國立交通大學

生物資訊研究所

碩士論文

藉由建立基因演化與抗原性漂移之關聯性

預測 A 型 H3N2 流行性感冒病毒之抗原性變異

Predicting Antigenic Variants of Influenza A H3N2 Viruses by

Building Relationships between Genetic Evolution and Antigenic Drift

研 究 生：黃章維

指導教授：楊進木　教授

中 華 民 國 九 十 五 年 六 月

藉由建立基因演化與抗原性漂移之關聯性

預測 A 型 H3N2 流行性感冒病毒之抗原性變異

學生:黃章維　　指導教授: 楊進木博士

國立交通大學 生物資訊研究所碩士班

Abstract

具有疾病性之禽類與人類流行性感冒病毒曾對人類文明社會帶來嚴重的傷害與經濟損失，因此了解流感病毒之抗原性演化對於預防流感與疫苗株之挑選是很重要的議題。大多數的相關研究在預測抗原性演化與預測未來造成流行之病毒株時只統計位於紅血球凝集素(HA)上之突變點數與使用演化式之分析方法。近來有幾份研究發現位於血球凝集素上之突變點數量與抗原-抗體親和力有關聯性，換句話說，發現了基因演化與抗原性演化之關聯性。此發現顯示抗原性演化比基因演化更具有不連續之跳躍性，且基因序列上的改變有時會造成不等價之鉅大抗原性影響。

在這份論文中，我們研究的重要議題是"位於HA的序列中，那些重要位置的改變會與HI滴定量改變有高度的相關性"。資訊獲得量被用來衡量並且代表基因演化與抗原性演化之關聯性。位於HA序列上之一個胺基酸位置若具有高的資訊獲得量則表示發生在此位置上之點突變會與代表抗原性特性之血球凝集抑制抗體效價高度相關。此顯示了每個位置的資訊獲得量可以用來預測HA序列上之基因改變與抗原性改變之相關性。決策樹方法(C4.5)根據資訊獲得量被用來選擇21個重要的位置。這21個位置被進一步分成6群，每一群內高度相關之位置具有共同演化之特性。根據每個位置之資訊獲得量與共同演化之資訊，在研究中建立了一個模組來預測基因演化與抗原性演化之關聯性。

我們的方法分別使用序列上之特徵值與結構上之特徵值(Contact Map)，此兩者在訓練模組之預測率分別91%與96%。此方法在同一組資料集上之預測率比傳統使用漢明距離法具有較高的預測率。大部分我們找到重要的位置都落在Epitope上並且與之前的相關研究有一致性。 最後該預測模組(使用資訊獲得量所選擇之重要位置)被應用於2個測試資料上。對於WER之50筆疫苗株資料之預測率為74%，對於5928筆歷史資料之預測率為87%並且能成功地預測流感病毒群體間之轉移(99%)。由以上的結果，顯示我們的方法具有robust之特性並且有助於預測基因與抗原性演化之關聯性，此方法亦具潛力助於疫苗發展。

Predicting Antigenic Variants of Influenza A H3N2 Viruses by Building Relationships between Genetic Evolution and Antigenic Drift

Student: Jhang-Wei Huang Advisor : Dr. Jinn-Moon Yang

Institute of Bioinformatics
National Chiao Tung University

Abstract

Pathogenic avian and human influenza virus could cause disastrous damage to human society and economics. Understanding antigenic evolution of influenza viruses is a very important issue for vaccine strain selection and prophylaxis. To predict antigenic drift most current approaches use only hemagglutinin protein (HA) sequences of influenza by number of mutations and phylogenetic analyses to select viruses which will probably be the progenitor of viruses in the next epidemic. Recently, several reports had indicated that there were relationships between mutations of HA protein sequences and antigen-antibody affinity, i.e., the relationships between the viral genetic evolution and antigenic drift. They observed that antigenic drift was more punctuated than genetic evolution, and genetic changes sometime had a disproportionately large antigenic effect.

In this thesis, we study an important issue: “whether certain amino acid positions change in the HA protein sequences are correlated to the change of binding HI titer values”. The information gain is used to calculate the degree of association between the genetic evolution and antigenic drift. An amino acid with high information gain at a specific position (i.e., 1 ~ 329 positions for a HA sequence) means that amino acid mutation on this position is highly correlated to antigenic change on HI titer value. This implied that the value of information gain in each position is able to predict the association between genetic and antigenic change for HA sequences. Here, a decision tree tool (C 4.5) was used to select 21 important positions based on information gain. These 21 positions are further clustered into 6 groups and the amino acid positions on the same cluster are high co-evolution. According to the information gain of each position and co-evolution, we have built a model to predict the association between the genetic and antigenic evolution.

Our method yielded both sequence features (amino acid position changes) and structure features (contact maps). The accuracies of our model were 91% and 96% by using sequence and structure features, respectively. The accuracy is much better than a traditional hamming distance method on the same data set. Most of the critical positions identified by our method are located on the epitope sites and are consistent with previous works. Finally, the predicted model (critical positions selected by information gain) was applied on two test sets. The predicting accuracy for 50 cases from WER vaccine strains was 74% and for 5928 historical real cases was 87%. These results demonstrate that our approach is robust and useful for predicting the relationship between genetic evolution and antigenic drift and is potential useful for vaccine development.

Acknowledgements

The most appreciation is for my advisor Dr. Jinn-Moon Yang. He always teaches us how to do a research project although which took a lot of his time. After a four-year training in BIOXGEM lab, I finally start to realize what research is and know that there still many works needed to be done. I think sincere interest and perseverance are the key factors for conducting a high quality research work and Dr. Jinn-Moon Yang is the one who perfectly matches the criterion ☺

I also want to thank Dr. Chwan-Chuen King. She initiated our group to the research of influenza field and thank for her spending a lot of time to discuss with us and sharing her viewpoints to this research. I hope this work could finally help to understand the genetic and antigenic evolution of human and avian influenza viruses and more practical.

Finally I want to thank Mr. Chun-Chen Chen for many of his sincere help. Parts of this work are done through our collaboration.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Influenza A viruses is a negative-stranded RNA virus which can cause epidemic, is common acute respiratory diseases. And influenza A has the potential to trigger pandemic infection. In Temperate Zone, influenza affected 1%-5% human population. Children were infected most easily, but the infected elders were at the highest risk of complication and death. In industrialized countries, influenza epidemics induced severe damages in economics. Each million people shared social cost of 10-60 million US dollars ([1]). Shortly, prevention and therapy of influenza are very important issues in medicine.

1.2 Motivations and Purposes

Since influenza A virus cause such important epidemic and economic impacts to human beings. The prevention of influenza virus has significant importance. Current strategy for prevention influenza virus is vaccination ([2]). The vaccination with the inactivated influenza vaccines can provide protection when the vaccine strain vaccine antigens and circulating strains share a high degree of similarity in antigenic property of hemagglutinin (HA) protein. But gradual mutations to the HA gene continually produce immunologically distinct strains.

Immune responses from infection by one influenza virus may not protect fully against antigenic or genetic variants of the same subtype (influenza A viruses). As a consequence, influenza outbreaks could occur every year. New influenza vaccine strain must be selected annually to match the circulating viruses. In order to help the selection of future vaccine candidate, to understanding and predicting the antigenic and genetic evolution pattern of the surface antigen HA are desired.

Currently there are abundant sequences data and HI titer values which are available from public databases. By the means of analyzing HA sequences and HI titers, we want to answer the important question: which amino acid positions change would have significant effects on host immune response in predicted HI titer values. In other words we want to build relationships between genetic and antigenic evolution.

1.3 Hemagglutinin and Epitope

The two surface glycoproteins Hemagglutinin (HA) and neuraminidase (NA) on the influenza virus are the most important targets for the human immune system [3]. Gradually mutations of HA gene produce immunologically distinct strains of the virus that cause annual epidemic outbreaks. Since the HA are the surface protein of influenza virus, HA is the key component of current influenza vaccine ([1]). From 1968 to 2006 the vaccine component have 22 times of changes (Fig. 1) (Table 1) .The HA protein consists of two chains, HA1 and HA2,

respectively 329 and 175 residues long. Mostly works focus on the HA1, which is the immunogenic part of HA. The protein 3D structure of HA is determined and deposit in PDB (ex: 1HGF).

There are 15 different subtypes of HAs among avian, which need HI test to be identified.

The definition of Epitope is the particular site within a macromolecule to which a specific antibody binds . Since HA is surface antigen protein of influenza virus, the epitope to which antibody binds is important to the immune system. Since the protein structure is determined and deposit in the PDB (Ex 1HGF 1KEN). The epitope sites could be determined. There are five epitope sites (A B C D E) contain 131 residues in the full HA1 329 residues [4, 5]. These five antigenic sites show structure clusters in the 3D space (Figure 2).

1.4 Hemagglutination Inhibition Test

The purpose of this test is to determine the function of serum antibody to inhibit the abilities for hemagglutination of flu virus. In the WHO Influenza manual [6], this test can be used to identify influenza isolates since there are 15 different subtype of HA among avian if one uses standardized HA antibody. After a serial HI tests, the result often record in a HI table (Figure 3). The list of first row represents different antisera and the list of first column represents antigen. The value in the table means number folds of dilution. If the value

between antisera A and antigen b is 320 (Figure 3), that means after a 320 folds serial dilution the antisera still can completely inhibit the hemagglutination.

The HI test is extremely reliable [6], provided reference antisera are available to all subtypes. Disadvantages of the HI test include the need to remove nonspecific inhibitors which naturally occur in sera, to standardize antigen each time a test is performed, and the need for specialized expertise in reading the results of the test. However, the HI assay remains the test of choice for WHO global influenza surveillance.

1.5 Related Works

The final purpose of all the influenza researches hope to answer the following questions: when to update vaccine? and how to choose future vaccine candidate [1]? To the purpose, there are many approaches to answer these two questions. Either from experiment or computational approach, scientists collect abundant data and want to discover the pattern or evolution trend of the influenza virus [7-10]. In the view of whether they considering HI titer value, they could be classified into two approaches. First kind of approach focus on the genetic evolution of HA protein [8, 9], and the second kind combine experiment data which could further consider the antigenic evolution of the HA protein [11, 12].

In the genetic level, there had been discovered that those sites of HA1 involved in antigen determination exhibit significantly more non-synonymous nucleotide substitutions

than synonymous substitutions [8], whereas the remaining sites show the more common pattern of primarily synonymous variation. These observations demonstrate that HA is undergoing positive Darwinian selection for new antigenic variants [13]. Bush *et al.* [8] have identified 18 HA1 codon sites with significantly higher non-synonymous to synonymous ratios.

In order to analysis the evolutional pattern of influenza virus, there had been propose a cluster method [10], which cluster 560 HA protein sequences into 174 clusters. According to the cluster result, there are several representative clusters. By the means of compare genetic variation between intra and inter of representative clusters, they found some evolution trend of influenza virus. They also proposed a method to predict the future vaccine candidate.

