 3.2B), Model two achieved the highest accuracy if more than two changed epitopes was considered as "antigenic variants". The accuracies were 74.9% (257/343) and 92.2% (29410/31878) for predicting antigenic variants on the training set and independent set, respectively. This result was similar to the previous work [9].

Model three: Changed epitopes on 64 selected positions

Model three considered a changed epitope when the number of mutations on the 64 selected critical positions is more than 2. In Model two, the numbers of "antigenic variants" and "similar viruses" with ≥ 3 changed epitopes were 119 and 16, respectively (Fig. 3.2B). The averages of changed epitopes with ≥ 2 mutations on 329 positions for "antigenic variants" and "similar viruses" were 3.8 and 3.2, respectively. The averages of changed epitopes with ≥ 2 mutations on 64 selected critical positions for "antigenic variants" and "similar viruses" were 3.2 and 1.5, respectively (Fig. 3.2C). These observations show that Model three using mutations on 64 critical positions is better than Model two to discriminate "antigenic variants" from "similar viruses". For the "similar viruses", A/Alaska/10/95 and A/France/75/97, there are 12 mutations to drive zero changed epitope because no epitope with ≥ 2 mutations on selected 64 positions (Table 3.3).

Three HA/antibody complex structures [10] can be used to provide structural evidences for the changed epitopes (Fig. 3.4). Among these complexes, two antibodies bind to epitopes A and B (PDB code 1KEN [58] and 2VIR [59]), while the third binds to epitopes C and E (PDB code 1QFU [60]). The antibodies consistently bind to two epitopes and this result agrees to Models two and three. HA/antibody structures and Models two and three show that two position mutations often induce the conformational change of an epitope to escape from the antibody recognition. However, the numbers of changed epitopes of 48 "similar viruses" pairs are 2 (35 pairs) and ≥ 3 (16 pair) for Model two (Fig. 3.2B). Conversely, 14 "similar viruses" pairs have more than 2 changed epitopes for Model three (Fig. 3.2C).

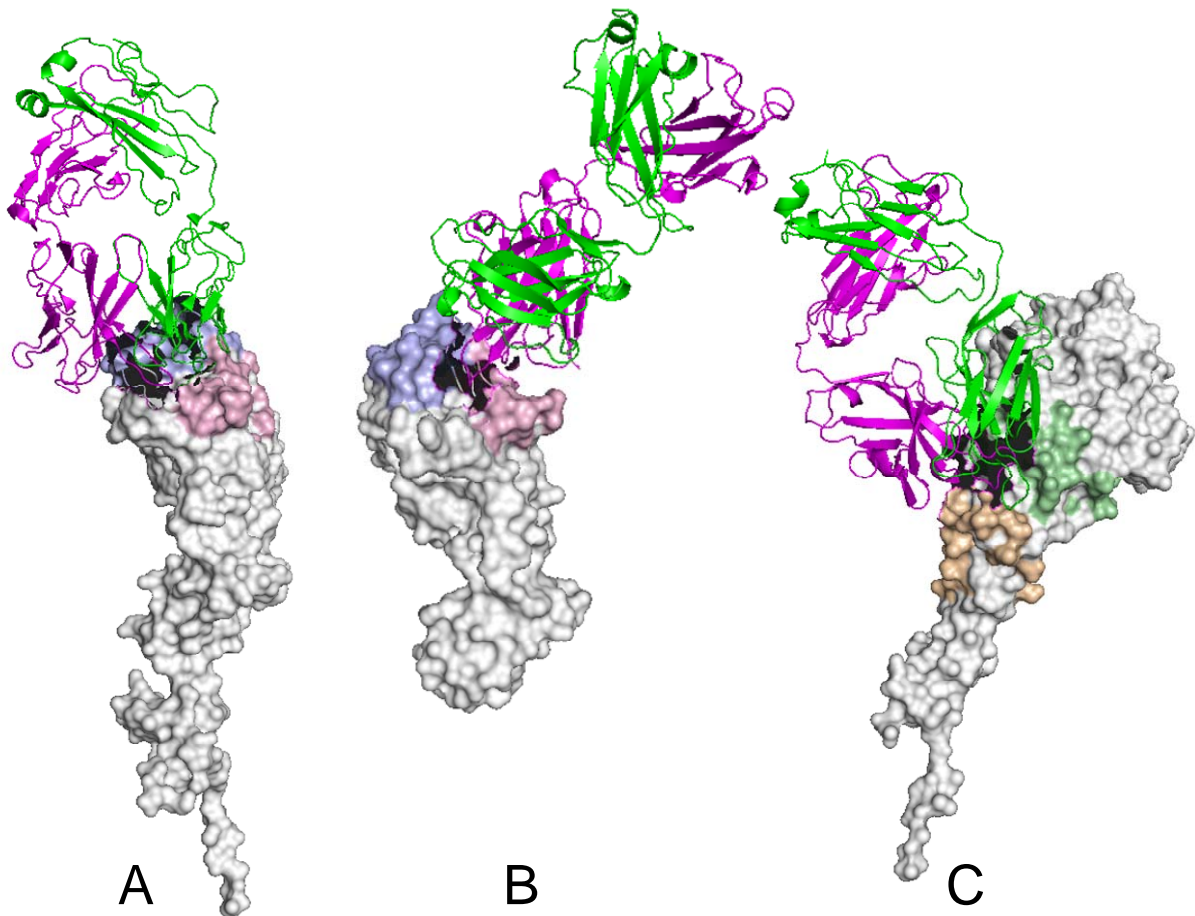


Figure 3.4 The three HA-antibody complex structures. PDB codes are (A) 1KEN [58] (B) 2VIR [59] and (C) 1QFU [60]. All of the three structures of antibodies bind on two epitopes on HA by heavy chain (pink) and light chain (green). The five epitopes on HA are labelled (Epitope A in red; B in purple; C in orange; D in cyan; E in green).

Model four

Among 72 "antigenic variants" pairs with one changed epitope based on Model three, 70 pairs change on epitopes A or B. The single changed epitope on A or B, which can cause "antigenic variants", agreed to HA/antibody complex structures and the experiments. The receptor-binding site, surrounded by epitopes A and B, is a basis for HA for the neutralizing mechanism [58, 61] (Fig. 3.1B).

Based on this observation, the epitopes A and B play a key role for neutralizing antibodies. Model four based on Model three considered a pair of HA sequences as "antigenic variants" when ≥ 2 changed epitopes or ≥ 1 changed epitope on A or B. In Model 4, a pair of HA sequences with ≥ 3 mutations on 64 critical positions for the epitope B is regarded as "antigenic variants". Thus, we annotated a virus-pair with single changed epitope on A or B as "1+" type (Fig. 3.3D). For example, the pair, A/Guizhou/54/89 and A/Beijing/353/89, occurs the changed epitope on A (i.e. mutation positions 135, 144 and 145) (Table 3.3). The accuracies of Model four were 81.6% and 94.0% on the training set and independent set, respectively. This model outperformed two compared methods, i.e. Wilson & Cox (89.7%) [9] and Lee & Chen (92.4%) [40], on the independent data set (Fig. 3.5).

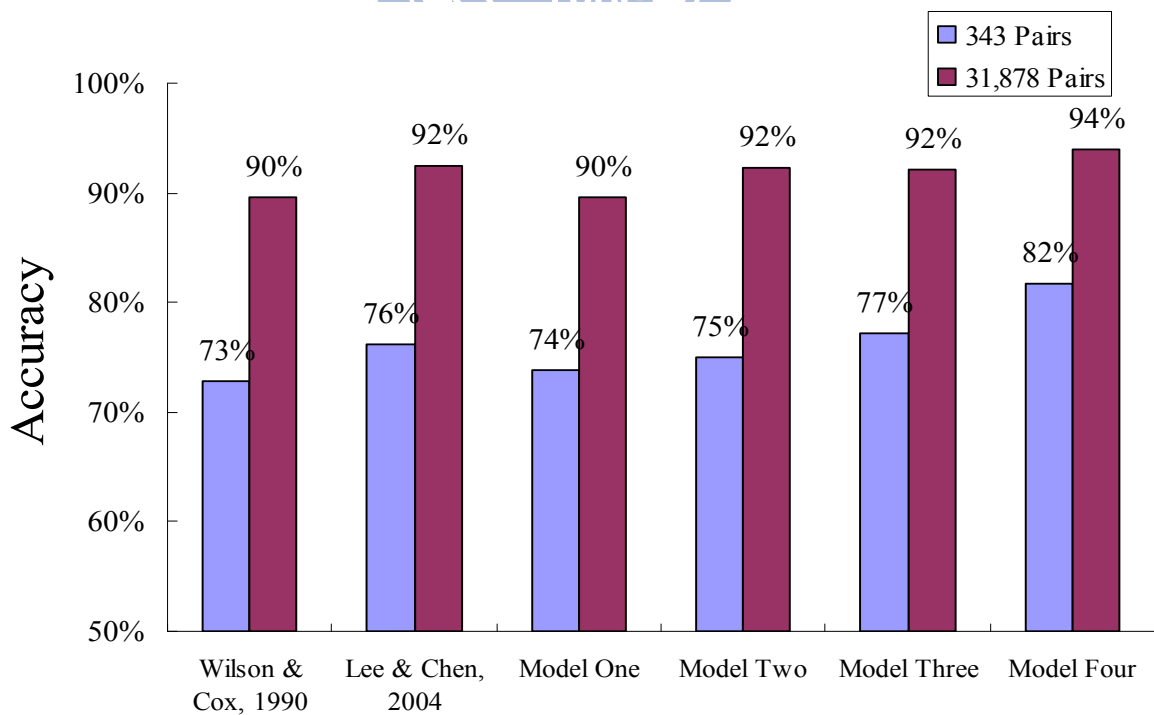


Figure 3.5 Comparison of our method with the other two methods (Wilson & Cox [9]; Lee and Chen [25]) on predicting antigenic variants on two data sets.

In the HA/antibody structure complex (PDB code 1KEN [58]), the antibody binds on epitopes A and B using two CDRs (i.e. CDR1 and CDR3) on the heavy chain and one CDR (i.e. CDR2) on the light chain (Fig. 3.6). The interface of antibody and HA consists of 13 and 5 contacted residues locating on epitopes B and A, respectively. Among these 13 positions, 7 positions were selected as critical positions. Based on Model four, 46 "antigenic variants" pairs have one changed epitope B with 3 mutations on epitope B, denoted as "B+". This result suggested a single changed epitope B can cause antigenic variants. For example, the pair virus strains, A/NewYork/55/2004 and A/Anhui/1239/2005, have three critical mutations on epitope B (i.e. positions 156, 160 and 193) (Table 3.3). According to the HA/antibody structure (Fig. 3.6), the residue 156 interacts to CDR2 (position 55 on the antibody) and the residue 193 interacts with three residues on CDR2 (positions 50, 55 and 57) and one residue on CDR3 (position 105). This structure suggested that mutations on residues 156, 160 and 193 can induce the conformation change on epitope B to escape from CDR2 and CDR3 of the neutralizing antibody.

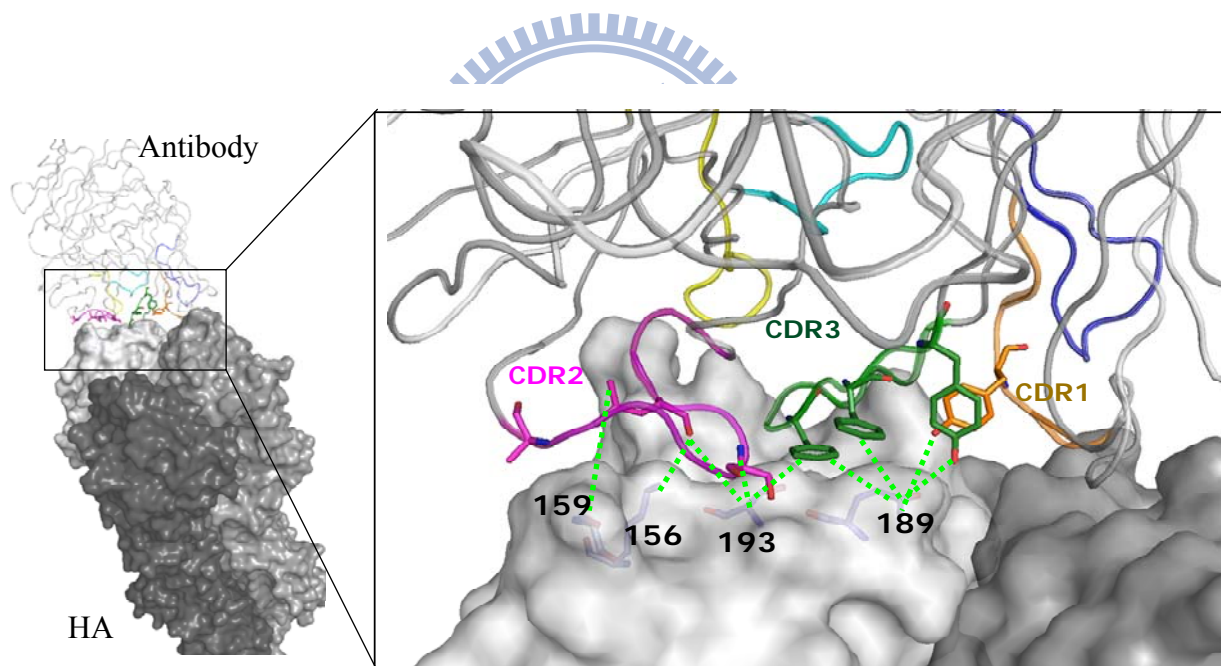


Figure 3.6 The HA/antibody structure and interface. (A) The antibody and HA trimer (PDB code 1KEN [58]). (B) The interface of the antibody and HA. The critical positions on epitope B and the CDRs of the antibody are labelled.