Before the year of 2004, mostly works focus on the genetic level on HA. Until the year of 2004, there began to have more efforts made on the comparison between genetic and antigenic evolution. The result shows that gradual genetic evolution, but punctuated antigenic evolution [11]. As a result they found that the genetic evolution could not directly correspond to antigenic evolution. Genetic change sometimes had a disproportionately large antigenic effect. The next question should be what are the relations between genetic and antigenic evolution.

By collect historical WHO vaccine HI titer tables and HA sequence from 1968 to 2002, a global prediction model is build [12]. The highest performance model for predict antigenic

variant shows that when there are more than 7 amino acid changes on the epitope sites then a antigenic variant strain is predicted (agreement = 83%). But the importance of these positions in terms of affecting cross-reactive antibody is unclear.

In order to find what key position changes would affect cross-reactive antibody interaction. We apply an index value from information theory. The information gain evaluates the relation between two variables (genetic and antigenic evolution). Here we take information gain as a index to represent relations between genetic and antigenic evolution. We hope to find out antigenically important positions and to understand the pattern between genetic and antigenic evolution.

Chapter 2

Materials and Methods

2.1 Overview of Research Steps

The research flowchart could be divided into two parts (Fig 4). In the first part we calculate the information gain of 329 HA positions from on a representative training dataset and evaluate the fitness for information gain to represent the relations between genetic and antigenic evolutions. Then in the second part we apply the important positions selected by information gain to predict antigenic variants on two unseen and meaningful application sets (test sets).

In the first part we first select a representative training set which was used in a published work [12]. Then we extract features from sequences and HA protein structure. The HI titers are transformed from folds of serial dilution to an antigenic distance between two influenza viruses. The large antigenic distance means more antigenic difference between two viruses. After we have two variables (genetic features and antigenic distance), we could calculate information gain for each 329 HA positions. By a well-known method (Decision Tree C4.5) based on information gain we could select several clusters of important positions and get a training model for predicting antigenic variants. After we found important positions, we discuss the fitness for information gain to represent the relations between genetic and

antigenic evolutions. Those selected positions are then used to predict antigenic variants and compare predicting performance to related works.

In the second part we find two unseen test sets which have antigenic properties. The first smaller set (51 cases) were all vaccine strains extracted from WER (1968~2006) and each case with known HI titer value. The second larger set (5928 cases) containing 181 influenza viruses from 1968 to 2003 which having an antigenic clustering label [11]. Then we apply the position and rules from training model on these two test sets.

In the following part we would first show that how materials are prepared and then the detail of methods.

2.2 Influenza Sequence Database

The influenza sequence database [14] is a well-known and frequently cited database, which collect the nucleotide sequence of influenza virus. They collect all 3 influenza species and 8 protein segments of various hosts (Appendix). This difference between NCBI database the ISD is that ISD deposit not only publish sequence but also un-publish sequences. The ISD also provide some useful information such as vaccine selection from 1999 to 2006 (Appendix I) and influenza virus activity in United States from 1981 to now (Appendix I) .Since all sequence is presented in nucleotide format, the translation is required. The EBI translation tool is recommended (http://www.ebi.ac.uk/emboss/transeq/).

2.3 Training Set

We need a representative and robust training set which should including representative influenza virus strain and the set should better to be complete and balanced. From the literature search we choose a set which was used in a related work [12]. This set consisted of six sets ferret serum HI cross-reactivity data which including 45 influenza virus strains and 181 pairwise ferret serum HI titers (Table 2). From 1968 to 2005 there were 21 influenza virus strains treat as WHO vaccine component (Table 1), and this set cover 17 virus strains of them.

The first set included 11 viruses (55 pairwise comparisons, virus ID: A to K) isolated from 1971 to 1979 [15]. The second set included 8 viruses (28 pairwise comparisons, virus ID: J, L to R) isolated from 1979 to 1987 [16]. The third set included 10 viruses (45 pairwise comparisons, virus ID: S to AB) isolated from 1989 to 1994 [17]. The fourth set included 8 viruses (28 pairwise comparisons, virus ID: AC to AJ) isolated from 1994 to 1996 [18]. The fifth set included 5 viruses (10 pairwise comparisons, virus ID: AE, AK to AN) isolated from 1995 to 1999 [19]. The sixth set included 6 viruses (15 pairwise comparisons, virus ID: AN to AT) isolated from 1999 to 2002 [20]. (Note : the strain TOK75's position 226 x is assign to amino acid Leucine, which is identical to other residue in table one. The sequence need manually key in table one is using template J02135). The information of all the sequences is listed on table 3.

After the feature extraction and calculation of antigenic distance, the training set have

181 cases and 125 of them are variant type (antigenic distance $\geqq 4$) while the other 56 cases are equal type(antigenic distance < 4) . Among all the 329 residues of HA protein, there are 101 positions have occurred change in this set.

2.4 Test Set

The purpose of test set is to evaluate the correctness of the positions and rules learning from the training set. Since our method integration both genetic and antigenic evolution, the test set should also containing antigenic property.

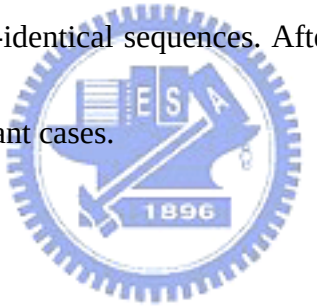
The first set was extracted from WER (1968~2006), from which we could found 62 reference pairs of HI titer value with both homologous and heterologous titer values and available HA sequences. We further filter these 62 cases with one more criterion: there should have at lease one vaccine strain for the pair comparison. Finally we could got 50 cases satisfy the condition.

The second dataset includes 253 sequences. All 253 sequences were grouped into 11 groups according to the K-mean result which using antigenic distances transform from HI titer [11] (Fig 5). After a simple all pairwise comparison, we identify 181 non-identical sequences treated in the test set (Table 4). Since the cluster result were based on antigenic data, we assume that two different cluster would have different antigenic properties (consider as variant) and members within a cluster would have similar antigenic properties (consider as

equal) (Fig 6).

According to the article, there are 273 isolates .But according to the final grouped table on the supporting material, there are only 253 sequences. According to the query condition in reference and supporting material we could get 255 sequences, but there are 3 sequences of A/SP/1/96(AY661200 AY661199 AY661198) and 1 outlier Dk/33/80. The three A/SP/1/96 sequences in which two are identical, and we adapt the first one AY661200. There is one sequence in the grouped table but not in the supporting material, which is A/Sydney/5/97. So the sequence number : 255-2(two identical)-1(Dk/33/80)+1(A/Sydney/5/97)=253 sequences.

The test set have 181 non-identical sequences. After the antigenic type assign, there are 2118 equal cases and 3810 variant cases.



2.5 The ISD set

This set was downloaded from Influenza Sequence Database at 2005/07/10. The query is " A type, HA, Human, H3" , so we could get 1744 sequences . The sequences download from ISD are in the nucleotide format and the length are not identical, so we need to translate them into protein sequence and modify their length to 329 residues. The flowchart is in recorded (Fig. 7). For some virus strains the isolation date is recorded, so those sequences could be clustered according to the influenza season.

2.6 Feature extraction from HA sequence and 3D protein structure.

The inputs of this question are two influenza virus strain's HA protein sequence, then a pairwise comparison is generated. The most common method to compare two influenza virus strains is hamming distance (HD) which counts the total number of changed amino acids [12]. But the HD method can't explain each position's different importance to determine antigenic property. We here apply the position-specific change (PSC) coding, the change of each position is independently recorded as a feature (Fig 8). For example, the number of changed amino acids between A/Panama/2007/99 and A/Fujian/411/2002 is 13 positions, so the HD is 13. But the position change method individually record which 13 positions are changed (Fig 8A).

Since the protein structure of HA is determined and deposit in the Protein Data Bank [21], we further want to utilities the information of structure environment to find important regions on HA structure. Here we apply the contact map coding which could consider each position's environment information. In the contact map coding, each position is considered as the center of a sphere (Fig 8B). The region here is defined as a sphere which center at each amino acid position. Since there are 329 positions in HA, there are 329 regions on the 3D structure of HA. If any position in a region is changes, then this region is considered as changed. The radius of the sphere region is test on the training set from 3 to 12 Å to determine what distance's performance is best

2.7 Antigenic Distance

We want to find out what positions change would affect HI titer value, so we need to define to what degree the HI titer value is considered as changed. In this work, we divide the degree of HI value difference into two categories: antigenic variant and antigenic equal cases.

The HI value from experiment was not convenient for analysis, so the HI values are usually transformed to antigenic distance for large scale analysis. We apply the equation used in the related work [12, 22, 23] to define antigenic variants. This equation calculate the antigenic distance between two virus strains and the equation is show as follows:

$$\sqrt{\frac{(\text{homologous I\_I})(\text{homologous J\_J})}{(\text{heterologous J\_I})(\text{heterologous I\_J})}} \qquad (1)$$

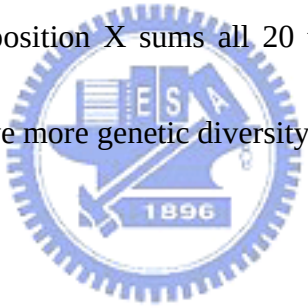
This equation need four cell of HI values that means both two antisera are needed for cross test. A antigenic variant is defined when antigenic distance is $\geqq 4$. That means both two homologous and heterologous HI test should have HI difference equal or more than 4 times. The example is illustrated in (figure 9).

2.8 Entropy

Entropy is used to measure the degree of disorder of one space. We use the entropy here to evaluate the disorder of each position as an index in the genetic level. The equation to calculate entropy is as follows:

$$H(X) = -\sum_{r=1}^{20} P_r \log(P_r) \qquad (2)$$

The H(X) is the entropy of position X and P_r is the probability the amino acid type r in this position. The entropy of position X sums all 20 types of amino acids. The higher the entropy means that position have more genetic diversity.



2.9 Information Gain

Information gain is an index value from information theory with statically meaning. Information gain measures the association between two variables. The higher the information gain means more association between two variables. In this case, a position with very high information gain means if that position is changed then an antigenic variant is expected. As a consequence we could use information gain to build relations between genetic and antigenic evolutions.

Here we use the information gain to measure the degree of each position change's effect to antigenic change. The information gain of a given attribute X with respect to the class attribute Y is the reduction in uncertainty about the value of Y when we know the value of X. The equation is show as follows:

$$I(Y, X) = H(Y) - H(Y \mid X) \qquad (3)$$

The uncertainty about the value of Y is measured by its entropy, H(Y). The uncertainty about the value of Y when we know the value of X is given by the conditional entropy of Y given X, H(Y |X). Equation (3) could translated into following form:

$$I(Y, X) = H(Y) - \sum_{V \in Value(X)} \frac{|Y_V|}{|Y|} H(Y_V) \qquad (4)$$

Equation (4) works when Y and X are discrete variables that take values in {y1...yk} and {x1...xl}.

2.10 Selecting Important Positions by Information Gain

The key idea for selecting important positions is as follows:

Suppose there are many possible HA mutation patterns for influenza virus to escape immune-selection. So we could classify those different HA mutations into several groups. Each group of mutations could explain part of antigenic change from 1968 to 2002.