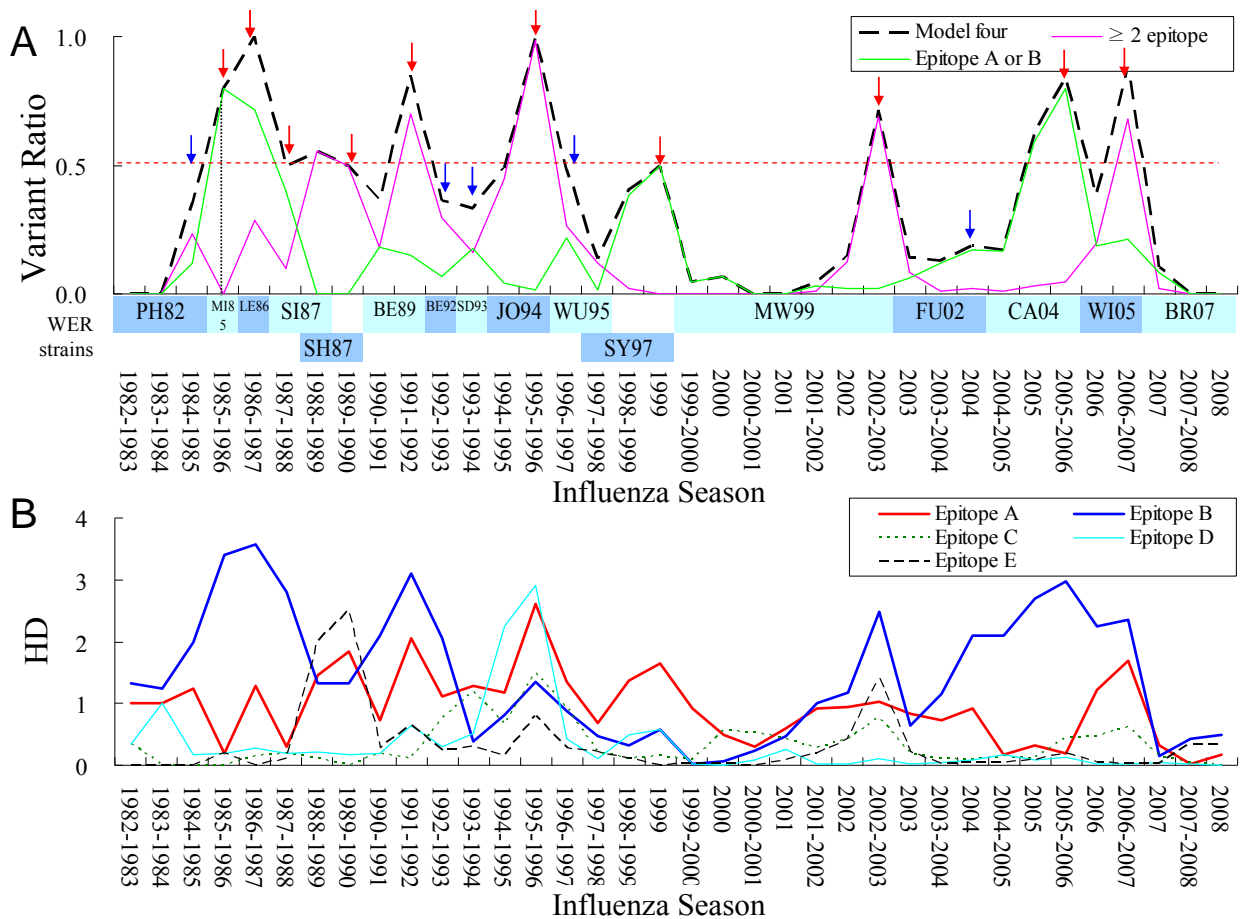


Figure 3.7 The epitope evolution and the antigenic drift from 1982-1983 to 2008 influenza season. (A) The distributions of variant ratios of WER strains from 1982-1983 to 2008 season. The match between Model four and WER are labelled (Match in red arrow; Not match in blue arrows). (B) The average hamming distances (HD) of 5 epitopes from 1982-1983 to 2008.

Antigenic drift and epitope evolution

We utilized the changed epitopes to study the antigenic drift on 2,789 circulating strains ranging from influenza season 1982-1983 to 2008 (36 influenza seasons). One of WHO surveillance network's purposes is to detect the emergence and spread of antigenic variants that may signal a need to update the composition of influenza vaccine [14-15]. Here, we considered an emerging antigenic variant according to WER strain, which was the dominant strain in each influenza season [34] (Table 3.1). For a selected season, we applied Model four, measuring changed epitopes for the pairs between the vaccine and circulating strains for "antigenic variants", and the variant ratio (VR) to detect the emerging antigenic variants.

Among 36 influenza seasons, our model detected 12 seasons with emerging antigenic variants ($VR \geq 0.5$) and 10 of them followed by the update of WER strain in the next season (Fig. 3.7A). For example, the 1985-1986 season, 80% of the circulating strains with changed epitope "B+" (Fig. 3.7B), is the first emerging antigenic variants and the WER strain updated in the next season (i.e. from A/Mississippi/1/85 to A/Leningrad/360/86). Moreover, among seven "emerging antigenic variants" seasons (matching WHO vaccine updates), four seasons (i.e. 1989-1990, 1991-1992, 1995-1996 and 2002-2003) matched the antigenic cluster transitions proposed by Smith *et al.* [15]. The other three seasons, which were detected by one changed epitope on A or B, are consistent to the WER strain updates (i.e. 1985-1986, 1987-1988 and 1999). These observations suggested that "emerging antigenic variants" with ≥ 2 changed epitopes may cause the major antigenic drift while "emerging antigenic variants" with one changed epitope on A or B may cause the minor antigenic drift.

To observe the epitope evolution, Figure 3.7B illustrates the hamming distance (HD) on 64 critical positions of all the five epitopes. For example, the VR of the season 1985-1986 was 0.8 (Fig. 3.7A) and the epitope with the largest HD was epitope B (HD is 3.4). For 15 seasons with WER strain updates, the average HD of epitopes A, B, C, D and E were 1.2, 2.1, 0.5, 0.4 and 0.4 respectively. These observations showed that epitopes A and B change more frequently in vaccine update seasons and play a key role for the antigenic drift.

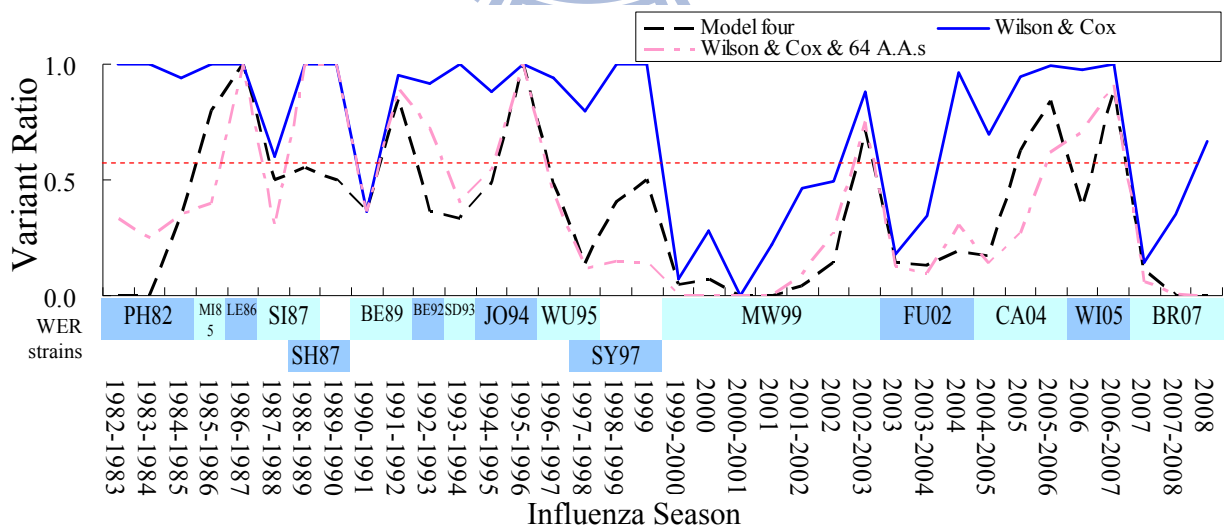


Figure 3.8 The comparison between our method and Wilson & Cox's model [9] in the antigenic drift from 1982-1983 to 2008 influenza season.

Table 3.5 Example of 13 antigenic variants without changed epitopes

Virus A	Virus B	Type (A to B) ¹	Type (B to A) ²	HD	Epitope A	Epitope B	Epitope C	Epitope D	Epitope E
Alaska/10/95	Idaho/4/95	V	S	3	145	165	312		
Alaska/10/95	Wuhan/359/95	V	S	5	135	165, 194	275		262
Alaska/10/95	Hongkong/55/95	V	S	7	135, 138	165	275	167, 226	262
Alaska/10/95	Shanghai/9/95	V	S	7	135	165, 193, 194	275	226	262
Anhui/1239/2005	Wisconsin/67/2005	V	S	5	122, 138	160			
Fujian/140/2000	Chile/6416/2001	V	S	12	144	186, 194	273	226, 246, 247	
Hong_Kong/1/94	Guangdong/25/93	V	S	8	124		47, 299	96, 216, 219, 226	92
Panama/2007/99	Chile/6416/2001	V	S	7	144	186		246	
Wellington/1/2004	Singapore/68/2004	V	S	9	145	189	50	226, 227	94
Wellington/1/2004	Victoria/513/2004	V	S	5	145	186, 190		167, 226	
Wellington/1/2004	Wisconsin/19/2004	V	S	5	138, 145	186	278	226	
Fujian/140/2000	NewYork/55/2001	V	V	12	144	186, 194	273	226, 229, 247	
Victoria/3/75	Victoria/112/76	V	V	2				229	

¹ the antigenic type of virus B relative to antisera against virus A.

² the antigenic type of virus A relative to antisera against virus B.

3.5. Discussion

Based on the accumulated HI assays from 1968 to 2008, we identified 64 critical positions on HA. Among the 64 critical positions, we observed that 10 positions are not located on all the five epitopes. Furthermore, 4 of the 10 positions were almost conserved from 1968 to 2000 and underwent frequency switch [41] after year 2000 (positions 25, 202, 222 and 225). These new emerging positions suggested that the previously conserved positions may become new antibody binding sites and IG can identify new emerging positions from HI assay. Moreover, the emerging mutations also revealed a need to update the epitope definition that proposed before 1999 [9, 31].

According to the distribution of antigenic variants of Model four (Fig. 3.3D), it is interesting that the main samples (209/225 pairs) of the antigenic variants have the changed epitope on epitopes A or B. In addition to this, among the 85 pairs of antigenic variants which had one or no changed epitopes, we observed that 56 pairs of them had ≥ 3 mutations in epitope A and B. These observations suggested that we may consider the epitopes A and B as one epitope and sum up the mutations in epitope A and B since these two epitopes were close to the receptor-binding site. Furthermore, many experiments suggested that the occlusion of the receptor-binding site by antibodies bound to the HA molecule forms the dominant neutralizing mechanism [58, 61].

Wilson & Cox [9] suggested that a viral variant usually contains more than 4 residue mutations located on \geq two of the five epitopes. Among the 225 antigenic variants, we observed

that 215 of them match Wilson and Cox's model. However, we also observed that 83 of the 118 similar viruses match their observation, which implied that their model had little ability to discriminate between antigenic variants and similar viruses. We also applied Wilson & Cox's model to detect the antigenic drift from seasons 1982-1983 to 2008 (Fig. 3.8). Among 36 influenza seasons, their model could detect 25 seasons with emerging variants ($VR \geq 0.5$) and 15 of them followed by the update of WER strain in the next season. Furthermore, their model detected only one season without emerging antigenic variants from seasons 1982-1983 to 1999, which suggested that their model had less ability to discriminate between seasons with and without emerging variants. We further compared our model with Wilson & Cox's model and we found that the major difference was due to the critical positions. If we applied Wilson & Cox's model only to the 64 critical positions, the false positive decreased from 10 seasons to only 3 seasons, which suggested that these critical positions were crucial for antigenic drift.

Gupta *et al.* proposed an antigenic distance between vaccine strain and circulating strains [30]. The proposed distance quantitatively measured the degree of change in the "dominant epitope", which was the epitope having largest fractional change in protein sequence. They further correlated the antigenic distance with the vaccine efficacies of 19 influenza A (H3N2) vaccines from 1971 to 2004. Among the 19 comparisons of vaccine strain and circulating strain, 13 of them were composed of different vaccine and circulating strain. It was interesting that all 13 pairs of them had dominant epitopes A or B and this suggested that the epitopes adjacent to the receptor-binding site were crucial for the antigenic drift. In addition, our model had been validated on 2,789 circulating strains while the proposed antigenic distance had been only validated only on 19 circulating strains.

Among 225 "antigenic variants" pairs, 13 pairs have no changed epitopes (Table 3.5). 11 of the 13 pairs have contradicting antigenic types by two antiseras, which suggested a more powerful experimental assay is required to verify the antigenic types. For example, the antibody against the A/Alaska/10/95 strain can't inhibit the A/Idaho/4/95 strain; while the antibody against the A/Idaho/4/95 strain inhibits the A/Alaska/10/95 strain.

The HA is a trimer protein and each subunit includes two chains, HA1 and HA2 [62]. Recently, Ekiert *et al.* identified an antibody that bound a new epitope in the stem of HA1 and HA2 chains [63]. The antibody blocks the conformational changes which are required for the fusion of viral membrane. This new epitope consists of two chains on HA and it can be regarded as two epitopes, this also implies that two changed epitopes can escape the neutralizing antibody.

3.6. Summary

This study demonstrates our model is robust and feasible for quantifying the changed epitopes. According to the distribution of antigenic variants in HI assays and HA/antibody complex structures, we found that two critical position mutations with high genetic diversity and antigenic scores can induce the conformation change of an epitope. Epitopes A and B, closing the receptor-binding site of HA, play a key role for neutralizing antibodies. Furthermore, two changed epitopes often drive the antigenic drift and can be used to explain the WHO vaccine strain selection. We believe that our method is useful for the vaccine development and understanding the evolution of influenza A viruses.



Chapter 4

A Bayesian Approach for Quantifying the Antigenic Distance of Influenza A (H3N2)

Viruses

4.1. Introduction

Influenza viruses often cause significant human morbidity and mortality [23]. The viral surface glycoproteins, HA and NA are the primary targets of the protective immune system. The viruses are able to continually evade host immune system through the accumulated mutations on the HA to change its antigenic properties through the time. The degree to which immunity induced by one (e.g. vaccine) strain is effective against another (e.g. circulating) strain is mainly dependent on the antigenic difference between two strains [14]. Thus, studies of antigenic difference among strains are important for the vaccine strain selection and many methods have been proposed to study the antigenic drift and vaccine development [15, 25, 31-32].

Among the sequence-derived methods measuring the antigenic difference crossing strains, hamming distance (HD) is one of the well-known methods. It counts the number of mutations between pairs of sequences and considers all amino acid positions as antigenic equivalents. However, not all positions on HA are on surface or on antibody combining sites that are recognizable by antibodies [9]. Moreover, some functional sites are evolutionarily conserved (e.g. serine 136 and tyrosine 98 in receptor-binding site) [11]. Furthermore, Smith *et al.* demonstrated that, "Antigenic evolution was more punctuated than genetic evolution, and genetic change sometimes had a disproportionately large antigenic effect". [41].

4.2. Motivation and aim

Based on the experimental results from previous two chapters, we observed that some positions were crucial for antigenic variants (e.g. position 145) while other positions had few effects for antigenic variants (e.g. position 226). We also noticed that mutations on epitope A and B seem more likely to cause antigenic variants. The above observations raise the question of whether the amino acid positions are antigenically equivalent or not.

Here, we proposed a Bayesian approach [64-66] to identify the antigenic drift of influenza A by quantifying the antigenic effect of each amino acid position on HA. We utilized the likelihood ratio (LR) to quantify the antigenic distance of an amino acid position. Based on naïve Bayesian network and LR, we developed an index, AD_{LR} , to quantify the antigenic distance of a given pair of HA sequences. Our experimental results show that the positions located on the epitopes and near the receptor-binding site are crucial to the antigenic drift. In addition to this, the AD_{LR} values are highly correlated to the HI assays and can explain WHO vaccine strain selection from 1968 to 2008.

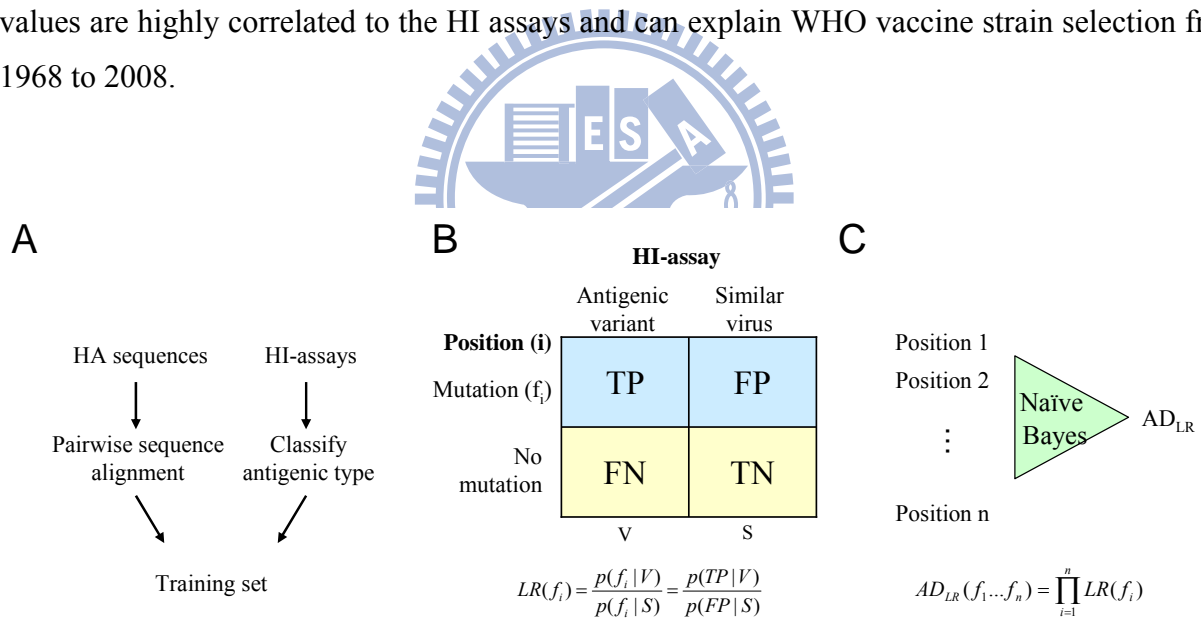


Figure 4.1 Overview of our method for quantifying the antigenic distance for amino acid positions and a pair of HA sequences. (A) The flowchart of training set preparation (B) The calculation of LR for a amino acid position to quantify its antigenic distance (C) The calculation of AD_{LR} for a pair of HA sequences to quantify their antigenic distance based on naïve Bayesian network [64-65].

4.3. Materials and methods

Figure 4.1 show the overview of our method for quantifying the antigenic distance for amino acid positions and a pair of HA sequences by calculating the likelihood ratio (LR) and AD_{LR} based on naïve Bayesian network [64-65].

4.3.1. Data sets

HI assays

For H3N2 virus, the HI assay data have had accumulated almost 40 years since 1968, which were selected in this study to quantify the antigenic distance of amino acid positions. We first collected influenza H3N2 virus HI assays from Weekly Epidemiological Record (WER) [Table 4.1], public documents from World Health Organization (WHO) collaborating center [Table 4.1] and publications [Table 4.1]. Then, we searched the H3N2 viruses with HI assays for their HA sequences in influenza virus resource [36] and influenza sequence database [35]. The number of collected HI assays with HA sequences available is 636 pairs and a subset of 343 pairs with 125 HA sequences were selected as a training set. In the training set, 183 pairs and 106 pairs of them are collected from WER and WHO collaborating center, respectively. The main samples (72%, 249 pairs among 343 pairs) consisted of pairs of vaccine-circulating strains and for each pair it was known whether there is inhibition of the circulating strain by antibodies against the vaccine strain ("antigenic variants" and "similar viruses"). Vaccine strains selected by WHO and are often the dominant strains of influenza seasons. Each pair includes a HI assay value (i.e. antigenic distance) and a bit string with 329 binary bits by aligning a pair of HA sequences (329 amino acids). For a specific position on a pair of HA sequences, the binary value is "1 (named as mutation)" if the residue types of the two sequences on this position are different; conversely, its binary value is "0 (named as no mutation)". In general, an influenza vaccine should be updated if an antigenic distance is more than 4.0 between the current vaccine strain and the strains expected to circulate in next season [15]. The antigenic distance between strains A and B is the reciprocal of the normalized HI assay of B relative to antisera raised against A [55]. Among 343 pairs of HA sequences, 225 pairs with antigenic distance ≥ 4 are considered as "antigenic variants" and 118 pairs with antigenic distance < 4 are considered as "similar viruses". For example, the antigenic distance of the pair of HA sequences, A/England/42/72 and A/Port_Chalmers/1/73, is 12 and this pair is considered as "antigenic variant". Conversely, the antigenic distance of the

pair of HA sequences, A/Wuhan/359/95 and A/Nanchang/933/95, is 1 and this pair is considered as "similar virus".

HA sequences

The HA sequences of the H3N2 virus were download from influenza virus resource [36] on April 5, 2008. After removing the sequences whose nucleotide shorter than 981 or repeated strain name, the number of sequences was 5,959. 4,548 of them can be further partitioned into influenza seasons according to their date of isolation or Plotkin's study [34]. For the sequences in the same influenza season, identical strains from the same geographic area were removed. Then, the number of sequences became 2,789 (Table 4.2) that distributed in 36 seasons ranged from 1983 to 2008. The influenza season is defined as 1 October through 30 September for the year before 1999. For example, the "1982-1983 season" refers to those sequences collected between 1 October 1982 and 30 September 1983. After 1999, the influenza season for northern and southern winter was defined as 1 November through 30 April and 1 May through 31 October, respectively. For example, the "1998-1999 season" refers to those sequences collected between 1 November 1998 and 30 April 1999 and the "1999 season" refers to those sequences collected between 1 May 1999 and 31 October 1999.

In addition to the training set, we prepared another data set proposed by Smith *et al.* [15] to compare with our method. This data set consists of 253 H3N2 viruses (Table 4.3) which are clustered into 11 antigenic groups. The sequences were extracted from supporting materials of publication [15].

Table 4.1 The data sources and composition of HI assay dataset from 1968 to 2007

HI assay source	Year	Number of HI assays (636)	Number of HI assays (343)	Reference
WER	1970	1	1	[67]
WER	1971	2	2	[68]
WER	1972	2	2	[69]
WER	1973	2	2	[70]
WER	1974	3	3	[71]
WER	1975	7	6	[72]
WER	1976	7	7	[73]
WER	1977	4	4	[74]
WER	1980	6	6	[75]
WER	1982	3	3	[76]
WER	1983	5	5	[77]
WER	1985	3	3	[78]
WER	1986	9	9	[79]
WER	1987	3	3	[80]
WER	1988	19	17	[81]
WER	1990	10	10	[82]
WER	1991	5	5	[83]
WER	1992	5	5	[84]
WER	1993	21	21	[85]
WER	1994	6	6	[86]
WER	1995	3	3	[87]
WER	1996	12	12	[88]
WER	1998	3	3	[89]
WER	1999	2	2	[90]
WER	2003	11	11	[91]
WER	2004	5	4	[92]
WER	2005	14	14	[93]
WER	2006	11	11	[94]
WER	2007	3	3	[95]
WHO collaborating center	1997	112	28	[56]
WHO collaborating center	2003	82	20	[96]
WHO collaborating center	2004	39	17	[97]
WHO collaborating center	2005	28	1	[98]
WHO collaborating center	2006	15	9	[99]
WHO collaborating center	2007	78	31	[100]
Publications	1983	46	23	[53]
Publications	1995	38	20	[52]
Publications	2001	11	11	[54]
Total		636	343	

Table 4.2 The number of sequences and WER strains from 1982-1983 to 2008 influenza season

Influenza season	Number of circulating strains	Number of circulating strains (non-identical ¹)	WER strains ²	WER strain ref
1982-1983	3	3	A/Phillipines/2/82	[77]
1983-1984	4	4	A/Phillipines/2/82	[101]
1984-1985	17	17	A/Phillipines/2/82	[78]
1985-1986	5	5	A/Christchurch/4/85, A/Mississippi/1/85 ³	[79]
1986-1987	7	7	A/Leningrad/360/86	[80]
1987-1988	10	10	A/Sichuan/2/87	[81]
1988-1989	9	9	Si/87; Sh/87 ³	[102]
1989-1990	6	6	A/Shanghai/11/87	[82]
1990-1991	11	11	A/Beijing/353/89	[83]
1991-1992	20	20	A/Beijing/353/89	[84]
1992-1993	96	61	A/Beijing/32/92	[85]
1993-1994	94	63	A/Shangdong/9/93	[86]
1994-1995	103	76	A/Johannesburg/33/94	[87]
1995-1996	89	73	A/Johannesburg/33/94	[88]
1996-1997	143	106	A/Wuhan/359/95	[103]
1997-1998	105	59	Wu/95; Sy/97 ³	[89]
1998-1999	99	47	A/Sydney/5/97	[90]
1999	22	14	A/Sydney/5/97	[90]
1999-2000	92	43	A/Moscow/10/99 ⁴	[104]
2000	90	43	A/Moscow/10/99	[104]
2000-2001	18	13	A/Moscow/10/99	[105]
2001	49	23	A/Moscow/10/99	[105]
2001-2002	151	91	A/Moscow/10/99	[106]
2002	173	105	A/Moscow/10/99	[106]
2002-2003	175	136	A/Moscow/10/99	[91]
2003	269	134	A/Fujian/411/2002	[91]
2003-2004	336	209	A/Fujian/411/2002	[92]
2004	274	159	A/Fujian/411/2002	[92]
2004-2005	365	256	A/California/7/2004	[93]
2005	225	130	A/California/7/2004	[93]
2005-2006	369	284	A/California/7/2004	[94]
2006	92	76	A/Wisconsin/67/2005	[94]
2006-2007	278	181	A/Wisconsin/67/2005	[95]
2007	71	48	A/Brisbane/10/2007	[95]
2007-2008	670	261	A/Brisbane/10/2007	[107]
2008	8	6	A/Brisbane/10/2007	[107]
Total	4548	2789		

¹ identical sequences in same season and same geographic area were removed.

² the dominant recommended virus based on HI assays, as reported by the WHO in WER

³ for the purpose of detecting emerging variants, the later strain is selected to comparing with circulating strains.

⁴ the wildy used vaccine strain A/Panama/2007/99 was used instead in following seasons.

Table 4.3 The number of sequences and WER strains in Smith's dataset from 1968 to 2003

Year	Number of strains	WER strains ¹	WER strain ref
1968	4	A/Hong Kong/1/68	[108]
1969	3	A/Hong Kong/1/68	[109]
1970	2	A/Hong Kong/1/68	[67]
1971	4	A/Hong Kong/1/68	[68]
1972	5	A/Hong Kong/1/68	[69]
1973	4	A/England/42/72	[70]
1974	5	A/Port Chalmers/1/73	[71]
1975	3	A/Port Chalmers/1/73	[72]
1976	6	A/Victoria/3/75	[73]
1977	5	A/Victoria/3/75	[74]
1978	0	A/Texas/1/77	[110]
1979	0	A/Texas/1/77	[111]
1980	2	A/Bangkok/1/79	[75]
1981	1	A/Bangkok/1/79	[112]
1982	4	A/Bangkok/1/79	[76]
1983	1	A/Phillipines/2/82	[77]
1984	1	A/Phillipines/2/82	[101]
1985	4	A/Phillipines/2/82	[78]
1986	2	A/Christchurch/4/85, A/Mississippi/1/85 ²	[79]
1987	3	A/Leningrad/360/86	[80]
1988	4	A/Sichuan/2/87	[81]
1989	16	Si/87; Sh/87 ²	[102]
1990	5	A/Shanghai/11/87	[82]
1991	17	A/Beijing/353/89	[83]
1992	45	A/Beijing/353/89	[84]
1993	43	A/Beijing/32/92	[85]
1994	10	A/Shangdong/9/93	[86]
1995	15	A/Johannesburg/33/94	[87]
1996	10	A/Johannesburg/33/94	[88]
1997	9	A/Wuhan/359/95	[103]
1998	4	Wu/95; Sy/97 ²	[89]
1999	3	A/Sydney/5/97	[90]
2000	1	A/Moscow/10/99 ³	[104]
2001	3	A/Moscow/10/99	[105]
2002	3	A/Moscow/10/99	[106]
2003	6	A/Moscow/10/99	[91]
Total	253		

¹ the dominant recommended virus based on HI assays in influenza season, as reported by the WHO in WER. Because most of these sequences are without influenza season assignment, the isolation year are used instead.

² for the purpose of detecting emerging variants, the later strain is selected to comparing with circulating strains.

³ the wildly used vaccine strain A/Panama/2007/99 was used instead in following seasons.

4.3.2. Quantifying the antigenic distance of amino acid positions

We applied the likelihood ratio (LR) [64], which is derived from Bayesian theorem [64-65], to quantify the antigenic distances of amino acid positions on HA. For each binary position f_i taking on a particular value (1 or 0; mutation or no mutation), the LR is the fraction of antigenic variants where the feature takes on the given value, divided by the fraction of similar viruses where the feature takes on the given value. The LR is then defined as follows:

$$LR(f_i) = \frac{p(f_i | V)}{p(f_i | S)} = \frac{(TP + K_v)/(V + K_v)}{(FP + K_s)/(S + K_s)}$$

where the true positives (TP) is the number of antigenic variants with feature takes on the given value and V is the total number of antigenic variant (V=225); false positive (FP) is the number of similar viruses with feature takes on the given value and S is the total number of similar viruses (S=118). For HI assays, the value of FP sometimes encounters zero probability and this difficulty may be surrounded when using Bayesian approach [113], here we added K_v and K_s as the pseudo counts to avoid of zero probability and we follow Lawrence's work [113], which selected K_v as \sqrt{V} ($K_v=15$) and K_s as \sqrt{S} ($K_s=10.8$). For example, for the position 145 is mutated, the number of "TP" and "FP" are 102 and 11, respectively, among 343 pair-wise HA sequences in the training set. According to these data, the $p(f_i | v)$ is 0.49 and $p(f_i | s)$ is 0.17. Finally, we obtained LR=2.87. Furthermore, when we called the "LR of a position", which means the LR for a position that is mutated.

4.3.3. Quantifying the antigenic distance of a pair of HA sequences

Given a pair of HA sequences, we use the LR for combined positions from naïve Bayesian network [64-65] to quantify its antigenic distance and defined as AD_{LR} . For each pair of sequences, the AD_{LR} is based on the calculation of the posterior odds of antigenic variant given the mutated and $LR>1$ positions. The posterior odds for antigenic variant are defined by integrating mutated and $LR>1$ positions $f_1 \dots f_n$ can be written as follows using the Bayes' rule:

$$\log_2 \left(\frac{P(V | f_1 \dots f_n)}{P(S | f_1 \dots f_n)} \right) = \log_2 \left(\frac{P(f_1 \dots f_n | V)}{P(f_1 \dots f_n | S)} \right) + \log_2 \left(\frac{P(V)}{P(S)} \right)$$

where V and S represents antigenic variants and similar virus, respectively. f_i through f_n are different mutated and $LR>1$ positions. $P(V | f_1 \dots f_n)$ is the probability that the pair is antigenic

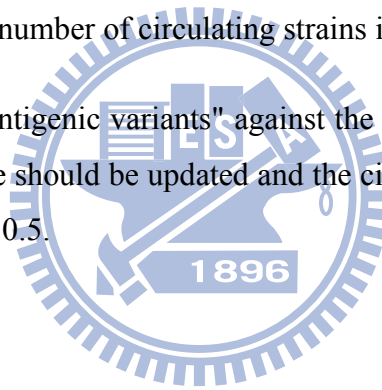
variant given these positions. $P(f_1 \dots f_n | V) / P(f_1 \dots f_n | S)$ is the likelihood ratio for the combined positions and defined as AD_{LR} . $P(V) / P(S)$ is the prior odds. We then assumed that the n positions are conditionally independent, the Bayesian network is so-called naïve Bayesian network and AD_{LR} can be simplified to

$$AD_{LR}(f_1 \dots f_n) = \prod_{i=1}^n LR(f_i)$$

A pair of sequences is predicted as antigenic variant if the calculated AD_{LR} is greater than a predetermined threshold.