The process is illustrated in figure (Figure 18). We adapt the greedy method to select important positions. In the level 1 we have full training dataset (181 cases) and then we select the position P_1 with highest antigenic association (highest information gain). Those cases in the level 1 which have mutation on P1 is considered as explained by position P1 and those explained cases are removed from the original dataset. Then in the level 2, the non-explained cases all have no mutation on P_1, so we find the position P_2 with highest information gain for the remain cases in level 2.

By recursively selecting positions with highest information gain and then remove explained cases, we could finally find several positions to explain all cases.

Decision tree are sophisticated data mining tools for discovering patterns and using them to make predictions. The kernel methodology of decision tree is information gain. Here we adapt the decision tree C4.5 [24] help us to select positions with highest information gain in each levels.

Chapter 3

Results and Discussions

The Results could be divided into two main parts. First part is the evaluation process for the suitability of information gain to represent genetic and antigenic evolution. This part also shows the process and result of selecting important positions via information gain.

Second part use the important positions selected by information gain to predict two unseen test set. The predicting performance and results are discussed.

3.1 The result and meaning of information gain value

In order to find out what position change would affect HI value. We calculate the information gain of 329 HA positions from 181 cases. We first to evaluate that whether information gain is a proper index for represent the association between genetic and antigenic evolutions.

The process is illustrated in (figure10). In the figure 10 (table A) we list eight cases of virus comparison, the left most column record the antigenic type between that two virus and the right most column record the genetic changed positions between that two virus's HA protein sequence. In the (table B), we statistic each 329 position's change frequency in the total 181 cases. The change frequency of one position is separated classified into two

categories, the change happened in variant type and in equal type. In the (table C), we do the calculation of entropy and information gain for each 329 positions. The top three information gain positions are 145, 189, 278 all have high association with antigenic type. For example, position 145 have total 62 frequencies of change is total 181 cases and 61 of them happened in the variant type. We may conclude that position with high information gain means high association between position change and antigenic type change. The three positions with high entropy are 226, 135, 124 show low association between genetic and antigenic relationship. For example, position 226 have total 61 frequencies of change in total 181 cases and 34 of them happened in variant type and 27 of them happened in equal type. The result shows that position 226 with very low association between genetic and antigenic relations. Specially note the position 145 with top information gain have been verified by experiment that could lead to cluster transition [11].

. The information gain of each position is plot in graph (Fig. 11).The entropy of each position is plot in graph (Fig. 12). We also plot the information gain on the HA structure (Figure 13). Figure 13 shows the information gain for 329 positions on the HA protein structure in the form of color. The red the color means more high the information gain and the top five information gain positions are labeled. Figure (A) is the front view of HA monomer. Figure (B) is the top view of HA trimer. Compare the red region between front view and top view shows that the top view show more high information gain positions

The comparison between information gain and entropy is also plot in graph (Fig 14). In the genetic view, residues with high entropy may be important. But from the view of information gain, positions with high entropy may have zero information (Ex: position 124). From this figure we may conclude that positions with highest information gain means high association between genetic and antigenic evolutions. The top 10 information gain and top 10 codon diversity positions are list in table (Table 5). The information gain of all positions are listed in appendix (Appendix II).

3.2 The Discuss for Information Gain

Since information gain associates genetic and antigenic evolution. We here to discuss the relationship between them.

The relation between information gain and genetic evolution is plot in figure (Figure 15). For each 181 cases, we compare the genetic changes and information gain for both All positions (329 positions) and epitope sites (131 positions). The linear regression R factor shows good relation between genetic change and information gain ($R > 0.9$) and epitope sites could better fit the genetic change. But for the same value of information gain, the genetic sequence may have high diversity change. For example the information gain value near 0.5, the position change number could range from 7 to 19. The result shows that information gain treat each position change with different weight, but not equal weight.

The relation between information gain and antigenic distance is plot in figure (Figure 16). For each 181 cases, we compare the antigenic distance and information gain for both All positions (329 positions) and epitope sites (131 positions). The result shows that sum of information gain could fit the linear relation to antigenic distance (R >0.74). The result also shows that epitope could better fit the antigenic distance than all positions.

Antigenic variants are defined when antigenic distance $\geqq$ 4, from this figure when sum of information gain > 0.1835 , we could get best predicting performance for predicting antigenic variant. The agreement is 87%.

3.3 The Ability for Information Gain to Predict Antigenic Variants

From figure 15 we found that information gain have the potential to predict antigenic variants. For each pair of viruses we calculate the sum of information gain of changed positions. And the result is compared with a related work [12] which based on hamming distance (The sum of different amino acid positions). The result is illustrated in figure (Figure 17). For each 181 cases, the information gain and number of sequence mutations of epitope sites is plot on the figure. When the sum of Information gain value > 0.1835, the case is predicted as antigenic variant and the agreement is 87 % (158/181). When the sum of sequence mutations $\geqq$ 7, the case is predicted as antigenic variant and the agreement is 83% (150/181). The different predicted cases are illustrated in figure 18.

Figure 18 further observe the different predicted cases by these two methods. Cases successfully predicted by information gain but false predicted by hamming distance are label with big circle. The cases in the circle A means little sequence mutations but which leads to antigenic variant pair. The cases in circle B means large sequence mutations but still antigenic equal pair. The result show that when sequence mutations are less than 11 positions, the position which actually changes would more important than the amount of total mutations. The detail information of 14 successful predicted cases are list in table 6.

Table 6 list the 14 cases successful predicted cases by information gain but false predicted by hamming distance. Eight cases of nine variant cases have changes on the top two information gain positions (145:0.1969 and 189:0.1286). And only the A/England/42/72 vs A/Port_Chalmers/1/73 pair do not have any position with information gain>0.1 but could reach the information gain threshold to antigenic variant at 0.1836. The five equal cases all have change on positions with low information gain.

3.4 Selecting Important Positions by Information Gain (IG Sets)

We divide the 181 cases into several ordered subgroups according the greedy selection method described in the method section. The flowchart is illustrated in figure (Figure 19).

We have 181 cases in the initial and then to find the position with highest information gain. Position 145 have the highest information gain in level one, so the first selected position

is position 145. There are 62 cases in level one have position changes on 145, so those 62 cases are considered as explained by position 145 and removed from the original set.

In each level we further consider those positions with information gain > average + 2*standard deviation. In level one there are other 4 positions satisfy the condition. The position 126 and 278 are considered as co-evolution to position 145 because when the 62 cases are removed from the dataset, their information drop significantly in the level two. The position 189 and 158 are considered as independent important positions would not drop information gain when the 62 cases are removed from the dataset. When the 62 cases are removed from the dataset, we use the same method to select the position with highest information gain in level two.

We further illustrate the detail method to select important positions and define co-evolution sites in figure (Figure 20). This figure plots the information gain in 6 levels of the 21 positions selected by information gain. In each level the selected position having the highest information gain. Positions with close information gain behavior are consider co-evolution groups are colored in the same color. For example the first group includes position 145, 278 and 126 are in green color. Specially note that when the position with highest information gain is selected and those cases have mutation on that position is removed from the dataset. The information of that selected position would drops to zero. As a consequence, some positions (Ex: 155) do not have high information gain in the level one but

it’s information gain gradually increase from level one to level four.

Decision tree tool could help us to select positions with highest information gain for each level in a easy way. The result is illustrated in figure (Figure 21). Figure (A) is the decision tree model of using decision tree tool C4.5 to select the positions with highest information gain. The nodes of decision tree are the positions with highest information gain in each level. The root is on the top of the tree and we should read the tree begin from root. The condition “145 > 0 : Variant (62/1)” means that when position 145 changes and the predicting type is variant. There are 62 cases have change on position 145 and 61 of them are variant type and only one cases disobey the rule. When position 145 have no change (145<=0), then go to check the next position 189. Figure (B) show the selected positions with highest information gain in each levels and the co-evolution positions with that selected position. The last three columns denote the biology meaning and related work’s remark of that selected position.

The positions selected by information gain are label on the HA protein structure (Figure 22). The co-evolution positions in each level are also label on HA structure (Figure 23). From the structure view, the co-evolutions all locate on at least two epitope sites. This result match Wilson’s conclusion for drift variant of epidemiologic importance.

All positions in 6 levels with top information gain are selected to build a model for predicting antigenic variants, the name of this model is “” IG Sets”.

3.5 The result of Contact Map

In the contact map coding, the HA protein structure is divided into many regions which form a sphere center at each amino acid position. The radius of the sphere is tested from 3Å to 12Å and to see which radius's performance of predicting antigenic variants is best. The test is performed by decision tree and the results are illustrated in figure (Figure 24). According the result we choose radius at 9Å

Figure 25 show the positions selected in contact map coding. .Figure (A) Each node in the tree represents a region. Figure (B) The center of each region is list in the table. Position label with "T" means have position change on this set. (C) There are 102 residues covered by all selected regions and 54 of them are in the five epitope sites.

Specially note the first rule : when region 160 and region 134 both occur changes then 90 cases of total 91 cases lead to antigenic change, in which this rule match Wilson's conclusion for drift variant of epidemiologic importance.[5]. This rule imply am important co-evolution relation between this two regions.

We further analysis the 91 cases apply the rule found in contact map. The result is at table 7. This table analysis the 91 cases apply the relation found by contact map. The positions are divided into two groups according to structure neighbors. The value represent the frequency that both two position changes. For example: the marked value 31 meant there are 31 times that position 133 and position 145 change in the same time. The average and standard

frequency is 7.6 and 8.5 respectively. Values higher than average+1STD are underlined. The result shows that position 133 and 145 in region 1 highly co-evolution to position 156 158 160 197 in the region 2. The positions are illustrated in figure 26. Figure A is the HA protein with neutralizing antibody binding (PDB 1KEN). The two regions selected by contact map are labeled in red and blue color. According to the figure, the two regions could directed block by antibody. Figure B is the top view of the HA trimer.

3.6 Compare Training Model Performance with Related works

Since we found important positions by information gain, we further evaluation the selected positions by predicting antigenic variants compare to related works. From related works we could find two methods to predict antigenic variant. Wilson & Cox (1990) [5] proposed that a drift variant of epidemiologic importance usually contains >4 amino acid changes located on >2 of the five antigenic sites. Lee & Chen (2004) [12] proposed that a variant strain is predicted when there are ≧ 7 amino acids change on the five antigenic sites (agreement= 83%). We compare the performance comparison between three of our method and two related works. The results are show in table 8. The important positions selected by IG and contact map both have rate > 90%. The false predicted cases are listed in table 9.

3.7 Application on test set

There are two test sets as our application sets. The summary performance comparison between all methods of there two sets are listed in Table 10. The analysis of information gain and sequence mutations of these tow sets are illustrated in figure 27, 28.