4.3.4. Variant ratio for studying the antigenic drift

We used the variant ratio (VR) to measure the vaccine efficiency on year y . The VR is defined as $VR(y) = \frac{V_y}{N_y}$, where N_y is total number of circulating strains in the year y and V_y is the number of circulating strains which are "antigenic variants" against the vaccine strain in the year. Here, we considered an influenza vaccine should be updated and the circulating strains are emerging if the VR value is more or equal than 0.5.



4.3.5. Shannon entropy

Shannon entropy was used to measure the genetic diversity of an amino acid position i ($i=1$ to 329) with 20 amino acid types and is defined as

$$H(i) = -\sum_{T=1}^{20} P(A_i = T) \log(P(A_i = T))$$

where $P(A_i=T)$ is the probability of the position i with amino acid type T .

4.3.6. Contact-pair distance on antigen-antibody interaction

To gather the statistics of the contact-pair distance on antigen in antigen-antibody interaction, we selected 54 antigen-antibody protein complexes (Fig. 4.2B) from BEID [114] based on two criteria: the antigens are proteins and dimer whose chains are longer than 30 residues [115]. Based on these complexes, 8,261 contact residue pairs are identified among 949 contact residues on antigens, which are identified by 3D-partner [115]. The contact-pair distance between two contact residues i and j on the same antigen is defined as follows:

$$\text{distance}(i, j) = \sqrt{(i_x - j_x)^2 + (i_y - j_y)^2 + (i_z - j_z)^2}$$

where (i_x, i_y, i_z) and (j_x, j_y, j_z) are the x, y, and z coordinates of C α atom on residue i and j . For example, there are 15 contact residues on antigen in the HA/antibody complex (PDB code 2VIR [59]) and the longest distance among them is 27.13Å that between position 129 and position 226 (Fig. 4.2A).

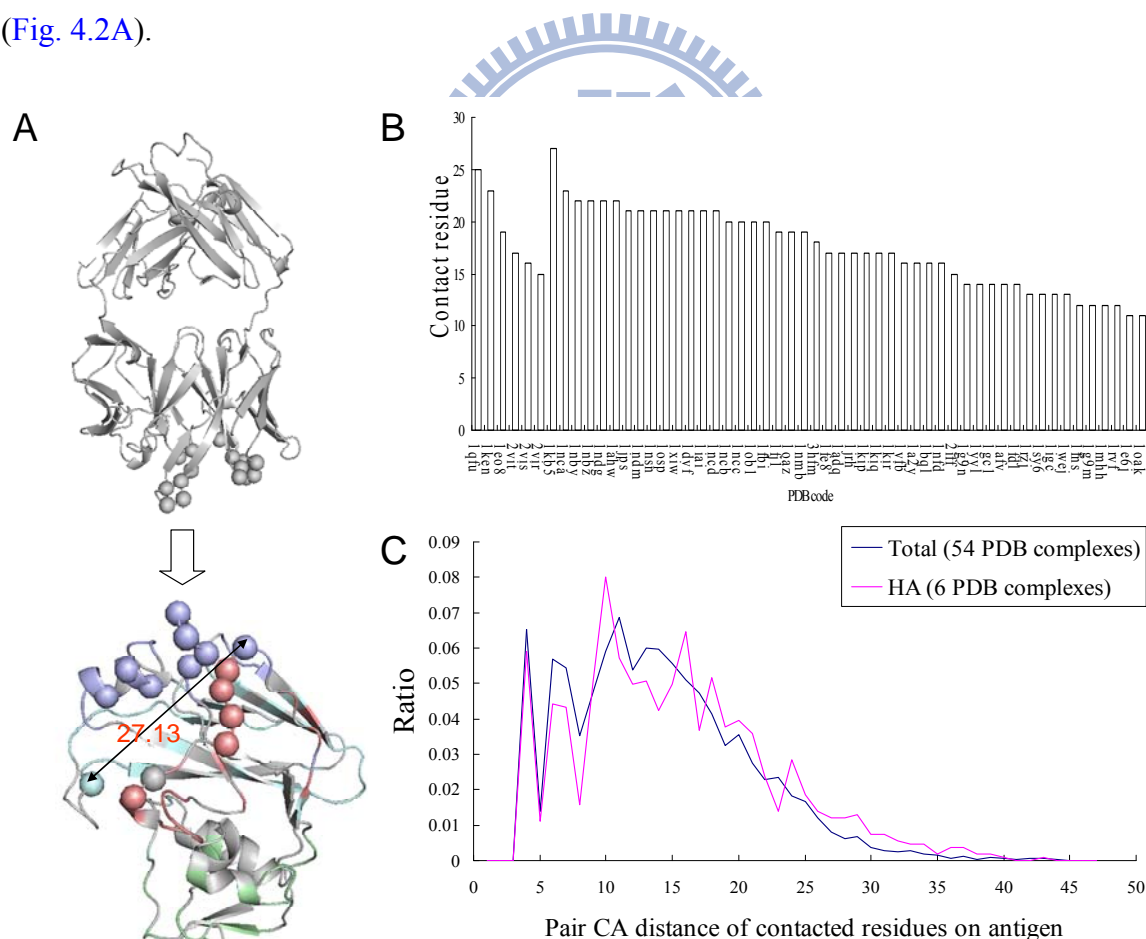


Figure 4.2 The statistics of 54 antigen-antibody complex structures. (A) The distribution of contact residues on antibody and HA (showed in spheres). (B) The distribution of contact residues among 54 complex structures. (C) The distribution of contact-residue distance.

4.3.7. Amino acid distance to sialic acid

The distance of an amino acid position i ($i=1$ to 329) to sialic acid on HA structure is defined as follows

$$D(i, sia) = \frac{\sum_{s=1}^{21} \text{distance}(c_{\alpha i}, atom_s)}{21}$$

where $C_{\alpha i}$ is the C_{α} atom of amino acid i and $atom_s$ is one of the 21 heavy atoms on sialic acid. The PDB code of the selected HA structure is 4hmg.

4.4. Results

4.4.1. Antigenic distance of amino acid positions

In this study, we used LR to quantify the antigenic distance of an amino acid position, which locates at the specific site on HA, for their different probability of antigenic variants in HI assay. An amino acid position with $LR > 1$ means that this position is more correlated to the antigenic variants than to similar viruses. The highest and lowest values of LR in this study are 2.87 and 0.49, respectively. Among the 329 amino acids of HA, 131 positions and 166 positions are considered to lie in the five antibody-binding sites (named as epitopes), which are labeled A through E [9, 31] and on the surface, respectively. Table 4.4 summarizes the LR of amino acid positions on HA. The LR of 69 positions are larger than 1 and most of these 69 positions are on the epitope (60/69) or surface (57/69), which suggested that positions accessible to antibodies have higher antigenic distance. The summary of 10 amino acid positions are listed in Table 4.5 and Figure 4.3. The first rank, position 145-A with $LR=2.87$, locates at the epitope A and surface of HA. Among 343 pairs of HA sequences in the training set, the position 145-A mutating on 113 pairs and 102 pairs are the antigenic variants. This result implied that a mutation on this position highly induces an antigenic drift. This observation is consistent to the results of Smith *et al.* [15], that is, the position 145 participated in four of ten antigenic cluster transitions (EN72-VI75, SI87-BE89, BE89-BE92 and BE92-WU95) and the single amino acid substitution N145K can be responsible for antigenic cluster transition (BE92-WU95). The LR of another position (position 140-A), which also locates on epitope A and surface of HA, is 1.16 and this position is not selected by any related works. The position 140-A is almost conserved at amino acid lysine (K) from 1968 to 2006 and suddenly mutated into isoleucine (I) at 2007 (Fig. 4.3). The K140I

mutation is also observed between the two vaccine strains A/Wisconsin/67/2005 and A/Brisbane/10/2007 that was used in recent influenza seasons. The LR also identified potential antigenic sites that are not identified as epitope before, such as position 222 and 225 that are almost conserved from 1968 to 2000 and started to mutate after 2000 (Fig. 4.3), which are also observed in antigenic cluster transition from SY97 to FU02 group [15]. The LR value can be smaller than 1 which means the mutation on these positions have smaller probability for antigenic variants than similar viruses, such as position 138 and 190, which are selected as positive selection codons. These observations suggested that not all frequently mutated positions have high antigenic distance [32].

Table 4.4 The summary of LR on 329 amino acid positions

	HA	Epitope	Surface ¹	Positive selection ²	Smith <i>et al</i> ³	Shih <i>et al</i> ⁴
LR>1	69	60	57	14	42	53
LR≤1	260	71	109	4	1	10
Total	329	131	166	18	43	63

¹ the position with relative accessibility > 0.16 [116] in any one of two HA structures (PDB: 4hmg and 1hgf)

² the position is under positive selection defined by Bush *et al.*[31].

³ the position is a cluster-difference substitution defined by Smith *et al.*[15].

⁴ the position is a frequency switch sites defined by Shih *et al.*[41].

Table 4.5 The TP, FP, $p(f_i | V)$, $p(f_i | S)$, LR and distance to sialic acid of 10 amino acid positions on HA

Position -epitope	TP	FP	$p(f_i V)$	$p(f_i S)$	LR	Distance to sialic	Epitope	Surface ¹	Positive selection ²	Smith <i>et al</i> ³	Shih <i>et al</i> ⁴
145-A	102	11	0.49	0.17	2.87	9.4	A	+	+	+	+
278-C	48	4	0.26	0.12	2.28	48.9	C	+		+	+
156-B	98	16	0.47	0.21	2.26	12.9	B	+	+	+	+
137-A	43	3	0.24	0.11	2.25	7.3	A	+		+	+
225-	35	4	0.21	0.12	1.81	10.3	-	+		+	+
193-B	81	20	0.40	0.24	1.67	10.8	B	+	+	+	+
222-	16	0	0.13	0.08	1.53	13.8	-	+		+	+
140-A	17	4	0.13	0.12	1.16	12.1	A	+			
138-A	49	27	0.27	0.29	0.91	8.9	A		+		
190-B	28	15	0.18	0.20	0.89	9.7	B	+	+	+	+

¹ the position with relative accessibility > 0.16 [116] in any one of two HA structures (PDB: 4hmg and 1hgf)

² the position is under positive selection defined by Bush *et al.*[31].

³ the position is a cluster-difference substitution defined by Smith *et al.*[15].

⁴ the position is a frequency switch sites defined by Shih *et al.*[41].

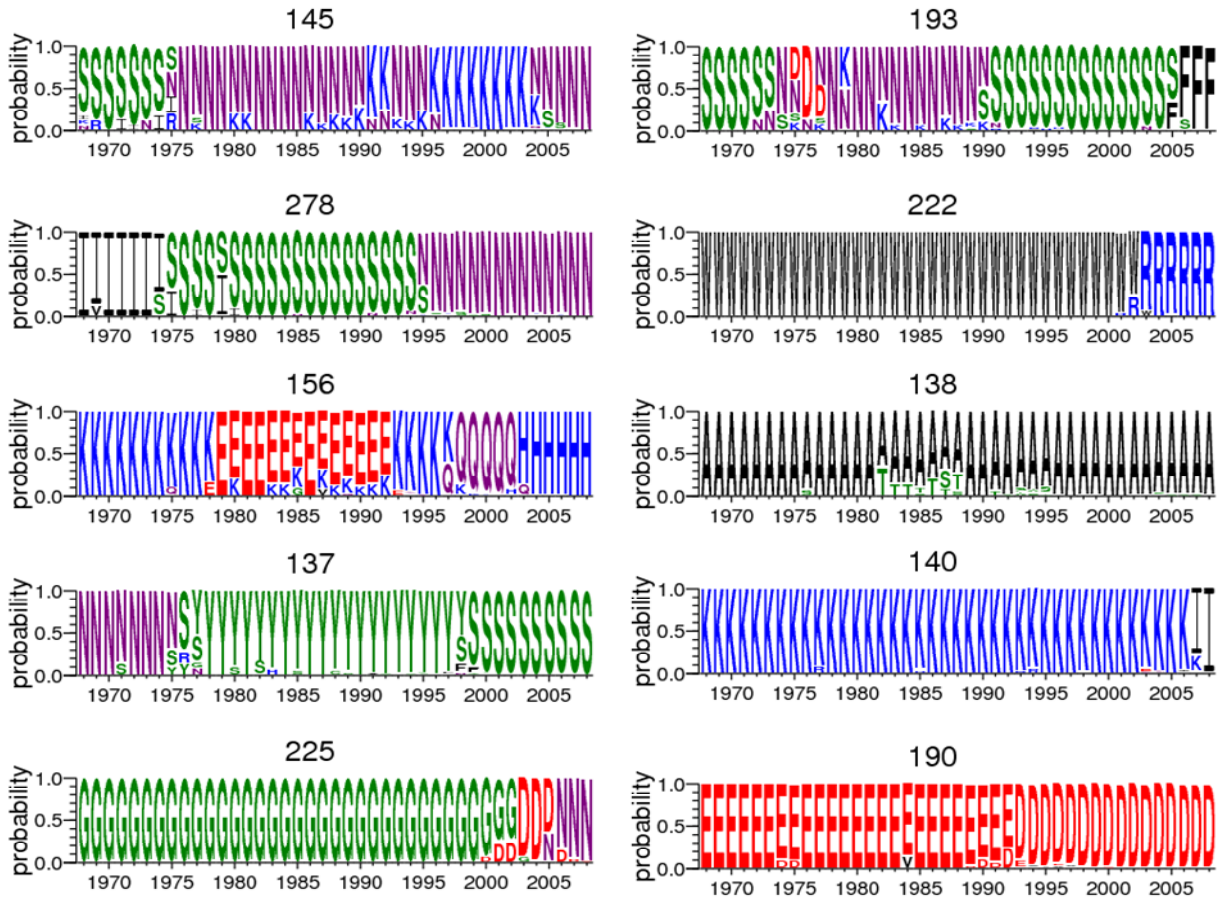


Figure 4.3 The frequency diagram of 10 amino acid positions on HA [41]

4.4.2. Antigen-antibody interaction

Most studies on the influenza virus mainly focused on the evolution of antigen, although the entry of the influenza virus is highly correlated to the escape of neutralizing antibodies. Here we proposed an antigen-antibody view to study the antigenic evolution on HA. It is suggested by many experiments that occlusion of the receptor-binding site (RBS) by antibodies bound to the HA molecule forms the dominant neutralizing mechanism [10, 59, 61]. From the view point of antigen-antibody interaction, we formulated the problem are as follows: on which distance to RBS may amino acid positions on HA affect the occlusion of RBS by antibodies and whether high LR positions are related to this event. Firstly, we collected 54 antigen-antibody complexes and gathered the statistics of contact pair distance on antigen. Figure 4.2 shows that the contact residues number on 54 antigens ranged from 11 to 27 and the contact residues number in 6 HA-antibody complexes (PDB code 2VIR [59], 2VIS [59], 2VIT [59], 1EO8 [120], 1KEN [58] and 1QFU [60]) are also within this range. Then, the contact pair distance was applied to calculate the contact area on the antigen. Among the 8,261 contact residue pairs on 54 antigens

(Fig. 4.2C), the contact pair distance of 8091 pairs (98%) are within 30Å, which means that most of the contacted area on antigen are within 30Å. In other words, the mutations that have potential to affect antibody occluding RBS are suggested to locate within 30Å to RBS.