The first test set was extracted from WER from 1968 to 2006 and all the pair comparison contains at least one vaccine strain. From 1968 to 2006, there are 6 influenza season's vaccines not including in the training set and they are now in the test set. The result is list in table 10. The result shows that both our three method have the best accuracy. And the contact map performed best in training set do not have best accuracy in the test set. The detail information about the 50 cases is listed in table 11. In table 11 the value for column 4,5,6 means if the prediction is successful. Value "1" means a successful prediction. The 13 false predicted cases by IG Sets are listed in top 13 rows. The first 7 antigenic equal cases are all have position change on 145, so they are applied to level one. The first three cases A/Hong_Kong/1/68 vs A/England/42/72 and A/Shanghai/31/80 vs A/Bangkok/1/79 and A/Texas/1/77 vs A/Belgium/2/81 all have epitope position changes more than 11 positions while the maximum position change for antigenic equal case in the training set is only 10.

The second test set including two type of cases, in which variant types represent cluster transition and equal types represent member of the same cluster should have related antigenic property. We here take the IG sets model to predict this set for further analysis. The accuracy

of these two types is 87.7%. The individual accuracy of variant type and equal type is 99.24% and 67.04% (Table 12). The detail of each rule's apply situation is illustrated in figure 29. The figure (A) is the important positions selected by information gain and each position is label with level number. (B) is the test result of test set 2 which having 5928 cases. Take level 1 for example, there are total 3098 cases having position changed on position 145 and 2986 of them are antigenic variant type which means a successfully prediction, in other wise if the antigenic type is equal then this is a false prediction. The majority false prediction were due to position 214 and the majority false predicting cases in test set 1 (WER 50 cases) also caused by position 214.

The prediction accuracy of variant type is 99.24% means that this model could detect cluster transitions and the comparisons between our model and related work are listed in table 13 and table 14 [11]. The cluster difference substitutions defined by Smith *et al* are comparison results of local two clusters. Each cluster transition including many position substitutions, so there are many candidate positions needed to be verify by single point mutation experiment. Our method based on a global view which may provide more confidence to determine the antigenic important positions.

The prediction accuracy of equal type is 67.04%. The majority of false predicted cases are in the "BE92" and "WU95"clusters which count 533/2118 = 25.16 % of all equal cases. In the "BE92" cluster there are 5 virus pair disobey the true class in the training set (Table 15).

There are total 9 cases in the test set disobey the true class in the training set which the "BE92" count for 5 cases. According to the WHO vaccine recommendation (Table 1), there are two vaccine candidates in "WU95" cluster which may imply that the "WU95" cluster should further divided into two clusters.

Chapter 4

Conclusions

4.1 Summary

In summary, we apply information gain to build relations between genetic evolution and antigenic drift. Information gain could also apply to be a good index value to predict antigenic variants and having good accuracy than traditional hamming distance method.

We select representative influenza strains from the WHO influenza vaccine strains. Then we extract the genetic sequence and protein structure information as features while the antigenic data from HI titer value are treated as the label(markers). In this problem biologists want to know what positions on HA are critical and what are the antigenic rules, as a consequence we apply the information gain to represent the degree of association between genetic evolution and antigenic drift.

The model based on genetic sequence (IG sets) and structure information (contact map) have 91% and 95% predicting accuracy on training model while *Lee et al* proposed a model based on hamming distance have 83% accuracy. The application on test set also have 87% accuracy and successfully predict the cluster transitions from 1968 to 2002..

4.2 Major contributions and Future Perspectives

By integrating genetic and antigenic evolution, we have successfully applying the information gain to represent the association between genetic and antigenic evolution. We also found that information gain is a good index to predict antigenic variants.

Compare to the traditional hamming distance which consider each positions are equal weighted, the information gain give each position a different weight which could help us to identify the important positions for antigenic evolution.

Recently the threat of avian influenza has to emerge and human know little about the it. Our method could also apply to avian influenza virus and have high potential to identify the important positions.

In the interaction between antigen and antibody, there are many interesting issues. What positions are immunodominant and what kind of residue changes are important? Will the protein structure help to explain some problems? Currently we have just try to find out what positions change are important but there remain many interesting problems for us. Next step we want to know what kinds of amino acid type changes are important and we need to put more effort to discuss the co-evolution between positions.

The methodology currently applied to influenza virus could also apply to other kind of molecule interaction which having experiment data to represent the degree of binding affinity.

Table 1. The influenza vaccine component recommended by WHO from 1968 to 2006

Year	Vaccine Strain [a]	Influenza Season
1968~1973	A/Hong_Kong/1/68	1968/10/01~1973/09/30
1971~1974	**A/England/42/72**	1972/10/01~1974/09/30
1975~1976	**A/Port_Chalmers/1/73**	1974/10/01~1976/09/30
1975	A/Scotland/840/74	1975/10/01~1976/09/30
1977~1978	**A/Victoria/3/75**	1976/10/01~1978/09/30
1979~1980	A/Texas/1/77	1978/10/01~1980/09/30
1981~1983	**A/Bangkok/1/79**	1980/10/01~1983/09/30
1984~1986	**A/Phillipines/2/82**	1983/10/01~1986/09/30
1987	**A/Mississippi/1/85**	1986/10/01~1987/09/30
	A/Christchurch/4/85	
1988	**A/Leningrad/360/86**	1987/10/01~1988/09/30
1989	**A/Sichuan/2/87**	1988/10/01~1989/09/30
1990	**A/Shanghai/11/87**	1989/10/01~1990/09/30
1991	A/Guizhou/54/89	1990/10/01~1991/09/30
1992~1993	**A/Beijing/353/89**	1991/10/01~1993/09/30
1994	**A/Beijing/32/92**	1993/10/01~1994/09/30
1995	**A/Shangdong/9/93**	1994/10/01~1995/09/30
1996	**A/Johannesburg/33/94**	1995/10/01~1996/09/30
1997~1998	**A/Wuhan/359/95**	1996/10/01~1998/09/30
1999~2000	**A/Sydney/5/97**	1999/05/01~2000/04/30
2000~2004	**A/Moscow/10/99**	2000/05/01~2004/04/30
2004~2005	**A/Fujian/411/2002**	2004/05/01~2005/04/30
2005_2	A/Wellington/1/2004	2005/05/01~2005/10/31
2006	A/Calfornia/7/2004	2005/11/01~2006/10/30
2007_1	A/Wisconsin/67/2005	2006/11/01~2007/04/30

a. Influenza virus strains in bold and blue font are present in training set

Table 2. The six HI titer table adapted in training set

Table	From To	Seq	Pairs	Variant [a]	Equal [b]	Period Years	Vaccine	Strain [c]
1	1971 1979	11	55	54	1	9	4	HK71 **ENG72 PC73** MC75 **VIC75** TOK75 ENG75 AC76 VIC76 **BAN179** BAN279
2	1979 1987	8	28	19	9	9	6	**BAN179 PHI82 MIS85 LEN86 SHA87 SIC87** SYD87 VIC87
3	1989 1994	10	45	34	11	6	3	**BEI89** HK90 **BEI92** HK92 GUA93 MAD93 SCO142 SCO160 **SHA93** HK94
4	1994 1996	8	28	7	21	3	2	**JOH94** ALA95 **NCH95** WHN95 AUC96 FUJ96 NY96 SA96
5	1995 1999	5	10	4	6	5	4	**NCH95 SYD97** IRE99 **MOS99 PAN99**
6	1999 2002	6	15	7	8	4	2	**PAN99** FUJ00 CHI01 NY01 **FUJ02** HK02

a. The antigenic type is variant when antigenic distance $\geqq 4$.

b. The antigenic type is equal when antigenic distance < 4.

c. The influenza virus strain names are in abbreviate from and names in blue and bold font are vaccine strains.

Table 3. The list of influenza virus strains in training set

Full name	Vaccine	ID	Abbreviation	Accession no.	Note
A/Hong Kong/107/71		A	HK71	ISDNHK71	
A/England/42/72	Yes	B	ENG72	ISDNENG72	
A/Port Chalmers/1/73	Yes	C	PC73	ISDNPC73	
A/Mayo Clinic/1/75		D	MC75	ISDNMC75	
A/Victoria/3/75	Yes	E	VIC75	Direct entry (12)	J02135
A/Tokyo/1/75		F	TOK75	ISDNTOK75	226x->226L
A/England/864/75		G	ENG75	ISDNENG75	
A/Allegheny County /29/76		H	AC76	Direct entry (12)	J02135
A/Victoria/112/76		I	VIC76	Direct entry (12)	J02135
A/Bangkok/1/79	Yes	J	BAN179	ISDNBK179	
A/Bangkok/2/79		K	BAN279	ISDNBK279	
A/Philippines/2/82	Yes	L	PHI82	ISDNPH282	
A/Mississippi/1/85	Yes	M	MIS85	AF008893	
A/Leningrad/360/86	Yes	N	LEN86	AF008903	
A/Shanghai/11/87	Yes	O	SHA87	AF008886	
A/Sichuan/2/87	Yes	P	SIC87	AF008884	
A/Sydney/1/87		Q	SYD87	AF008882	
A/Victoria/7/87		R	VIC87	AF008888	
A/Beijing/353/89	Yes	S	BEI89	Z46391	
A/Hong Kong/34/90		T	HK90	Z46409	
A/Beijing/32/92	Yes	U	BEI92	Direct entry (13)	
A/Hong Kong/23/92		V	HK92	Direct entry (13)	
A/Guangdong/25/93		W	GUA93	Z46406	
A/Madrid/252/93		X	MAD93	Z46411	
A/Scotland/142/93		Y	SCO142	Z46413	
A/Scotland/160/93		Z	SCO160	Z46414	
A/Shangdong/9/93	Yes	AA	SHA93	Z46417	
A/Hong Kong/1/94		AB	HK94	Z46407	
A/Johannesburg/33/94	Yes	AC	JOH94	AF008774	
A/Alaska/10/95		AD	ALA95	AF008748	
A/Nanchang/933/95	Yes	AE	NCH95	AF008725	
A/Wuhan/359/95	Yes	AF	WHN95	AF008722	
A/Auckland/5/96		AG	AUC96	AF008714	
A/Fujian/47/96		AH	FUJ96	AF008726	
A/New York/37/96		AI	NY96	AF180650	
A/South Africa/1147/96		AJ	SA96	Direct entry (14)	
A/Sydney/5/97	Yes	AK	SYD97	ISDNASYD97	
A/Ireland/10586/99		AL	IRE99	Direct entry (15)	
A/Moscow/10/99	Yes	AM	MOS99	ISDN13277	
A/Panama/2007/99	Yes	AN	PAN99	ISDNCDA001	
A/Fujian/140/2000		AO	FUJ00	Direct entry (16)	
A/Chile/6416/2001		AP	CHI01	Direct entry (16)	
A/New York/55/2001		AQ	NY01	Direct entry (16)	
A/Fujian/411/2002	Yes	AR	FUJ02	ISDN38157	
A/Hong Kong/ 1550/2002		AT	HK02	Direct entry (16)	

Table 4. The sequence number and name of the 11 clusters in the test set [11].

Group	Name	Sequence number	Non-identical
1	HK68	14	12
2	EN72	15	10
3	VI75	9	8
4	TX77	3	3
5	BA79	16	14
6	SI87	25	18
7	BE89	64	29
8	BE92	57	43
9	WU95	28	25
10	SY97	16	16
11	FU02	6	3
total		253	181

Table 5. The positioins with top 10 information gain and positionss with top 10 codon diversity. Residues prefer to change in special antigenic type would have higher information gain. Residues with high rank codon diversity possible have zero information gain.