The relationships between LR values and HA-antibody complexes are shown in Fig. 4.4 by using PyMOL [47]. Currently, there are two antibodies that mainly bind the RBS (Fig. 4.4A), and one of them prevents the HA transition that is required for fusion of the virus [58]. To model the occlusion of RBS by antibody, we selected the sialic acid as the center of RBS and calculated each position's distance to sialic acid as its distance to RBS. Figure 4.4B shows the relationship between LR values and structural locations of positions. These positions can be roughly divided into two groups according to they are surrounded the RBS ($\leq 20\text{\AA}$ to sialic acid) or not ($> 20\text{\AA}$ to sialic acid, such as position 122-A, 124-A and 126-A). The LR values at different distance to sialic acid is plot in Fig. 4.4D and the mean of LR value are larger than 1.33 for positions $\leq 20\text{\AA}$ to sialic acid while the mean of LR is near 1 for positions located at 20\AA to 30\AA to sialic acid. The distribution of entropy value has similar behaviors as LR in Fig. 4.4C, which shows that the positions $\leq 20\text{\AA}$ to sialic acid are frequently mutated. These result showed that the positions $\leq 20\text{\AA}$ to sialic acid have higher antigenic distance and are frequently mutated, which implies that mutations on these high LR positions are highly possible to let the HA to escape the occlusion of RBS by antibodies. For the $LR > 1$ positions within 20\AA to sialic acid, we further divided them into 5 segments according to their epitope and secondary structure. The segment "I" and "II" include positions in two loops in epitope A (131-A, 133-A, 135-A and 137-A) and (140-A, 142-A, 143-A, 144-A, 145-A and 146-A), respectively. The segment "III" and "IV" include positions in one loop and helix in epitope B (155-B, 156-B, 157-B, 158-B, 159-B and 160-B) and (186-B, 188-B, 189-B, 193-B and 196-B), respectively. The segment "V" includes positions in non-epitope region (222-other and 225-other). The segment II and III are considered as important antibody-binding site because the projected from HA surface and could accept mutations without affecting the framework of HA [11].

Figure 4.5 shows the LR of Smith's 43 positions [15] and the distribution shows that these 43 positions have higher LR than other positions, which implies that positions in antigenic cluster transition have larger antigenic distance than other positions.

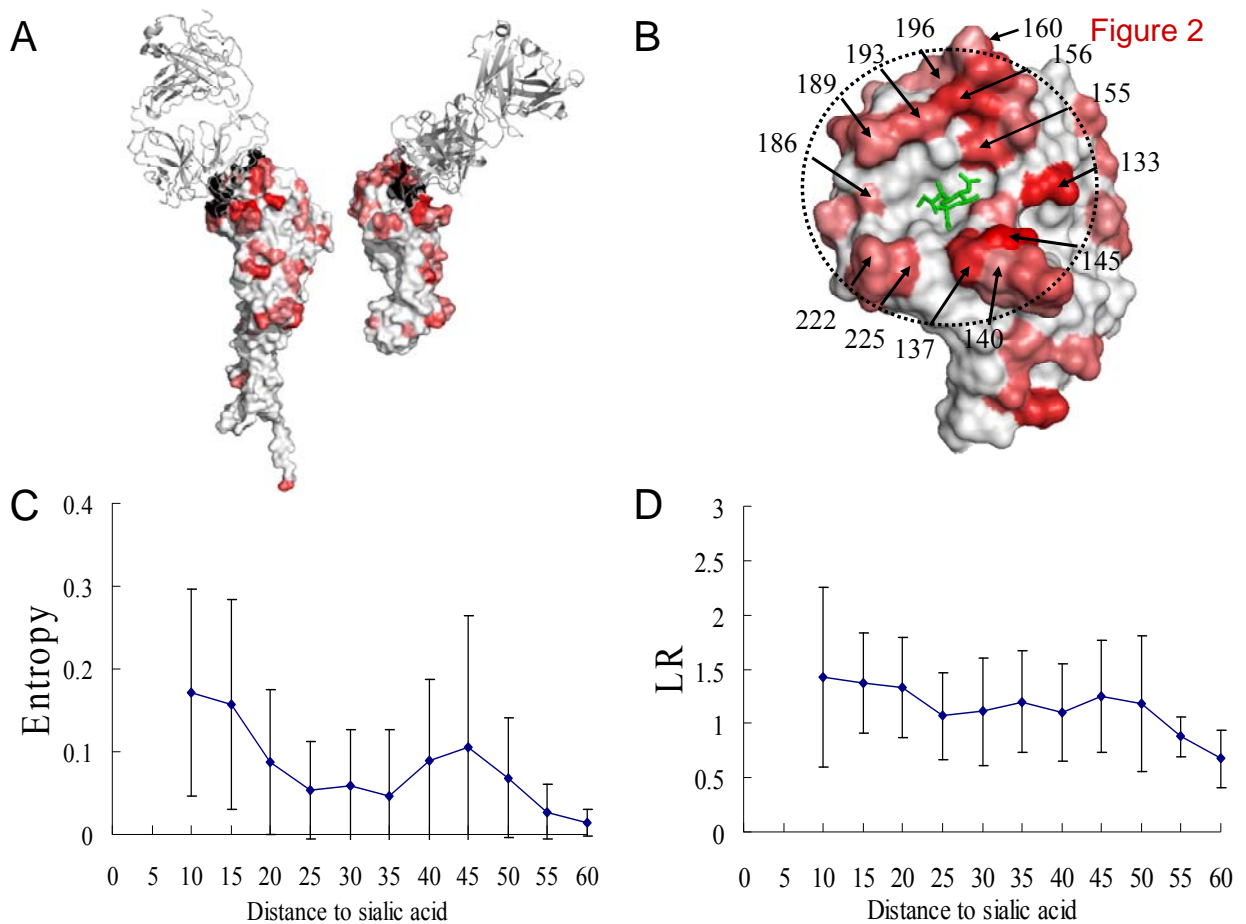


Figure 4.4 The relationships between LR and HA-antibody complexes. (A) The HA-antibody complexes and LR values distribution on HA structure. The red and gray indicate the highest LR value and the lowest LR value, respectively. The antibodies are shown in gray (PDB code 1KEN [58] and 2VIR [59] for left and right complex, respectively). (B) The structural locations and LR values of positions within 30Å to sialic acid (in green) (PDB code 4HMG [117]). The radius of the circle is roughly 20Å. (C) The distribution of entropy at different distance to sialic acid. (D) The distribution of LR values at different distance to sialic acid. The structure is presented by using PyMOL [47].

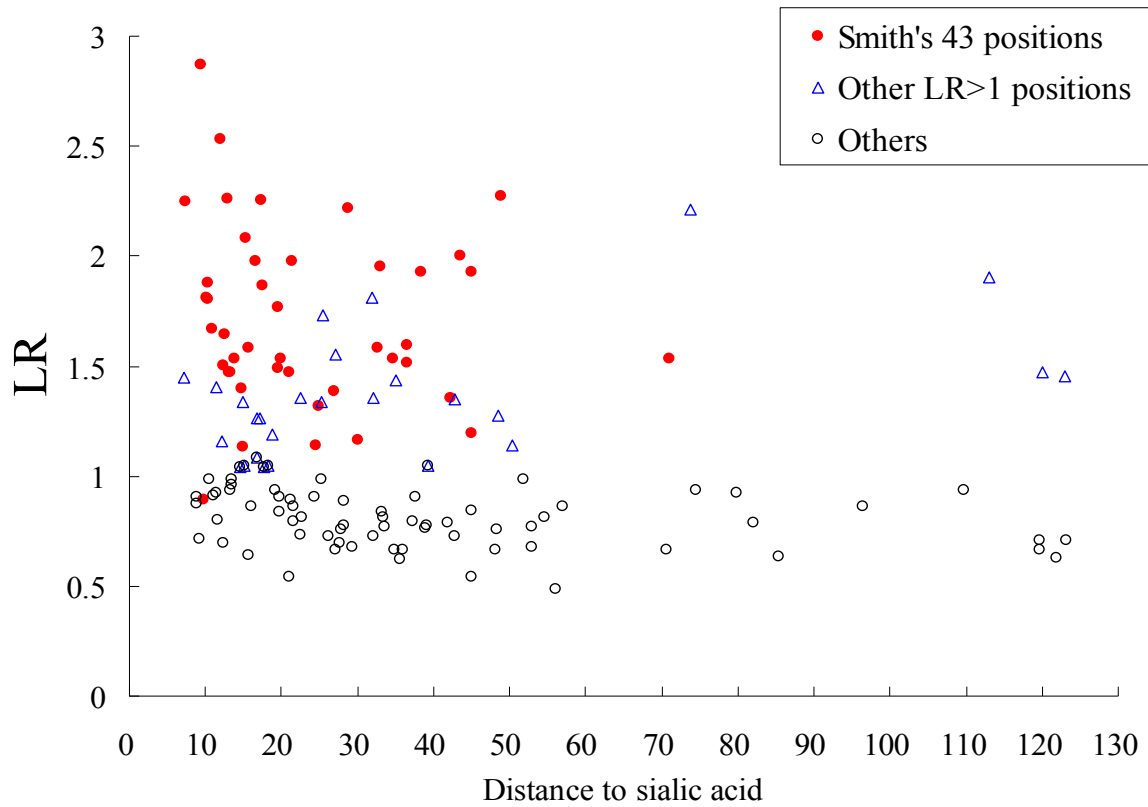


Figure 4.5 The LR values distribution of Smith's 43 positions [15].

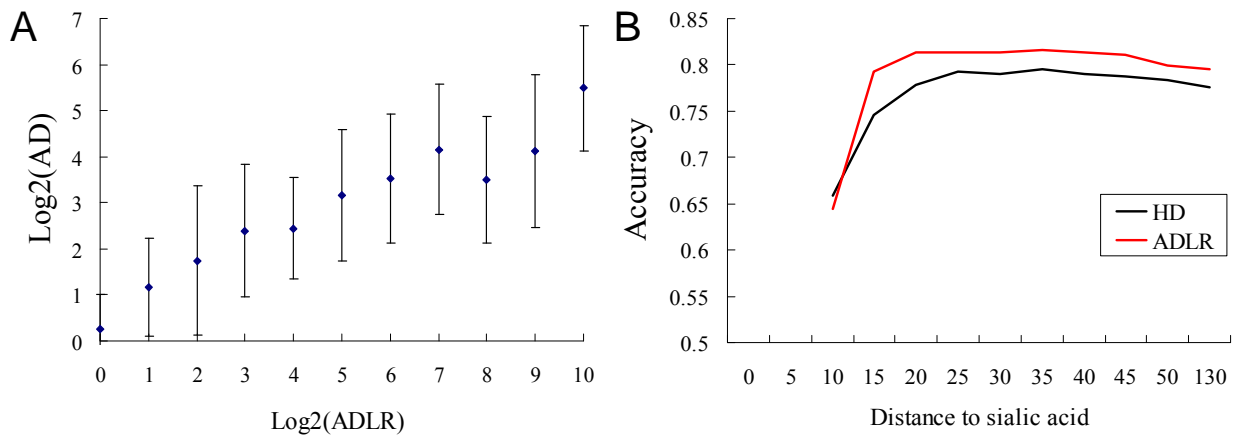


Figure 4.6 The relationships between AD_{LR} and HI assays. (A) The correlation between AD_{LR} and antigenic distance from HI assay. (B) The relationships between accuracy of predicting antigenic variants and models based on amino acid positions at different distance to sialic acid

4.4.3. Antigenic distance for a pair of HA sequences

The antigenic distance from HI assay is the main tool for epidemiologists to quantify the antigenic difference of circulating strains [14]. Furthermore, antigenic distance is correlated to the antigenic evolution of influenza virus [15]. Here, we defined AD_{LR} to quantify the antigenic distance for a pair of HA sequences by naïve Bayesian network [64-65], which based on the LR for each amino acid position. Based on the positions $\leq 20\text{\AA}$ to sialic acid, the correlation between antigenic distance and AD_{LR} is shown in Fig. 4.6A and the Pearson correlation is 0.66.

4.4.4. Predicting antigenic variants

In the global surveillance system, the emerging of antigenic variants often implies that vaccines should be updated to correspond with the dominant epidemic strains [14], thus the predicting of antigenic variants is crucial for vaccine update. The relationships between the accuracy for predicting antigenic variant and the positions with different distance to sialic acid are shown in Fig. 4.6B. The accuracy for AD_{LR} increased steadily from 10\AA to 20\AA positions, at which distance the accuracy reached a stable value at 81.3%. The threshold for $\log_2(AD_{LR})$ to predict a pair of viruses as antigenic variant is 1.8. The minimum number of mutations observed to reach threshold is 2. For example, there are two mutations (135-A and 145-A) between the strains A/Shangdong/9/93 and A/Madrid/252/93 within 20\AA to sialic acid and the value of $\log_2(AD_{LR})$ is 2.0. On the other hand, the accuracy for HD increased steadily from 10\AA to 25\AA positions and reaches a stable value at 79.3%. Based on the LR and entropy of positions that discussed in above section and the predicting accuracy, we selected AD_{LR} and the positions $\leq 20\text{\AA}$ to sialic acid as our model to predicting antigenic variants.

4.4.5. Vaccine-vaccine transitions

The WHO have had updated the component for A (H3N2) influenza vaccine 23 times from the 1968 to 2008 influenza season to ensure the vaccine effectiveness [14, 118]. Vaccine strains are often the dominant strains of influenza seasons, thus it is important to see whether AD_{LR} can detect the transition between vaccine strains. The comparison between successive vaccine strains were listed in Table 4.6. Take the most recent transition for example, there are 6 mutations between the A/Wisconsin/67/2005 strain and A/Brisbane/10/2007 strain within 20\AA to sialic acid

and the $\log_2(\text{AD}_{\text{LR}})$ of three $\text{LR}>1$ positions is 1.88 (140-A, 156-B and 186-B), which is predicted as antigenic variant. Among the 23 transitions, the $\log_2(\text{AD}_{\text{LR}})$ of 18 of them are larger than the threshold and are detectable by our method.

These transitions can be further divided into two classes according to the antigenic type between two viruses. The first class includes 10 virus-pairs, in which the two viruses are both antigenic variants to the other, while the other 13 pairs have at least one antisera that could inhibit the other virus. The *t*-test of comparing these two groups of pairs are 0.04, 0.01 (two-tail) for HD and ADLR, respectively, which means that the pairs in which are antigenic variant to each other usually have higher antigenic distance.

Table 4.6 The HD, AD_{LR} and mutated positions of 23 vaccine-vaccine pairs

Vaccine A	Vaccine B	HD	Mutated positions $\leq 20\text{\AA}$ to sialic acid		$\text{LR} \leq 1$ mutations	Sera A to V_B ¹	Sera B to V_A ²
			\log_2 (AD_{LR})	$\text{LR}>1$ mutations			
Hong Kong/1/68	England/42/72	7	5.4	133, 144, 145, 146, 155, 199	139	V ³	V
England/42/72	Port Chalmers/1/73	3	2.1	160, 188, 193		V	S
Port Chalmers/1/73	Victoria/3/75	9	7.8	137, 145, 157, 160, 189, 193, 201, 21, 230		V	V
Victoria/3/75	Texas/1/77	6	4.8	137, 157, 158, 193, 201, 230		V	V
Texas/1/77	Bangkok/1/79	7	7.2	133, 143, 146, 156, 160, 197, 217		V	S
Bangkok/1/79	Philippines/2/82	5	1.2	144, 182, 196, 248	138	V	S
Philippines/2/82	Mississippi/1/85	6	2.1	144, 156, 182, 196	138, 226	S	S
Mississippi/1/85	Leningrad/360/86	5	2.0	156, 159, 188	138, 226	V	S
Leningrad/360/86	Sichuan/2/87	6	3.6	155, 156, 186, 188, 189	138	V	V
Sichuan/2/87	Shanghai/11/87	3	1.7	156, 186	247	V	S
Shanghai/11/87	Guizhou/54/89	5	2.1	131, 144, 159, 186	247	V	S
Guizhou/54/89	Beijing/353/89	4	3.1	135, 144, 145, 159		V	V
Beijing/353/89	Beijing/32/92	9	5.8	133, 135, 145, 156, 186, 193	190, 214, 226	V	V
Beijing/32/92	Shangdong/9/93	2	1.0	157, 189		V	S
Shangdong/9/93	Johannesburg/33/94	4	0.9	135, 216, 219	214	V	V
Johannesburg/33/94	Wuhan/359/95	7	3.6	135, 145, 197, 216, 219	194, 226	V	V
Wuhan/359/95	Sydney/5/97	6	4.5	133, 142, 144, 156, 158, 196		V	V
Sydney/5/97	Moscow/10/99	5	2.4	137, 142, 160, 196	194	S	S
Moscow/10/99	Fujian/411/2002	13	7.6	75, 131, 144, 155, 156, 160, 186, 196, 202, 222, 225	192, 226	V	V
Wyoming/3/2003 ⁴	Wellington/1/2004	6	1.4	159, 186, 189, 219	226, 227	V	S
Wellington/1/2004	California/7/2004	5	2.2	145, 188, 196	138, 226	V	S
California/7/2004	Wisconsin/67/2005	7	3.9	156, 186, 188, 193, 196, 225	223	V	S
Wisconsin/67/2005	Brisbane/10/2007	6	1.9	140, 156, 186	138, 194, 223	S	S

¹ the antigenic type of vaccine B relative to antisera against vaccine A.

² the antigenic type of vaccine A relative to antisera against vaccine B.

³ the abbreviation of antigenic type, "V" is antigenic variant and "S" is similar virus.

⁴ the HI assay of recommended virus (Fujian/411/2002) were not available, so the Fujian/411/2002 like virus was used instead.

4.4.6. Antigenic cluster change

We considered the AD_{LR} together with the data on antigenic cluster change. Smith *et al.* [15] proposed 11 antigenic clusters of H3N2 virus based on 253 virus strains and HI assays. One of WHO surveillance network's purpose is to detect the emergence and spread of antigenic variant that may signal a need to update the formulation of the influenza vaccine [14-15]. We modeled the detection of emerging antigenic variants by measuring the match of WER strain and 253 strains and the results are shown in Fig 4.7. The WER strains are the dominant antigenic type in each influenza season and the summary are listed in Table 4.3. The average change of AD_{LR} from 1968 to 2003 is shown in Fig. 4.7B, in which the AD_{LR} captures the replacements between antigenic clusters. For example, the H3N2 virus first emerged in human population in 1968 (HK68 cluster) and then the AD_{LR} increased steadily from 1968 to 1972, in which year the second antigenic cluster (EN72) emerged. Unlike the steadily pattern, the third antigenic cluster (VI75) suddenly emerged in 1975. The two kinds of patterns agreed with Shih's observations [41], which suggested that positive selection has been ongoing most of the time and sometimes multiple mutations at antigenic sites cumulatively enhance antigenic drift.

Figure 4.7B also illustrates the antigenic evolution of 5 segments surrounded the RBS. The first changed segments in 1968 were segment I and II that belongs to epitope A. Then the interactions between first 4 segments drove cluster change from 1968 to 2001, which suggested that the interactions between epitope A and B dominated the antigenic drift before 2001 in these segments. The segment V, that includes the two non-epitope positions (222 and 225) was conserved until 2002 in this dataset and interacted with segment IV in the SY97 to FU02 antigenic cluster transition [15]. The variant ratio is shown in Fig. 4.7A, in which the AD_{LR} detected the ratio of antigenic variants against the vaccine strain in each year. The variant ratio usually shows peaks between antigenic cluster transitions, moreover, there are several peaks within the cluster, such as the HK68, BA79 and SY97 clusters. For the HK68 cluster, in which the H3N2 virus first entered human population, the high variant ratio implies the virus underwent many changes for rapid adaptation to a new host. Although all the sequences during 1979 to 1986 are clustered into BA79 cluster, the AD_{LR} detected the variant ratio as 100% in 1984, in which year there was only one sequence (A/Caen/1/84) and virus-pair A/Philippine/2/82 and A/Caen/1/84 is an antigenic variant according to the HI assay, which shows that AD_{LR} can detect the within cluster change. Again, within the SY97 cluster, the AD_{LR} detected the emerging of A/Mowcow/1/99, which is the vaccine strain that replaced the vaccine strain

A/Sydney/5/97, in the year of 1999. The detail information of BA79 and SY97 cluster are shown in Fig. 4.8.

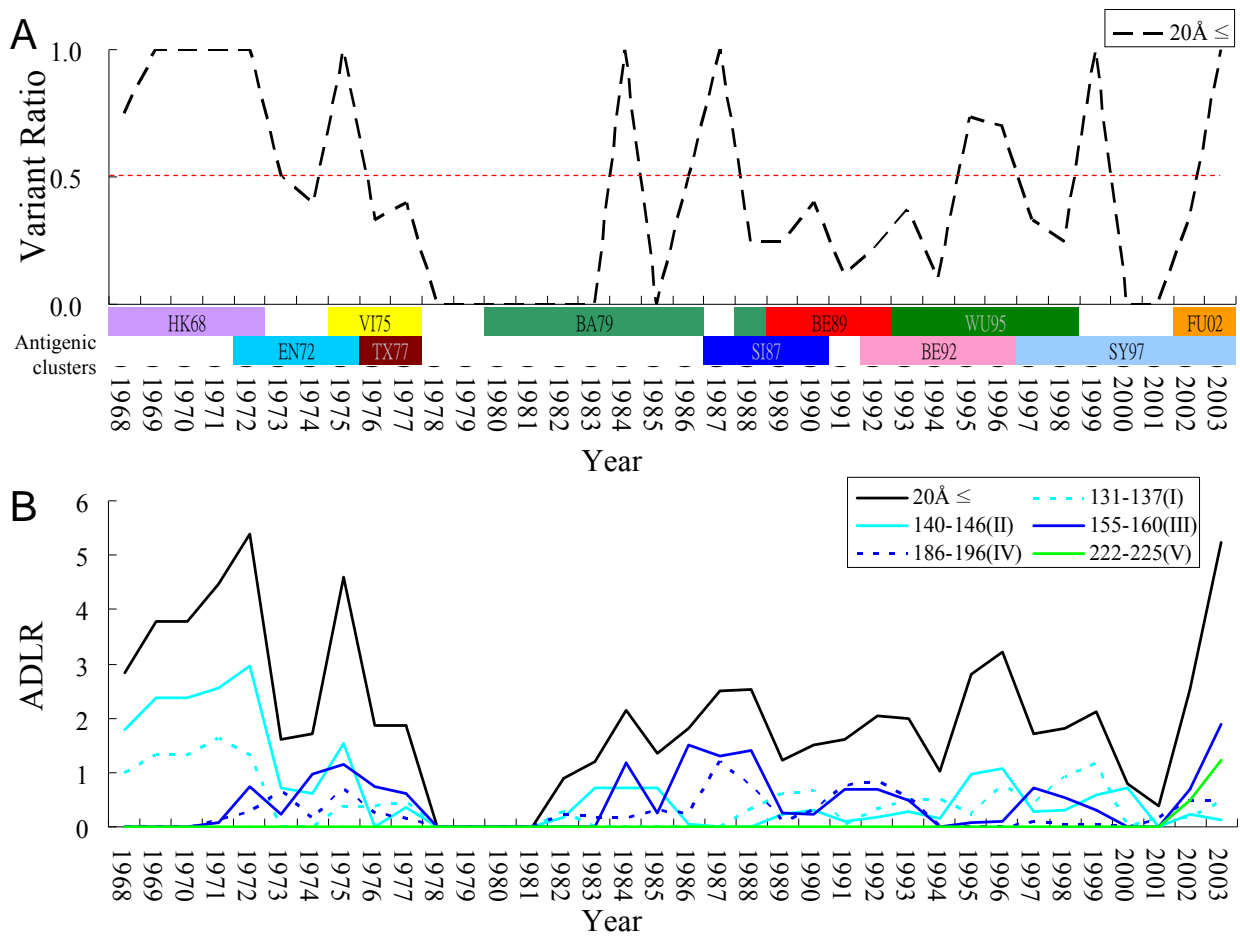


Figure 4.7 The distribution of AD_{LR} and the antigenic drift from 1968 to 2003. (A) The distributions of variant ratios of antigenic clusters from 1968 to 2003. (B) The average AD_{LR} from 1968 to 2003.

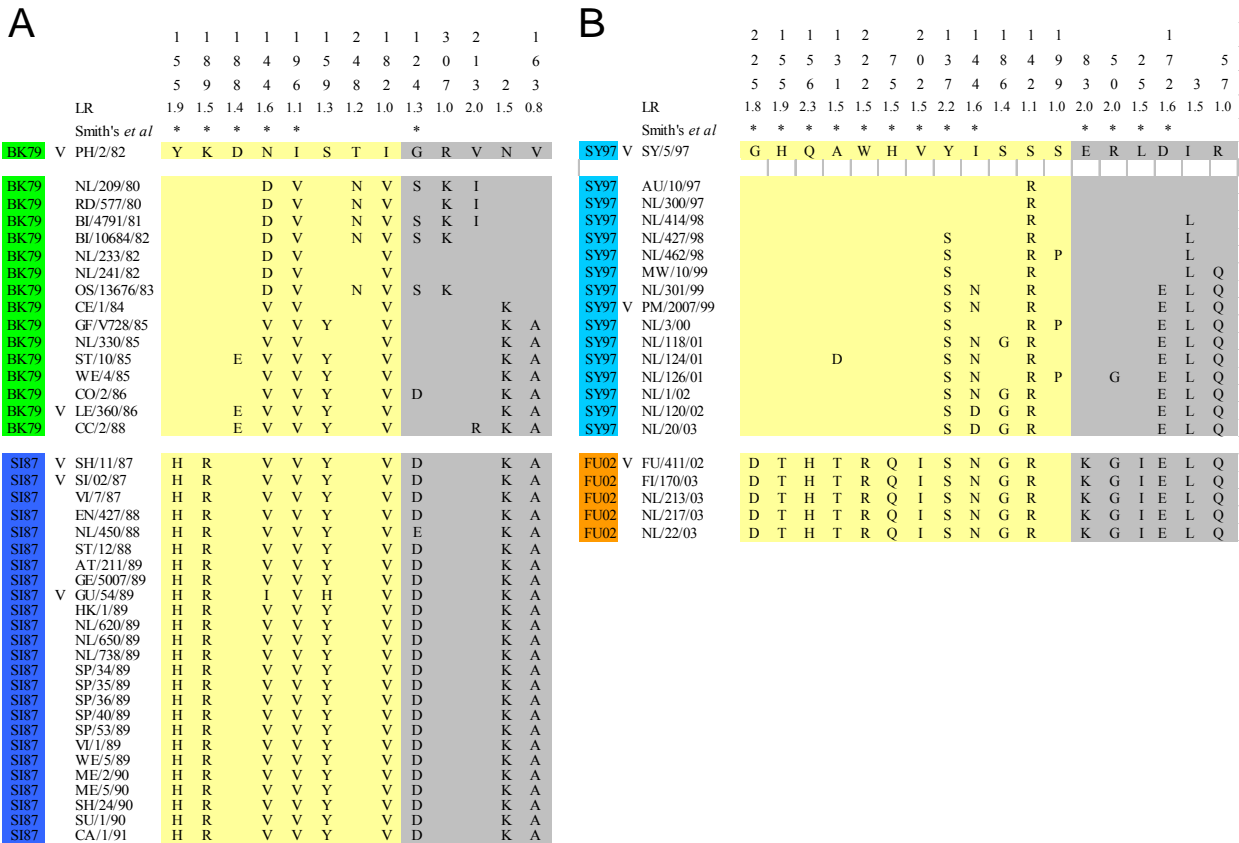
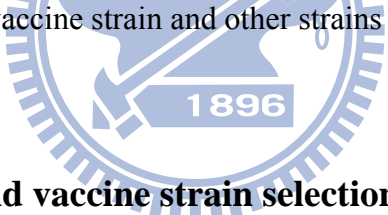


Figure 4.8 The comparison of vaccine strain and other strains in BK79 and SY97 cluster [15].



4.4.7. Antigenic drift and vaccine strain selection

We considered the AD_{LR} together with the data on antigenic drift. The WHO has updated the component for A (H3N2) influenza vaccine 23 times from the 1968 to 2008 influenza season [14, 118]. We detected the antigenic drift as well as to detect the emerging of antigenic variants, which measures the match of WER strain and circulating strains, and we judged the variants are emerging when the variant ratio ≥ 0.5 . The WER strains are the dominant antigenic type in each influenza season and the summary are listed in Table 4.2. The average change of AD_{LR} from 1982-1983 to 2008 is shown in Fig. 4.9B, in which the AD_{LR} shows that the antigenic variant usually emerges before the season that WER strain replacement. For example, the AD_{LR} between circulating strains and WER strain (A/Sydney/5/97) increased from 1997-1998 to 1999, in which season the mean of AD_{LR} is 2.10 and VR is 0.57, we judged the variants emerged in 1999. Indeed the next WER strain (A/Moscow/10/99) dominated at next influenza season.

Figure 4.9B shows more details of the evolution of 5 segments than Fig. 4.8B. For example,

the segment V started to mutate in 1999-2000 season in 2,789 HA sequences, while the segment started to mutate in 2002 in Smith's 253 sequences. There are 7 remarkable peaks for AD_{LR} before the season of 2004 (1987-87, 1989-90, 1991-92, 1995-1996, 1999, 2002-2003 and 2004) and 5 of them followed by the dominant of new WER strains that in different antigenic clusters. This result implies that antigenic cluster change have higher AD_{LR} than the update of other WER strain. Furthermore, the variant ratio for 14 seasons are larger or equal to 0.5 and 12 of them followed by a WER strain replacement in next season, which suggested that AD_{LR} can detected the emerging of antigenic variants.