Residue	Epitope	Variant [a]	Equal [b]	InfoGain [c]	Plotkin,2003 [d]	Receptor [e]	Positive [f]	Smith,2004 [g]
145	A	61	1	0.1969	11		P	+
189	B	66	6	0.1286	13			+
278	C	37	0	0.1255	10			+
158	B	32	0	0.1063	17		P	+
126	A	31	0	0.1025				
217	D	30	0	0.0988				+
174	D	36	1	0.0925				+
31	-	28	0	0.0915				
164	B	28	0	0.0915				+
156	B	51	5	0.0845	9		P	+
135	A	33	8	0	1	R	P	
226	D	34	27	0.0297	2	R	P	
124	A	30	12	0	3		P	+
262	E	15	6	0	4		P	+
133	A	32	6	0	5		P	+
121	D	18	8	0	6		P	
276	C	24	4	0	7			+
172	D	27	4	0.0259	8			+
156	B	51	5	0.0845	9		P	+
278	C	37	0	0.1255	10			+

a. The times of residue change occur in variant type.
b. The times of residue change occurs in variant type.
c. Information gain.
d. The rank of codon diversity of each residues [7].
e. If the residue is in the receptor binding sites.
f. If the residue is under positive selection [8].
g. If this residue leads to cluster transition [11].

Table 6. There the 14 cases successful predicted cases by information gain but false predicted by hamming distance. Eight cases of nine variant cases have changes on the top two information gain positions (145:0.1969 and 189:0.1286). And only the A/England/42/72 vs A/Port_Chalmers/1/73 do not have any position with information gain>0.1 but could reach the 0.1836 threshold. The five equal cases all have change on positions with low information gain.

Virus1	Virus2	Antigenic Type	HD Epitope	IG Epitope	Aminoacidchanges
A/Leningrad/360/86	A/Shanghai/11/87	Equal	8	0.166	I88V F94Y G124D T138A Y155H E188D K189R S247R
A/Hong_Kong/1/94	A/Guangdong/25/93	Equal	8	0.0578	P47S K92E N96S N124D D216N Y219S L226Q R299K
A/Alaska/10/95	A/South_Africa/1147/96	Equal	10	0.0962	T121N G124S D133N K135T G142R D165N D190V I226V N262S D275L
A/Wuhan/359/95	A/South_Africa/1147/96	Equal	8	0.1178	T121N G124S D133N G142R D190V I194L I226V G275L
A/Fujian/47/96	A/South_Africa/1147/96	Equal	8	0.1029	T121N G124S D133N G142R D190V R193S I226V D275L
A/England/42/72	A/Port_Chalmers/1/73	Variant	6	0.2926	L3F **D63N T160A** N188D **S193N A198T G208R**
A/Allegheny_County/29/76	A/Victoria/112/76	Variant	4	0.2112	S157L **R189K** S209N R240G
A/Hong_Kong/34/90	A/Hong_Kong/23/92	Variant	5	0.2732	S157L V174F **R189S** I214T T276N
A/Hong_Kong/34/90	A/Shangdong/9/93	Variant	6	0.3139	D53G S157L V174F **R189S** I214T T276N
A/Beijing/32/92	A/Scotland/142/93	Variant	6	0.2815	H75N I121T S157L **R189S** R201K Q226L
A/Hong_Kong/23/92	A/Madrid/252/93	Variant	4	0.2569	G135K **N145K** R208K T214I
A/Shangdong/9/93	A/Madrid/252/93	Variant	5	0.2976	G53D G135K **N145K** R208K T214I
A/Scotland/160/93	A/Madrid/252/93	Variant	4	0.2866	**N145K** R208K F219S L226Q
A/Madrid/252/93	A/Guangdong/25/93	Variant	5	0.285	K92E N96S **K145N** K208R R299K

Table 7. This table analysis the 91 cases apply the relation found by contact map. The positions are divide into two groups according to structure neighbors.The value represent the frequency that both two position changes. For example: the marked value 31 meant there are 31 times that position 133 and position 145 change in the same time.The average and standard frequency is 7.6 and 8.5 respectively. Values higher than average+1STD are underlined. The result shows that position 133 and 145 in region 1 highly co-evolution to position 156 158 160 197 in region 2.

			Region 1							Region 2					
Position	131	132	133	135	145	146	155	129	156	157	158	159	160	196	197
131	5							0	5	0	0	0	0	0	0
132		10						10	3	1	4	10	4	0	2
133			32					2	31	9	18	2	19	4	19
135				23				0	9	15	0	0	0	0	6
145					53			7	28	14	22	7	24	0	19
146						18		2	18	2	14	2	18	0	18
155							28	10	14	1	4	22	4	0	2
129	0	10	2	0	7	2	10	10							
156	5	3	31	9	28	18	14		50						
157	0	1	9	15	14	2	1			23					
158	0	4	18	0	22	14	4				30				
159	0	10	2	0	7	2	22					22			
160	0	4	19	0	24	18	4						32		
196	0	0	4	0	0	0	0							4	
197	0	2	19	6	19	18	2								25
sum	5	34	104	30	121	74	57	31	108	42	62	43	69	4	66
average	7.6														
STD	8.5														
AVE+1STD	16														

Table 8. The comparison between our models and related works.

	Wilson & Cox, 1990	Lee & Chen, 2004	Sum of Information Gain	IG Sets	Contact Map
Coverage of Positions	131/329	131/329	131/329	101/329	229/329
Method	Hamming Distance	Hamming Distance	Information Gain	Information Gain	Information Gain
Number of Rules	1	1	1	7	10
Error Cases	48	31	23	16	7
Performance	73.5%	82.9%	87.2%	91.2%	96.1%

Table 9.The false predicted cases by Four methods (Lee,2004 ; Sum of IG; IG Sets; Contact Map) .The column “Antigenic Type” means the true class. In the three columns (Lee , 2004 ; Sum of IG;IG Sets ; Contact Map) only value “1” means a successful prediction and “0” means false prediction. Rows in orange color means our three models better than Lee, 2004 and rows in pink color means IG Sets and contact map better than other two models.

Virus1	Virus2	Antigenic Type	Antigenic distance	Lee 2004	Sum of IG	IG Sets	Contact Map	Epitope Different	Total Different	Position Changes
A/England/42/72	A/Port_Chalmers/1/73	Variant	4	0	1	0	1	6	7	L3F D63N T160A N188D S193N A198T G208R
A/Mayo_Clinic/1/75	A/England/864/75	Variant	45.25	1	1	1	0	8	8	K50R E82K N137Y I145N Q189K K193N M260I Y308N
A/Victoria/3/75	A/Allegheny_County/29/76	Equal	1.89	0	0	0	1	7	7	N53D S145N K189R N193D I213V I230V G240R
A/Allegheny_County/29/76	A/Victoria/112/76	Variant	9.24	0	1	1	1	4	4	S157L R189K S209N R240G
A/Bangkok/1/79	A/Bangkok/2/79	Variant	9.24	0	0	0	0	3	3	D188Y N193K S278I
A/Bangkok/1/79	A/Philippines/2/82	Variant	11.31	1	0	1	1	7	7	A138T D144N N173K V182I I213V N248T K307R
A/Bangkok/1/79	A/Leningrad/360/86	Variant	11.31	1	0	1	1	10	11	N2K V88I A138T D144V S159Y V163A N173K D188E I213V N248T K307R
A/Mississippi/1/85	A/Sydney/1/87	Equal	1.41	0	0	0	0	9	9	F94Y G124D A138S Y155H K156E S159Y K189R N193K Q226L
A/Leningrad/360/86	A/Victoria/7/87	Variant	5.66	1	0	1	0	7	7	I88V F94Y G124D T138A Y155H E188D K189R
A/Leningrad/360/86	A/Shanghai/11/87	Equal	2	0	1	0	1	8	8	I88V F94Y G124D T138A Y155H E188D K189R S247R
A/Leningrad/360/86	A/Sydney/1/87	Equal	2	0	0	0	1	8	8	I88V F94Y G124D T138S Y155H E188D K189R N193K
A/Victoria/7/87	A/Sichuan/2/87	Variant	5.66	0	0	0	0	2	2	E156K S186V
A/Victoria/7/87	A/Shanghai/11/87	Variant	5.66	0	0	0	0	1	1	S247R

Virus1	Virus2	Antigenic Type	Antigenic distance	Lee 2004	Sum of IG	IG Sets	Contact Map	Epitope Different	Total Different	Position Changes
A/Victoria/7/87	A/Sydney/1/87	Variant	4	0	0	0	0	2	2	A138S N193K
A/Hong_Kong/34/90	A/Beijing/32/92	Equal	1.41	1	1	0	1	2	2	V174F I214T
A/Hong_Kong/34/90	A/Hong_Kong/23/92	Variant	5.66	0	1	1	1	5	5	S157L V174F R189S I214T T276N
A/Hong_Kong/34/90	A/Shangdong/9/93	Variant	5.66	0	1	1	1	6	6	D53G S157L V174F R189S I214T T276N
A/Beijing/32/92	A/Hong_Kong/23/92	Equal	2	1	1	0	1	3	3	S157L R189S T276N
A/Beijing/32/92	A/Shangdong/9/93	Equal	2	1	0	0	1	4	4	D53G S157L R189S T276N
A/Beijing/32/92	A/Scotland/142/93	Variant	5.66	0	1	1	1	6	6	H75N I121T S157L R189S R201K Q226L
A/Hong_Kong/23/92	A/Scotland/160/93	Equal	2.83	1	1	0	1	4	4	G135K T214I S219F Q226L
A/Hong_Kong/23/92	A/Hong_Kong/1/94	Equal	2.83	0	1	0	1	7	7	S47P D124N G135K T214I N216D S219Y Q226L
A/Hong_Kong/23/92	A/Madrid/252/93	Variant	22.63	0	1	1	1	4	4	G135K N145K R208K T214I
A/Hong_Kong/23/92	A/Guangdong/25/93	Variant	5.66	0	0	1	1	5	5	K92E N96S G135K T214I R299K
A/Shangdong/9/93	A/Scotland/160/93	Variant	4	0	0	1	1	5	5	G53D G135K T214I S219F Q226L
A/Shangdong/9/93	A/Hong_Kong/1/94	Variant	4	1	0	1	1	8	8	S47P G53D D124N G135K T214I N216D S219Y Q226L
A/Shangdong/9/93	A/Madrid/252/93	Variant	16	0	1	1	1	5	5	G53D G135K N145K R208K T214I
A/Shangdong/9/93	A/Guangdong/25/93	Variant	11.31	0	0	1	1	6	6	G53D K92E N96S G135K T214I R299K
A/Scotland/160/93	A/Scotland/142/93	Variant	4	1	0	1	1	7	7	H75N I121T K135G R201K I214T F219S N276T
A/Scotland/160/93	A/Madrid/252/93	Variant	8	0	1	1	1	4	4	N145K R208K F219S L226Q
A/Hong_Kong/1/94	A/Scotland/142/93	Variant	5.66	1	0	1	1	10	10	P47S H75N I121T N124D K135G R201K I214T D216N Y219S N276T
A/Hong_Kong/1/94	A/Guangdong/25/93	Equal	2	0	1	1	1	8	8	P47S K92E N96S N124D D216N Y219S L226Q R299K
A/Scotland/142/93	A/Guangdong/25/93	Variant	11.31	1	0	1	1	10	10	N75H K92E N96S T121I G135K K201R T214I L226Q T276N R299K
A/Madrid/252/93	A/Guangdong/25/93	Variant	11.31	0	1	1	1	5	5	K92E N96S K145N K208R R299K

A/Alaska/10/95	A/South_Africa/1147/96	Equal	2	0	1	1	1	10	10	T121N G124S D133N K135T G142R D165N D190V I226V N262S D275L

Virus1	Virus2	Antigenic Type	Antigenic distance	Lee 2004	Sum of IG	IG Sets	Contact Map	Epitope Different	Total Different	Position Changes
A/Wuhan/359/95	A/South_Africa/1147/96	Equal	2	0	1	1	1	8	8	T121N G124S D133N G142R D190V I194L I226V G275L
A/New_York/37/96	A/South_Africa/1147/96	Equal	2	0	1	1	1	7	7	T121N G124S D133N G142R D190V I229R D275L
A/Fujian/47/96	A/South_Africa/1147/96	Equal	2.83	0	1	1	1	8	8	T121N G124S D133N G142R D190V R193S I226V D275L
A/South_Africa/1147/96	A/Auckland/5/96	Equal	1	0	1	1	1	7	7	N96D N121T S124G N133D R142G V190D L275G
A/Sydney/5/97	A/Moscow/10/99	Equal	1.41	1	0	1	1	6	8	I3L R57Q Y137S S142R K160R I194L A196T H233Y
A/Sydney/5/97	A/Panama/2007/99	Equal	1.41	0	0	1	1	8	12	I3L P21S R57Q Y137S S142R I144N D172E H183L T192I I194L I226V H233Y
A/Sydney/5/97	A/Ireland/10586/99	Equal	1.41	0	0	1	1	7	12	I3L P21S R57Q Y137S S142R D172E H183L T192I I194L I226V H233Y D271N
A/Fujian/140/2000	A/Chile/6416/2001	Variant	4	1	0	0	1	7	12	G14C A43V A106V N144D S186G V194L P199S P221H I226V N246K C247S S273P
A/Fujian/140/2000	A/Hong_Kong/1550/2002	Equal	2	0	1	1	1	8	14	G14C A43V R50G E83K N96S S186V V194I P199S V202I W222R G225D I226V C247S S273P
A/Fujian/140/2000	A/New_York/55/2001	Variant	5.66	1	0	0	1	7	12	G14C A43V G49S A106V N144D S186G V194I P199S I226V R229G C247S S273P
A/Chile/6416/2001	A/Hong_Kong/1550/2002	Equal	2	0	1	1	1	7	12	R50G E83K N96S V106A D144N G186V L194I V202I H221P W222R G225D K246N
False Cases				31	23	16	7			
Accuracy Rate				0.83	0.9	0.91	0.961			

Table1 10

The predicting accuracy comparison between various methods. The result shows that both our three method have the best accuracy. And the contact map performed best in training set do not have best accuracy both in the test set1 and 2.

		Wilson & Cox, 1990	Lee & Chen, 2004	Sum of IG	IG Sets	Contact Map
Training Set	Right Cases	136	150	158	165	174
	Accuracy	73.5%	82.9%	87.2%	91.2%	96.1%
Test Set 1 (WER)	Right Cases	29	32	38	37	34
	Accuracy	58%	64%	76%	74%	68%
Test Set s (Smith,2004)	Right Cases	4191	4710	5180	5201	4504
	Accuracy	70.7%	79.5%	87.4%	87.7%	76.0%

Table 11

The detail information of the 50 test cases from WER. The value for column 4,5,6 means if the prediction is successful. Value "1" means a successful prediction. The 13 false predicted cases by IG Sets are listed in top 13 rows. The first 7 antigenic equal cases are all have position change on 145, so they are applied to level one. The first three cases A/Hong_Kong/1/68 vs A/England/42/72 and A/Shanghai/31/80 vs A/Bangkok/1/79 and A/Texas/1/77 vs A/Belgium/2/81 all have epitope position changes more than 11 positions while the maximum position change for antigenic change in the training set is only 10.

Virus1	Virus2	Antigenic Type	Antigenic distance	Lee 2004	Sum of IG	IG Sets	Apply Level	Epitope Changes	Total Changes	Position Changes
A/Hong_Kong/1/68	A/England/42/72	Equal	1.63	0	0	0	1	15	18	T122N,D133N,G144D,K145S,R146G,T155Y,R50K,S54N,D275G,R207K,R208G,V242I,K62I,V78G,K80Q,V31N,F139C,P199S
A/Shanghai/31/80	A/Bangkok/1/79	Equal	1.41	0	0	0	1	13	16	K145N,T155Y,K156E,G158E,T160K,D275G,I278S,D172G,R207K,P219S,V242I,V244L,M260I,D2N,N9S,D31N
A/Texas/1/77	A/Belgium/2/81	Equal	2.00	0	0	0	1	11	12	G124S,N133S,P143S,N145K,G146S,K156E,T160R,Q197R,V217I,I67V,I260M,N2K
A/Belgium/2/81	A/Bangkok/1/79	Equal	2.00	1	0	0	1	5	6	S124G,K145N,R160K,V67I,M260I,K2N
A/Belgium/2/81	A/Philippines/2/82	Equal	1.41	0	0	0	1	12	13	S124G,A138T,D144N,K145N,R160K,K307R,N173K,V182I,I213V,N248T,V67I,M260I,K2N
A/Shanghai/24/90	A/Beijing/353/89	Equal	2.83	0	0	0	1	8	8	G135E,N145K,S186I,D190E,S193N,Q226L,E62K,N262T
A/Wyoming/3/2003	A/Wellington/1/2004	Equal	2.00	0	0	0	2	9	9	A128T,Y159F,V186G,S189N,S193N,Y219S,I226V,S227P,N246S
A/Texas/1/77	A/Bangkok/1/79	Variant	5.66	1	1	0	6	7	7	N133S,P143S,G146S,K156E,T160K,Q197R,V217I
A/Texas/1/77	A/Bangkok/2/79	Variant	11.31	1	1	0	6	10	10	N133S,P143S,G146S,K156E,T160K,D188Y,N193K,Q197R,S278I,V217I
A/Caen/1/84	A/Philippines/2/82	Variant	5.66	0	0	0	6	6	7	A138T,V144N,K156E,Y279S,V182I,Y94F,K2N
A/Caen/1/84	A/Christchurch/4/85	Variant	4.00	0	0	0	6	6	6	K156G,S159Y,V163A,Y279S,N246S,Y94F
A/Christchurch/4/85	A/Philippines/2/82	Variant	5.66	1	0	0	6	7	8	A138T,V144N,G156E,Y159S,A163V,V182I,S246N,K2N
A/Washington/15/91	A/Beijing/353/89	Variant	4.00	0	0	0	6	2	2	D158E,S193N
A/Hong_Kong/1/68	A/England/878/69	Variant	5.33	1	1	1	1	10	13	D133N,G144D,K145S,R146G,R50K,S54N,K62I,D63N,K80Q,N81D,V31N,F139C,P199S
A/Hong_Kong/1/68	A/Hong_Kong/107/71	Variant	13.86	1	1	1	1	13	18	Q132E,D133N,G144D,K145S,R146G,G129E,S159R,N188D,A198T,R50K,S54N,K62I,K80Q,T10K,V31D,I34T,F139C,P19

										9S
Virus1	**Virus2**	**Antigenic Type**	**Antigenic distance**	**Lee,2004**	**Sum of IG**	**IG Sets**	**Apply Level**	**Epitope Changes**	**Total Changes**	**Position Changes**
A/England/878/69	A/England/42/72	Variant	6.53	1	1	1	4	9	9	T122N,T155Y,D275G,R207K,R208G,V242I,N63D,V78G,D81N
A/Hong_Kong/1/68	A/Port_Chalmers/1/73	Variant	22.63	1	1	1	1	19	23	T122N,D133N,G144D,K145S,R146G,T155Y,T160A,N188D,S193N,A198T,R50K,S54N,D275G,R207K,V242I,K62I,D63N,V78G,K80Q,L3F,V31N,F139C,P199S
A/Hong_Kong/1/68	A/Victoria/3/75	Variant	30.17	1	1	1	1	26	29	T122N,T126N,D133N,N137S,G144D,K145S,R146G,T155Y,L164Q,N188D,Q189K,S193N,R50K,S54N,D275G,I278S,F174S,R201K,R207K,I217V,V242I,K62I,D63N,V78G,K80Q,T83K,V31N,F139C,P199S
A/Hong_Kong/1/68	A/England/864/75	Variant	156.77	1	1	1	1	26	30	T122N,T126N,D133N,N137Y,G144D,K145N,R146G,T155Y,G158E,N188D,Q189K,S193N,S54N,D275G,I278S,Y308N,R207K,I217V,V242I,K62I,D63N,V78G,K80Q,E82K,T83K,M260I,S9N,V31D,F139C,P199S
A/Hong_Kong/1/68	A/Tokyo/1/75	Variant	64.00	1	1	1	1	20	24	T122N,T126I,D133N,G144D,K145R,R146G,T155Y,K156Q,T160S,L164Q,N188D,S193D,R50K,S54N,F174S,R201K,R207K,I217V,K62I,K80Q,S9N,V31D,F139C,P199S
A/Texas/1/77	A/Shanghai/31/80	Variant	8.00	1	1	1	1	16	19	N133S,P143S,N145K,G146S,Y155T,E158G,Q197R,G275D,S278I,G172D,K207R,V217I,S219P,I242V,L244V,I260M,N2D,S9N,N31D
A/Texas/1/77	A/Philippines/2/82	Variant	5.66	1	1	1	5	14	14	N133S,A138T,P143S,D144N,G146S,K156E,T160K,Q197R,K307R,N173K,V182I,I213V,V217I,N248T
A/Caen/1/84	A/Texas/1/77	Variant	4.00	1	1	1	5	13	14	S133N,S143P,V144D,S146G,K160T,R197Q,Y279S,R307K,K173N,V213I,I217V,T248N,Y94F,K2N
A/Caen/1/84	A/Bangkok/1/79	Variant	11.31	1	1	1	5	8	9	V144D,K156E,Y279S,R307K,K173N,V213I,T248N,Y94F,K2N
A/Caen/1/84	A/Mississippi/1/85	Equal	2.83	1	1	1	6	4	4	V163A,Y279S,L226Q,Y94F
A/Christchurch/4/85	A/Wellington/4/85	Equal	2.00	1	1	1	6	2	2	G156E,S246N
A/Christchurch/4/85	A/Mississippi/1/85	Equal	2.83	1	1	1	6	4	4	G156K,Y159S,L226Q,S246N