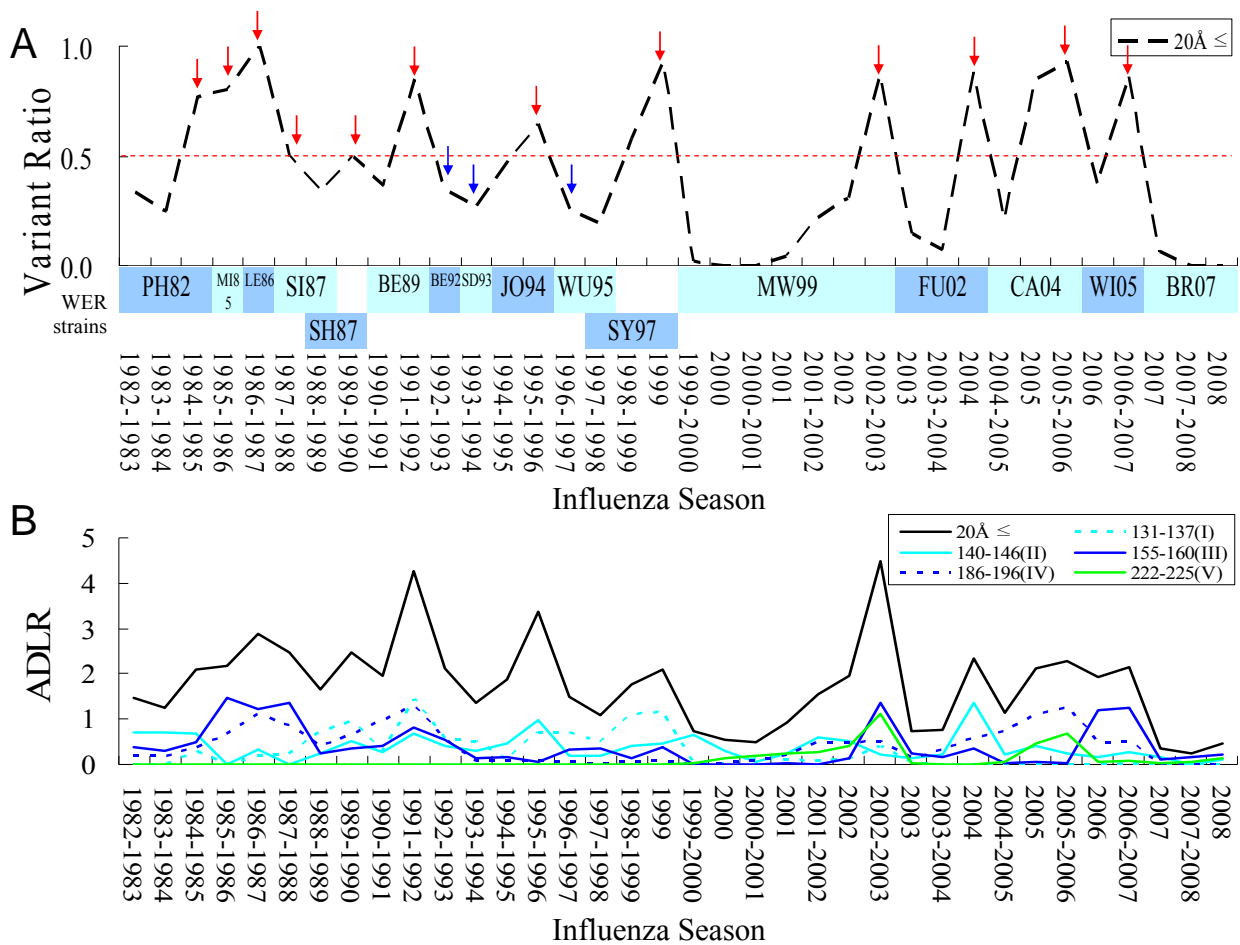


Figure 4.9 The distribution of AD_{LR} and the antigenic drift from 1982-1983 to 2008 influenza season. (A) The distributions of variant ratios of WER strains from 1982-1983 to 2008 season. The match between Model four and WER are labelled (Match in red arrow; Not match in blue arrows). (B) The average AD_{LR} from 1982-1983 to 2008.

4.5. Discussion

The LR quantify the antigenic distance of amino acid position from HI assays and identified 69 positions that with LR larger than 1. Among these positions, there were 6 positions that are almost conserved from 1968 to 2000 and underwent frequency switch [41] after year 2000 (position 25-Other, 75-E, 140-A, 202-Other, 222-Other and 225-Other). The new emerging positions suggest that the previously conserved positions may become new antibody binding sites and LR can identify these positions from HI assay. Due to the emerging of new mutations, the 131 positions that were previously identified as epitope in year before year 1999 [9, 31] are suggested to be updated to incorporate new positions. These new mutations also indicated that the Bayesian method similar to our method must incorporate new HI assay data to capture the emerging mutations. Recently, Lees *et al.* proposed 109 additional positions to extend the original 131 epitope positions [119], which also suggested the need of update the definition of epitope.

We used the five identified segments, which are located within 20Å to the sialic acid, to model the antigenic evolution and the results showed some interesting patterns. The first pattern is the interactions among ≥ 3 positions in segment III and IV that both located on epitope B, which are observed in the seasons 1985-1986, 1986-87 and 1987-1988 and the 10 vaccine-pairs. In the training set, there are 129 virus-pairs match this pattern and 118 of them are antigenic variant (91%). For example, there are three mutations in the vaccine-pair A/Mississippi/1/85 and A/Leningrad/360/86 (156-B, 159-B and 188-B). The other pattern is the interactions among ≥ 3 positions in three segments, which are observed in 12 vaccine-pairs. In the training set, there are 158 virus-pairs matching this pattern and 138 of them are antigenic variants (87%). For example, there are three mutations in the vaccine-pair A/Wisconsin/67/2005 and A/Brisbane/10/2007 (140-A, 156-B, 186-B).

From the view of occlusion of BRS by antibodies, LR identified positions within 20Å to sialic acid have higher antigenic distance and AD_{LR} correlated to antigenic distance well based on these positions. Most of the positions with $LR > 1$ within this distance are located on epitopes A, B and D (30/35), and previous studies suggest epitopes A and B are more antigenic important than other epitopes. Although the epitopes C and E that located more than 20Å to sialic acid were observed to be recognized by antibodies (PDB code 1EO8 [120] and 1QFU [60]), the neutralizing efficiency of them are lower than epitopes A and B [11]. Recently, Ndifon *et al.* studied the neutralization efficiency of epitopes and proposed that mutations on epitopes C and E

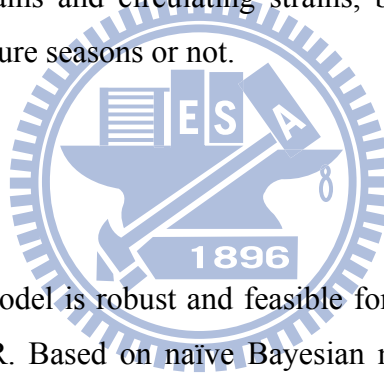
possibly increased the neutralization efficiency [61]. In other words, mutations on these two epitopes possibly increased the efficiency of antibodies neutralization.

Recently, Ekiert *et al.* identified an antibody recognizing a highly conserved epitope among several subtypes of influenza viruses (Subtypes H1, H2, H5, H6, H8 and H9) [63]. This identified antibody presents a new prospective to design the influenza vaccines against diverse subtypes of influenza viruses [121]. The conserved epitope are composed of residues from HA1 and HA2 chains. However, most studies focused on the evolution of HA1 domain and the HA sequences in database often lack the HA2 domain. The proposed index, LR, may provide new insights for the development of influenza vaccine when we consider both HA1 and HA2 sequences.

For the modeling of vaccine update, we proposed the variant ratio, which is an index to detect the emerging of antigenic variants. However, the variant ratio can only measure the degree of match between vaccine strains and circulating strains, but can not determine whether the variants will be dominant in future seasons or not.

4.6. Summary

This study demonstrates our model is robust and feasible for quantifying the antigenic distance of amino acid positions by LR. Based on naïve Bayesian network and LR, we developed an index, AD_{LR} , to quantify the antigenic distance of a given pair of HA sequences. According to the LR values and entropies of positions, we found that the positions locating on the epitopes and near the receptor-binding site are crucial to the antigenic variants. The accumulated critical mutations, which are near ($\leq 20 \text{ \AA}$) to the receptor-binding site, often drive the antigenic drift due to the conformation change to escape from neutralizing antibodies. The AD_{LR} values are highly correlated to the HI assays and can explain the selection of WHO vaccine strains.



Chapter 5

Conclusion

5.1. Summary

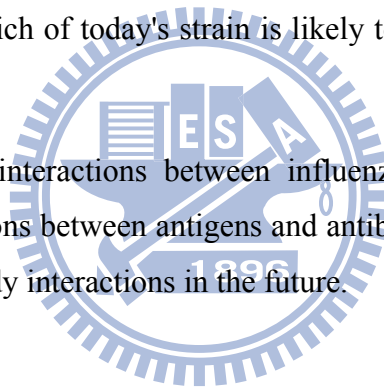
In this thesis, we study the relationships between genetic evolution and antigenic evolution focusing on three dimensions. In short, the major contributions of this thesis can be summarized as follows:

1. We identified critical amino acid positions, rules, and co-mutated positions for antigenic variants. The information gain (IG) and the entropy are used to select critical positions. The co-mutated positions can infer the co-evolution between amino acid positions on HA. The rules, which are derived from the decision tree, describe when one (e.g. circulating) strain will not be recognized by antibodies against another (e.g. vaccine) strain based on a given pair of HA sequences.
2. We developed an epitope-based method for identifying the antigenic drift of influenza A utilizing the conformation changes on antigenic sites (epitopes). Our experimental results show that two critical mutations can induce the conformation change of an epitope. The epitopes (A and B), which are near the receptor-binding site of HA, play a key role for neutralizing antibodies. Two changed epitopes often drive the antigenic drift and can explain the WHO vaccine strain selection.
3. We developed a Bayesian method for identifying the antigenic drift of influenza A by quantifying the antigenic effect of each amino acid position on HA. We utilized the likelihood ratio (LR) and a developed index, AD_{LR} , to quantify the antigenic distance of an amino acid position and a given pair of HA sequences, respectively. Our experimental results show that accumulated critical mutations, which are near ($\leq 20 \text{ \AA}$) the receptor-binding site, often drive the antigenic drift due to the conformation change to evade the recognition by immune system. The AD_{LR} can predict antigenic variants; detect vaccine-vaccine transitions and explain WHO vaccine strain selection.

5.2. Future work

There are several directions for the future work

1. For the antigenic drift, which is related to the recognition between antigen (HA) and antibodies, the structural information should be incorporated to improve the current understanding of the structural change on HA for antigenic drift.
2. Due to the high degree of structural similarity between HA of different subtypes of influenza viruses, our findings on the H3N2 virus can be mapped to other subtypes of influenza viruses (e.g. H1N1 and H5N1 viruses) for comparison. We expect the findings from H3N2 virus to provide new insights for the studies of other subtypes of influenza virus.
3. For the vaccine strain selection, our models for predicting antigenic variants may provide new insights to select which of today's strain is likely to be dominant in the coming year's epidemic
4. Based on our study of interactions between influenza viruses and antibodies, we are interested in the interactions between antigens and antibodies. We may extend our research to general antigen-antibody interactions in the future.



References

1. Stohr K: **Influenza--WHO cares.** *The Lancet infectious diseases* 2002, **2**(9):517.
2. Johnson NP, Mueller J: **Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic.** *Bulletin of the history of medicine* 2002, **76**(1):105-115.
3. Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, Sessions WM, Xu X, Skepner E, Deyde V *et al*: **Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans.** *Science* 2009, **325**(5937):197-201.
4. Webster RG, Bean WJ, Jr.: **Genetics of influenza virus.** *Annual review of genetics* 1978, **12**:415-431.
5. Hayashida H, Toh H, Kikuno R, Miyata T: **Evolution of influenza virus genes.** *Molecular biology and evolution* 1985, **2**(4):289-303.
6. Fouchier RA, Munster V, Wallensten A, Bestebroer TM, Herfst S, Smith D, Rimmelzwaan GF, Olsen B, Osterhaus AD: **Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls.** *Journal of virology* 2005, **79**(5):2814-2822.
7. Neumann G, Noda T, Kawaoka Y: **Emergence and pandemic potential of swine-origin H1N1 influenza virus.** *Nature* 2009, **459**(7249):931-939.
8. McHardy AC, Adams B: **The role of genomics in tracking the evolution of influenza A virus.** *PLoS Pathog* 2009, **5**(10):e1000566.
9. Wilson IA, Cox NJ: **Structural basis of immune recognition of influenza virus hemagglutinin.** *Annual Review of Immunology* 1990, **8**:737-771.
10. Knossow M, Gaudier M, Douglas A, Barrere B, Bizebard T, Barbey C, Gigant B, Skehel JJ: **Mechanism of neutralization of influenza virus infectivity by antibodies.** *Virology* 2002, **302**(2):294-298.
11. Skehel JJ, Wiley DC: **Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin.** *Annual review of biochemistry* 2000, **69**:531-569.
12. Nelson MI, Holmes EC: **The evolution of epidemic influenza.** *Nat Rev Genet* 2007, **8**(3):196-205.
13. Both GW, Sleight MJ, Cox NJ, Kendal AP: **Antigenic drift in influenza virus H3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential amino acid changes at key antigenic sites.** *J Virol* 1983, **48**(1):52-60.
14. Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, Gust ID, Hampson AW, Hay AJ, Hurt AC *et al*: **Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses.** *Vaccine* 2008, **26**:0p.
15. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA: **Mapping the antigenic and genetic evolution of influenza virus.** *Science* 2004, **305**(5682):371-376.
16. Treanor J: **Influenza vaccine--outmaneuvering antigenic shift and drift.** *The New England journal of medicine* 2004, **350**(3):218-220.
17. Gething MJ, Bye J, Skehel J, Waterfield M: **Cloning and DNA sequence of double-stranded copies of haemagglutinin genes from H2 and H3 strains elucidates antigenic shift and drift in human influenza virus.** *Nature* 1980, **287**(5780):301-306.
18. Belshe RB: **The origins of pandemic influenza--lessons from the 1918 virus.** *N Engl J Med* 2005, **353**(21):2209-2211.
19. Kawaoka Y, Krauss S, Webster RG: **Avian-to-Human Transmission of the Pb1 Gene of Influenza-a Viruses in the 1957 and 1968 Pandemics.** *Journal of virology* 1989,

- 63(11):4603-4608.
20. Viboud C, Grais RF, Lafont BAP, Miller MA, Simonsen L, M MIS: **Multinational impact of the 1968 Hong Kong influenza pandemic: Evidence for a smoldering pandemic.** *J Infect Dis* 2005, **192**(2):233-248.
 21. Uyeki TM: **2009 H1N1 virus transmission and outbreaks.** *The New England journal of medicine* 2010, **362**(23):2221-2223.
 22. Carrat F, Flahault A: **Influenza vaccine: the challenge of antigenic drift.** *Vaccine* 2007, **25**(39-40):6852-6862.
 23. WHO: **WHO position paper influenza vaccines.** *The Weekly Epidemiological Record* 2005, **80**:277-288.
 24. WHO: **WHO guidelines on the use of vaccines and antivirals during influenza pandemics.** 2004:5.
 25. Lee MS, Chen JS: **Predicting antigenic variants of influenza A/H3N2 viruses.** *Emerging infectious diseases* 2004, **10**(8):1385-1390.
 26. Finkenstadt BF, Morton A, Rand DA: **Modelling antigenic drift in weekly flu incidence.** *Stat Med* 2005, **24**(22):3447-3461.
 27. Cox NJ, Brammer TL, Regnery HL: **Influenza - Global Surveillance for Epidemic and Pandemic Variants.** *European Journal of Epidemiology* 1994, **10**(4):467-470.
 28. Fitch WM, Bush RM, Bender CA, Cox NJ: **Long term trends in the evolution of H(3) HA1 human influenza type A.** *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94**(15):7712-7718.
 29. Gerdil C: **The annual production cycle for influenza vaccine.** *Vaccine* 2003, **21**(16):1776-1779.
 30. Gupta V, Earl DJ, Deem MW: **Quantifying influenza vaccine efficacy and antigenic distance.** *Vaccine* 2006, **24**(18):3881-3888.
 31. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM: **Predicting the evolution of human influenza A.** *SCIENCE* 1999, **286**(5446):1921-1925.
 32. Huang JW, King CC, Yang JM: **Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses.** *BMC Bioinformatics* 2009, **10** Suppl 1:S41.
 33. Bush RM, Fitch WM, Bender CA, Cox NJ: **Positive Selection on the H3 Hemagglutinin Gene of Human Influenza Virus A.** *Molecular Biology and Evolution* 1999, **16**:1457-1465.
 34. Plotkin JB, Dushoff J, Levin SA: **Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(9):6263-6268.
 35. Macken C, Lu H, Goodman J, Boykin L: **The value of a database in surveillance and vaccine selection.** *Options for the Control of Influenza Iv* 2001, **1219**:103-106.
 36. Bao YM, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D: **The influenza virus resource at the national center for biotechnology information.** *Journal of Virology* 2008, **82**(2):596-601.
 37. Salzberg S: **The contents of the syringe.** *Nature* 2008, **454**(7201):160-161.
 38. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM: **Predicting the evolution of human influenza A.** *Science* 1999, **286**(5446):1921-1925.
 39. Shih AC, Hsiao TC, Ho MS, Li WH: **Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:6283-6288.
 40. Lee MS, Chen JS: **Predicting antigenic variants of influenza A/H3N2 viruses.** *Emerging Infectious Diseases* 2004, **10**(8):1385-1390.
 41. Arthur Chun-Chieh S, Tzu-Chang H, Mei-Shang H, Wen-Hsiung L: **Simultaneous**

- amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(15):6p.
42. Du X, Wang Z, Wu A, S L, C Y, H H, J T: **Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution.** *GENOME RESEARCH* 2007, **18**(178-187).
 43. Simonsen L: **The global impact of influenza on morbidity and mortality.** *Vaccine* 1999, **17**:S3-S10.
 44. Blackburne BP, Hay AJ, Goldstein RA: **Changing selective pressure during antigenic changes in human influenza H3.** *Plos Pathogens* 2008, **4**(5):-.
 45. Quinlan JR: **C4.5: Programs for Machine Learning.** San Mateo, CA: Morgan Kaufmann; 1993.
 46. Sauter NK, Hanson JE, Glick GD, Brown JH, Crowther RL, Park SJ, Skehel JJ, Wiley DC: **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* 1992:9609-9621.
 47. DeLano WL: **The PyMOL Molecular Graphics System.** Palo Alto, CA, USA: DeLano Scientific; 2002.
 48. Gupta V, Earl DJ, Deem MW: **Quantifying influenza vaccine efficacy and antigenic distance.** *Vaccine* 2006, **24**(18):3881-3888.
 49. Taylor HP, Dimmock NJ: **Competitive binding of neutralizing monoclonal and polyclonal IgG to the HA of influenza A virions in solution: only one IgG molecule is bound per HA trimer regardless of the specificity of the competitor.** *Virology* 1994, **205**(1):360-363.
 50. WHO: *Weekly Epidemiological Record* 1970-2007, **45, 46, 48, 49, 50, 51, 52, 55, 57, 58, 60, 61, 62, 63, 65, 66, 67, 68, 69, 70, 71, 73, 74, 78, 79, 80, 71, 82.**
 51. Centers for Disease Control and Prevention: **Information for the Vaccines and Related Biological Products Advisory Committee, CBER, FDA.** 2003-2007:p. 28, p.18, p.19, p.17.
 52. Ellis JS, Chakraverty P, Clewley JP: **Genetic and antigenic variation in the haemagglutinin of recently circulating human influenza A (H3N2) viruses in the United Kingdom.** *Arch Virol* 1995, **140**(11):1889-1904.
 53. Both GW, Sleight MJ, Cox NJ, Kendal AP: **Antigenic Drift in Influenza Virus-H3 Hemagglutinin from 1968 to 1980 - Multiple Evolutionary Pathways and Sequential Amino-Acid Changes at Key Antigenic Sites.** *Journal of Virology* 1983, **48**(1):52-60.
 54. Coiras MT, Aguilar JC, Galiano M, Carlos S, Gregory V, Lin YP, Hay A, Perez-Brena P: **Rapid molecular analysis of the haemagglutinin gene of human influenza A H3N2 viruses isolated in Spain from 1996 to 2000.** *Archives of virology* 2001, **146**(11):2133-2147.
 55. Webster R, Cox NJ, Stohr K: **WHO Manual on Animal Influenza Diagnosis and Surveillance.** *WHO/CDS/CSR/NCS/20025* 2002, **Rev.1.**
 56. Centers for Disease Control and Prevention: **Information for FDA vaccine advisory panel meeting. Atlanta: The Centers.** 1997:p. 30.
 57. Centers for Disease Control and Prevention: **Options for Live Attenuated Influenza Vaccines (LAIV) in the Control of Epidemic and Pandemic Influenza, Geneva, 12-13 June, 2007.** 2007:p. 11.
 58. Barbey-Martin C, Gigant B, Bizebard T, Calder LJ, Wharton SA, Skehel JJ, Knossow M: **An antibody that prevents the hemagglutinin low pH fusogenic transition.** *Virology* 2002, **294**(1):70-74.
 59. Bizebard T, Gigant B, Rigolet P, Rasmussen B, Diat O, Bosecke P, Wharton SA, Skehel

- JJ, Knossow M: **Structure of influenza virus haemagglutinin complexed with a neutralizing antibody**. *Nature* 1995, **376**(6535):92-94.
60. Fleury D, Barrere B, Bizebard T, Daniels RS, Skehel JJ, Knossow M: **A complex of influenza hemagglutinin with a neutralizing antibody that binds outside the virus receptor binding site**. *Nat Struct Biol* 1999, **6**(6):530-534.
61. Ndifon W, Wingreen NS, Levin SA: **Differential neutralization efficiency of hemagglutinin epitopes, antibody interference, and the design of influenza vaccines**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(21):8701-8706.
62. Tsurudome M, Gluck R, Graf R, Falchetto R, Schaller U, Brunner J: **Lipid Interactions of the Hemagglutinin Ha2 Nh2-Terminal Segment during Influenza Virus-Induced Membrane-Fusion**. *J Biol Chem* 1992, **267**(28):20225-20232.
63. Ekiert DC, Bhabha G, Elsliger MA, Friesen RHE, Jongeneelen M, Throsby M, Goudsmit J, Wilson IA: **Antibody Recognition of a Highly Conserved Influenza Virus Epitope**. *Science* 2009, **324**(5924):246-251.
64. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data**. *Science* 2003, **302**(5644):449-453.
65. Wang Y, Zhang XS, Xia Y: **Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data**. *Nucleic acids research* 2009, **37**(18):5943-5958.
66. Dujardin B, Van den Ende J, Van Gompel A, Unger JP, Van der Stuyft P: **Likelihood ratios: a real improvement for clinical decision making?** *European Journal of Epidemiology* 1994, **10**(1):29-36.
67. WHO: *Weekly Epidemiological Record* 1970, **45**:100.
68. WHO: *Weekly Epidemiological Record* 1971, **46**:517, 518.
69. WHO: *Weekly Epidemiological Record* 1972, **45**:82, 381.
70. WHO: *Weekly Epidemiological Record* 1973, **48**:389, 468.
71. WHO: *Weekly Epidemiological Record* 1974, **49**:44, 107.
72. WHO: *Weekly Epidemiological Record* 1975, **50**:52, 139.
73. WHO: *Weekly Epidemiological Record* 1976, **51**:41.
74. WHO: *Weekly Epidemiological Record* 1977, **52**:39.
75. WHO: *Weekly Epidemiological Record* 1980, **55**:74.
76. WHO: *Weekly Epidemiological Record* 1982, **57**:58.
77. WHO: *Weekly Epidemiological Record* 1983, **58**:54, 55.
78. WHO: *Weekly Epidemiological Record* 1985, **60**:53, 54.
79. WHO: *Weekly Epidemiological Record* 1986, **61**:62.
80. WHO: *Weekly Epidemiological Record* 1987, **62**:90.
81. WHO: *Weekly Epidemiological Record* 1988, **63**:58.
82. WHO: *Weekly Epidemiological Record* 1990, **65**:54.
83. WHO: *Weekly Epidemiological Record* 1991, **66**:58.
84. WHO: *Weekly Epidemiological Record* 1992, **67**:58.
85. WHO: *Weekly Epidemiological Record* 1993, **68**:58.
86. WHO: *Weekly Epidemiological Record* 1994, **69**:54.
87. WHO: *Weekly Epidemiological Record* 1995, **70**:54.
88. WHO: *Weekly Epidemiological Record* 1996, **71**:58, 59.
89. WHO: *Weekly Epidemiological Record* 1998, **73**:57, 58.
90. WHO: *Weekly Epidemiological Record* 1999, **74**:58, 322, 323.
91. WHO: *Weekly Epidemiological Record* 2003, **78**:59, 377.
92. WHO: *Weekly Epidemiological Record* 2004, **79**:89, 371.

93. WHO: *Weekly Epidemiological Record* 2005, **80**:72, 73, 344.
94. WHO: *Weekly Epidemiological Record* 2006, **81**, **392**:83, 84.
95. WHO: *Weekly Epidemiological Record* 2007, **82**:71, 353.
96. Centers for Disease Control and Prevention: **Information for the Vaccines and Related Biological Products Advisory Committee, CBER, FDA**. 2003:p. 28.
97. Centers for Disease Control and Prevention: **Information for the Vaccines and Related Biological Products Advisory Committee, CBER, FDA**. 2004:p. 18.
98. Centers for Disease Control and Prevention: **Information for the Vaccines and Related Biological Products Advisory Committee, CBER, FDA**. 2005:p. 19.
99. Centers for Disease Control and Prevention: **Information for the Vaccines and Related Biological Products Advisory Committee, CBER, FDA**. 2006:p. 17.
100. Centers for Disease Control and Prevention: **Information for the Vaccines and Related Biological Products Advisory Committee, CBER, FDA**. 2007:p. 19.
101. WHO: *Weekly Epidemiological Record* 1984, **59**:54.
102. WHO: *Weekly Epidemiological Record* 1989, **64**:54.
103. WHO: *Weekly Epidemiological Record* 1997, **72**:58.
104. WHO: *Weekly Epidemiological Record* 2000, **75**:62, 331.
105. WHO: *Weekly Epidemiological Record* 2001, **76**:59, 312.
106. WHO: *Weekly Epidemiological Record* 2002, **77**:63, 346.
107. WHO: *Weekly Epidemiological Record* 2008, **83**:83, 369.
108. WHO: *Weekly Epidemiological Record* 1968, **43**:448.
109. WHO: *Weekly Epidemiological Record* 1969, **44**:619.
110. WHO: *Weekly Epidemiological Record* 1978, **53**:67.
111. WHO: *Weekly Epidemiological Record* 1979, **54**:69.
112. WHO: *Weekly Epidemiological Record* 1981, **56**:58.
113. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment**. *Science* 1993, **262**(5131):208-214.
114. Tong JC, Song CM, Tan PT, Ren EC, Sinha AA: **BEID: database for sequence-structure-function information on antigen-antibody interactions**. *Bioinformation* 2008, **3**(2):58-60.
115. Chen YC, Lo YS, Hsu WC, Yang JM: **3D-partner: a web server to infer interacting partners and binding models**. *Nucleic acids research* 2007, **35**(Web Server issue):W561-567.
116. Pollastri G, Baldi P, Fariselli P, Casadio R: **Prediction of coordination number and relative solvent accessibility in proteins**. *Proteins* 2002, **47**(2):142-153.
117. Weis WI, Brunger AT, Skehel JJ, Wiley DC: **Refinement of the influenza virus hemagglutinin by simulated annealing**. *J Mol Biol* 1990, **212**(4):737-761.
118. Hay AJ, Gregory V, Douglas AR, Lin YP: **The evolution of human influenza viruses**. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2001, **356**(1416):1861-1870.
119. Lees WD, Moss DS, Shepherd AJ: **A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2**. *Bioinformatics* 2010, **26**(11):1403-1408.
120. Fleury D, Daniels RS, Skehel JJ, Knossow M, Bizebard T: **Structural evidence for recognition of a single epitope by two distinct antibodies**. *Proteins-Structure Function and Genetics* 2000, **40**(4):572-578.
121. Steel J, Lowen AC, Wang T, Yondola M, Gao Q, Haye K, Garcia-Sastre A, Palese P: **Influenza virus vaccine based on the conserved hemagglutinin stalk domain**. *MBio* 2010, **1**(1).

Appendix A

List of Publications

● Journal papers

1. **J.-W. Huang**, C.-C. King and J.-M. Yang*, "Co-evolution positions and rules for antigenic variants of influenza A/H3N2 viruses," *BMC Bioinformatics*, vol. 10 (Suppl 1):S41, 2009
2. K.-P. Liu, K.-C. Hsu, **J.-W. Huang**, L.-S. Chang and J.-M. Yang*, "ATRIPPI: An Atom-Residue Preference Scoring Function for Protein-Protein Interactions," *International Journal on Artificial Intelligence Tools*, vol. 19, pp. 251-266, 2010
3. C.-H. Tung, **J.-W. Huang** and J.-M. Yang*, "Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for fast protein structure database search," *Genome Biology*, vol. 8, pp. R31.1~R31.16, 2007
4. **J.-W. Huang**, and J.-M. Yang*, " A Bayesian approach for quantifying antigenic distance of influenza A (H3N2) viruses ," (Preparing)

● Conference papers

1. **J.-W. Huang**, and J.-M. Yang*, " Changed epitopes drive the antigenic drift for influenza A (H3N2) viruses," (Submitted)
2. **J.-W. Huang**, C.-C. Chen, J.-M. Yang "Identifying critical positions and rules of antigenic drift for influenza A/H3N2 viruses," *The 2nd International Conference on Bioinformatics and Biomedical Engineering*, pp. 249-252, 2008