Virus1	Virus2	Antigenic Type	Antigenic distance	Lee,2004	Sum of IG	IG Sets	Apply Level	Epitope Changes	Total Changes	Position Changes
A/Wellington/4/85	A/Philippines/2/82	Equal	2.83	1	1	1	6	5	6	A138T,V144N,Y159S,A163V,V182I,K2N
A/Wellington/4/85	A/Mississippi/1/85	Equal	2.00	1	1	1	6	3	3	E156K,Y159S,L226Q
A/England/427/88	A/Shanghai/11/87	Equal	1.89	1	1	1	6	5	6	A131T,R299K,S247R,K82E,E83K,S15L
A/England/427/88	A/Guizhou/54/89	Equal	1.63	1	1	1	6	4	5	V144I,Y159H,S186I,Q44H,S15L
A/England/427/88	A/Beijing/353/89	Variant	5.66	0	1	1	1	3	4	G135E,N145K,S186I,S15L
A/Guizhou/54/89	A/Shanghai/11/87	Equal	2.67	0	1	1	6	9	9	A131T,I144V,H159Y,I186S,H44Q,R299K,S247R,K82E,E83K
A/Guizhou/54/89	A/Guandong/39/89	Equal	1.41	1	1	1	6	3	3	I144V,H44Q,I260M
A/Guizhou/54/89	A/Beijing/353/89	Variant	22.63	0	1	1	1	5	5	G135E,I144V,N145K,H159Y,H44Q
A/Guandong/39/89	A/Shanghai/11/87	Equal	3.46	0	1	1	6	8	8	A131T,H159Y,I186S,R299K,S247R,K82E,E83K,M260I
A/Guandong/39/89	A/Beijing/353/89	Variant	13.06	0	1	1	1	4	4	G135E,N145K,H159Y,M260I
A/Shanghai/11/87	A/Beijing/353/89	Variant	7.54	1	1	1	1	8	8	T131A,G135E,N145K,S186I,K299R,R247S,E82K,K83E
A/Shanghai/16/89	A/Beijing/353/89	Variant	5.66	0	1	1	1	3	3	G135E,N145K,H159Y
A/England/261/91	A/Beijing/353/89	Equal	1.41	1	1	1	6	5	5	T138A,S186I,S193N,I196V,D172G
A/Washington/15/91	A/Beijing/32/92	Variant	8.00	1	1	1	1	10	10	S133D,E135G,K145N,E156K,D158E,I186S,E190D,I214T,L226Q,T262N
A/Shangdong/9/93	A/Johannesburg/33/94	Variant	5.66	1	0	1	7	7	7	D124N,G135K,S47P,G53D,T214I,N216D,S219Y
A/Guangdong/25/93	A/Johannesburg/33/94	Equal	1.00	0	1	1	6	7	7	D124N,S47P,K299R,S96N,N216D,S219Y,E92K
A/Wuhan/359/95	A/Sydney/5/97	Variant	16.00	1	1	1	3	10	12	G124S,D133N,G142S,V144I,K156Q,E158K,V196A,N276K,T121N,K62E,L3I,Y233H
A/Moscow/10/99	A/Fujian/411/2002	Variant	11.31	1	1	1	4	13	17	A131T,I144N,H155T,Q156H,R160K,S186G,T192I,T196A,R50G,D172E,I226V,H75Q,E83K,L25I,V202I,W222R,G225D
A/Moscow/10/99	A/Chile/6416/2001	Equal	2.83	0	1	1	6	8	10	I144D,R160K,S186G,T192I,T196A,D172E,I226V,N246K,A106V,P221H
A/California/7/2004	A/Wyoming/3/2003	Variant	4.00	1	1	1	1	8	8	N145K,T128A,F159Y,G186V,N188D,N189S,S219Y,P227S
A/California/7/2004	A/Wellington/1/2004	Variant	4.00	0	1	1	1	5	5	N145K,N188D,S193N,I226V,N246S
A/Wisconsin/67/2005	A/California/7/2004	Equal	2.83	1	1	1	6	3	6	D122N,D188N,F193S,H195Y,I223V,N225D
A/Wisconsin/67/2005	A/New_York/55/2004	Equal	2.00	1	1	1	6	5	8	D122N,A138S,G186V,F193R,S219Y,H195Y,I223V,N225D

Table 12. The result of apply training model on test set2. The total performance is 87.7%. The left part of table evaluates the cluster transition and the right part of table evaluate that strains in the same cluster should have close antigenic properties. The most false cases are [b][a] type which means we predict the cases in the same cluster to antigenic variants. The majority false cases are in the "BE92" cluster.

Cluster-Transition	**[a][a]**	**[a][b]**	**[b][a]**	**[b][b]**	**Cluster**	
HK68-EN72	120	0	44	22	HK68	1
EN72-VI75	80	0	29	16	EN72	2
VI75-TX77	24	0	13	15	VI75	3
TX77-BA79	40	2	2	1	TX77	4
BA79-SI87	252	0	45	46	BA79	5
SI87-BE89	522	0	17	136	SI87	6
BE89-BE92	1247	0	0	406	BE89	7
BE92-WU95	1048	27	486	417	BE92	8
WU95-SY97	400	0	47	253	WU95	9
SY97-FU02	48	0	15	105	SY97	10
	0	0	0	3	FU02	11
Rate	99.24%	0.76%	32.96%	67.04%		

[a][a]: Variant type and predict Variant
[a][b]: Variant type and predict Equal
[b][a]: Equal type and predict Variant
[b][b]: Equal type and predict Equal

Table 13. Analysis the rules and positions lead to cluster transitions.
The Result shows that position 145 plays an important role in 6 cluster transitions. Position 189 only counts for BA79-SI87 cluster transition. Position 62 count for theWU95-SY97 cluster transition

	Level1	Level2	Level3	Level4	Level5	Level6	Level7	
	145	189	62	155	213	214	214	sum
HK68-EN72	67	0	7	46	0	0	0	120
EN72-VI75	73	7	0	0	0	0	0	80
VI75-TX77	3	0	8	0	13	0	0	24
TX77-BA79	0	0	28	0	12	2	0	42
BA79-SI87	0	252	0	0	0	0	0	252
SI87-BE89	522	0	0	0	0	0	0	522
BE89-BE92	1247	0	0	0	0	0	0	1247
BE92-WU95	1032	2	2	0	0	27	12	1075
WU95-SY97	39	0	361	0	0	0	0	400
SY97-FU02	3	0	0	45	0	0	0	48
Total	2986	261	406	91	25	29	12	3810

Table 14. Cluster-difference amino acid substitutions, and distances between antigenic clusters [11].

Cluster transition	Genetic distance (aa changes)	Antigenic distance (units)	Genetic antigenic ratio	Cluster-difference substitutions					
				Site A	Site B	Site C	Site D	Site E	Other
HK68-EN72	12.1	3.4	3.6	T122N G144D	T155Y* N188D		R207K		
EN72-VI75	14.6	4.4	3.3	N137S*† S145N‡	L164Q Q189K S193D‡	N53D I278S	F174S R102K‡ I213V I217V I230V		
V175-TX77	14.8	3.4	4.4	S137Y*†	G158E‡ Q164L D193N‡	K50R† D53N	S174F K201R‡ V213I V230I	E82K M260I	
TX77-BA79	16.0	3.3	4.8	N133S‡ P143S G146S	K156E‡ T160K Q197R‡	N53D N54S	D172G† V217I V244L	162K K82E	
BA79-SI87	11.9	4.9	2.4	G124D‡	Y155H* K189R				
SI87-BE89	6.9	4.6	1.5	N145K‡					
BE89-BE92	13.7	7.8	1.8	S133D† K145N‡	E156K‡ E190D*‡			T262N‡	
BE92-WU95	9.9	4.6	2.2	N145K‡					
WU95-SY97	16.0	4.7	3.4		K156Q‡ E158K‡ V196A†	N276K†		K62E	
SY97-FU02	16.0	3.5	4.5	A131T	H155T* Q156H‡	R50G†		H75Q E83K	L25I V202I W222R G225D*
Total	131.9	44.6							
Average	13.2	4.5	3.2						
SD	2.9	1.3	1.1						

Table 15. This table list all the 39 cases appear both in training set and test set. The “TrainingType” means the true class and the “TestType” mean our assumption. If trainingtype and testtype are different then the conflict column will denote “X”.

Index	Virus1	Virus2	Cluster1	Cluster2	TrainingType	TestType	Conflict
1	A/HK/107/71	A/EN/42/72	HK68	EN72	Variant	Variant	
2	A/HK/107/71	A/PC/1/73	HK68	EN72	Variant	Variant	
11	A/EN/42/72	A/PC/1/73	EN72	EN72	Variant	Equal	X
13	A/EN/42/72	A/VIC/3/75	EN72	VI75	Variant	Variant	
21	A/PC/1/73	A/VIC/3/75	EN72	VI75	Variant	Variant	
64	A/PH/2/82	A/LE/360/86	BA79	BA79	Equal	Equal	
65	A/PH/2/82	A/VI/7/87	BA79	SI87	Variant	Variant	
66	A/PH/2/82	A/SI/2/87	BA79	SI87	Variant	Variant	
67	A/PH/2/82	A/SH/11/87	BA79	SI87	Variant	Variant	
74	A/LE/360/86	A/VI/7/87	BA79	SI87	Variant	Variant	
75	A/LE/360/86	A/SI/2/87	BA79	SI87	Variant	Variant	
76	A/LE/360/86	A/SH/11/87	BA79	SI87	Equal	Variant	X
78	A/VI/7/87	A/SI/2/87	SI87	SI87	Variant	Equal	X
79	A/VI/7/87	A/SH/11/87	SI87	SI87	Variant	Equal	X
81	A/SI/2/87	A/SH/11/87	SI87	SI87	Equal	Equal	
85	A/BE/353/89	A/BE/32/92	BE89	BE92	Variant	Variant	
89	A/BE/353/89	A/HK/1/94	BE89	BE92	Variant	Variant	
90	A/BE/353/89	A/SC/142/93	BE89	BE92	Variant	Variant	
92	A/BE/353/89	A/GU/25/93	BE89	BE92	Variant	Variant	
104	A/BE/32/92	A/HK/1/94	BE92	BE92	Variant	Equal	X
105	A/BE/32/92	A/SC/142/93	BE92	BE92	Variant	Equal	X
106	A/BE/32/92	A/MA/252/93	BE92	WU95	Variant	Variant	
107	A/BE/32/92	A/GU/25/93	BE92	BE92	Variant	Equal	X

Index	Virus1	Virus2	Cluster1	Cluster2	TrainingType	TestType	Conflict
123	A/HK/1/94	A/SC/142/93	BE92	BE92	Variant	Equal	X
124	A/HK/1/94	A/MA/252/93	BE92	WU95	Variant	Variant	
125	A/HK/1/94	A/GU/25/93	BE92	BE92	Equal	Equal	
126	A/SC/142/93	A/MA/252/93	BE92	WU95	Variant	Variant	
127	A/SC/142/93	A/GU/25/93	BE92	BE92	Variant	Equal	X
128	A/MA/252/93	A/GU/25/93	WU95	BE92	Variant	Variant	
130	A/JO/33/94	A/WU/359/95	BE92	WU95	Variant	Variant	
131	A/JO/33/94	A/NA/933/95	BE92	WU95	Variant	Variant	
142	A/WU/359/95	A/NA/933/95	WU95	WU95	Equal	Equal	
157	A/NA/933/95	A/SY/5/97	WU95	SY97	Variant	Variant	
158	A/NA/933/95	A/MO/10/99	WU95	SY97	Variant	Variant	
159	A/NA/933/95	A/PA/2007/99	WU95	SY97	Variant	Variant	
161	A/SY/5/97	A/MO/10/99	SY97	SY97	Equal	Equal	
162	A/SY/5/97	A/PA/2007/99	SY97	SY97	Equal	Equal	
164	A/MO/10/99	A/PA/2007/99	SY97	SY97	Equal	Equal	
171	A/PA/2007/99	A/FU/411/2002	SY97	FU02	Variant	Variant	

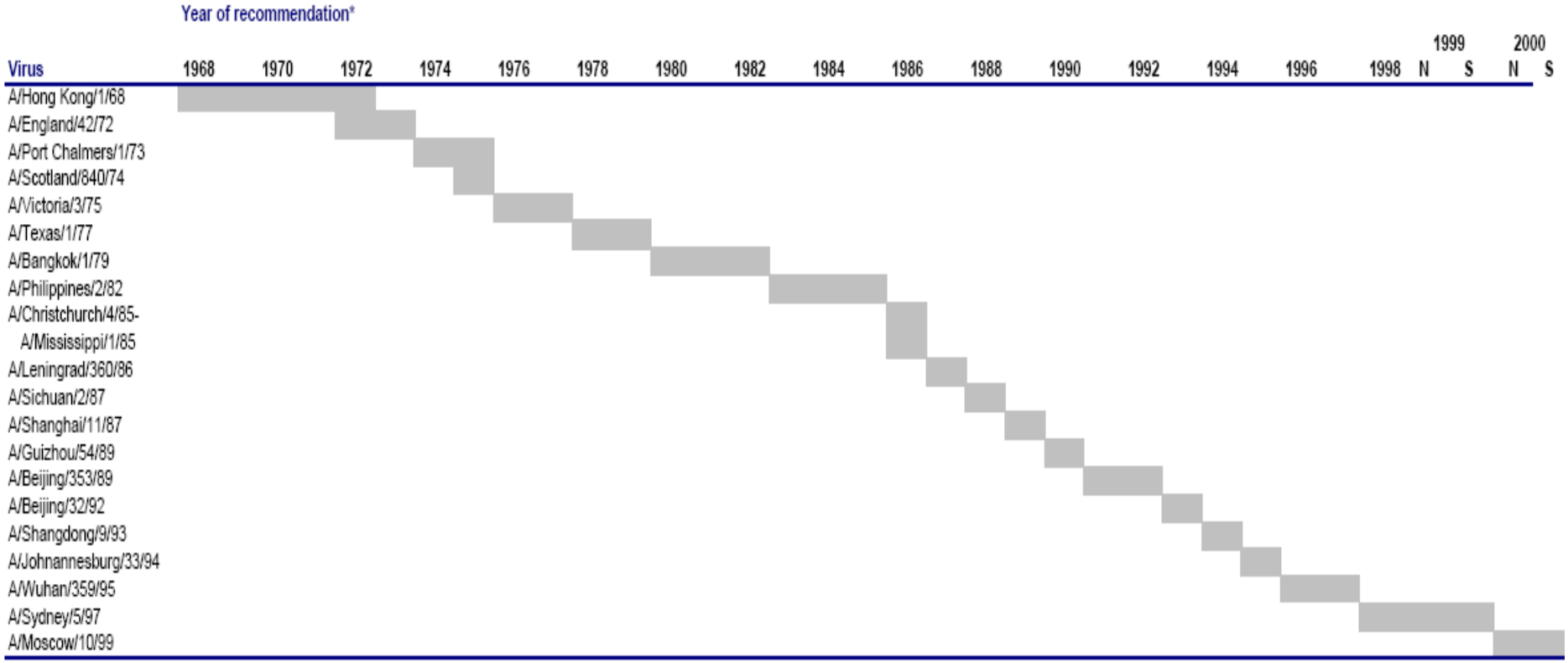


Figure 1. Viruses recommended for inclusion in the influenza H3N2 virus vaccines, 1968-2000 [25]. The influenza HA protein is the major influenza vaccine component.

(A) (B)

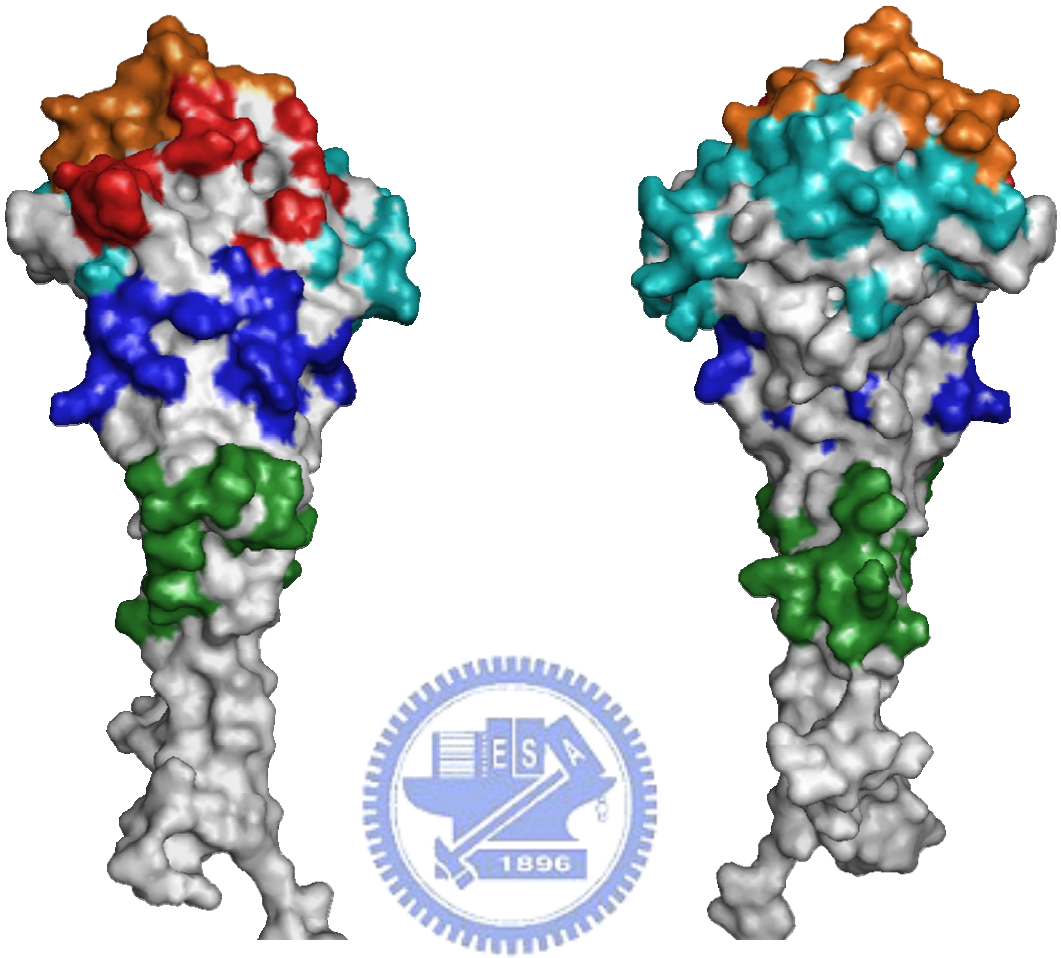


Figure 2. The 3D structure of HA monomer (Protein Data Bank ID 1HGF:A). (A) the front view and (B) the back view. The five epitope sites are space fill in different colors.

HEMAGGLUTINATION INHIBITION TEST OF INFLUENZA H3 VIRUSES (01/31/03)

STRAIN DESIGNATION	REFERENCE FERRET ANTISERA										
REFERENCE ANTIGENS	A SYD/5	B MOS/10	C PN/2007	D FUJ/140	E CI/6416	F HK/1550	G NY/55	H GUA/405	I FUJ/411	Date collected	Passage
1 A/SYDNEY/05/97	320	80	80	80	160	160	320	160	20	06/23/97	CK2E2/E2
2 A/MOSCOW/10/99	320	320	160	40	320	160	320	160	20	01/15/99	E2/E2
3 A/PANAMA/2007/99*	160	160	320	320	320	320	640	320	40	07/12/99	E5
4 A/FUJIAN/140/2000	40	20	160	640	160	320	160	320	20	7/30/00	E2/E4
5 A/CHILE/6416/2001	80	40	80	80	320	80	320	160	20	6/14/01	E7
6 A/HONG KONG/1550/2002	640	320	320	320	640	640	1280	640	160	01/19/02	E5
7 A/NEW YORK/55/2001	320	160	320	160	640	320	1280	320	40	11/22/01	E4
8 A/GUANGZHOU/405/2002	40	40	160	80	160	160	80	320	20	21/2002	C2/C4
9 A/FUJIAN/411/2002	80	80	160	320	320	320	160	640	1280	08/11/02	X/C2
TEST ANTIGENS											
10 A/EGYPT/919441/2002	320	640	640	320	1280	640	1280	640	80	11/07/02	C2/C1
11 A/ALASKA/4/2002	160	160	320	640	320	320	320	640	1280	12/02/02	X/C1
12 A/TEXAS/53/2002	160	160	320	640	640	320	320	1280	2560	12/18/02	M1/C1
13 A/ENGLAND/455/2002	320	160	320	160	640	320	640	320	320	12/30/02	C1/C2
14 A/EGYPT/914454/2002	320	320	320	320	1280	640	1280	640	80	09/04/02	C2/C1
15 A/EGYPT/915167/2002	320	160	320	160	1280	320	1280	640	40	10/02/02	C2/C1
16 A/EGYPR/919475/2002	320	320	320	320	1280	640	1280	640	80	11/11/02	C2/C1
17 A/EGYPT/921138/2002	160	160	320	320	1280	640	1280	640	80	10/22/02	C2/C2
18 A/HAWAII/19/2002	160	160	160	320	320	320	320	640	1280	2/30/02	M2/C1
19 A/HAWAII/20/2002	80	160	160	320	160	320	160	640	640	12/04/02	M2/C2
20 A/BEIJING/253/2002	160	320	160	20	640	160	320	640	2560	12/12/02	C1/C1

Figure 3. The Hemagglutination inhibition test table.

The Hemagglutination inhibition test results are present in a table form. The first column represent antigen and the first row represent antisera. The HI test always need reference tests when the test is applied to new isolates. The reference antigens are always in the top rows of the table. The HI titer values represent serial dilution folds. The antigenic distance is calculated from HI titer values.

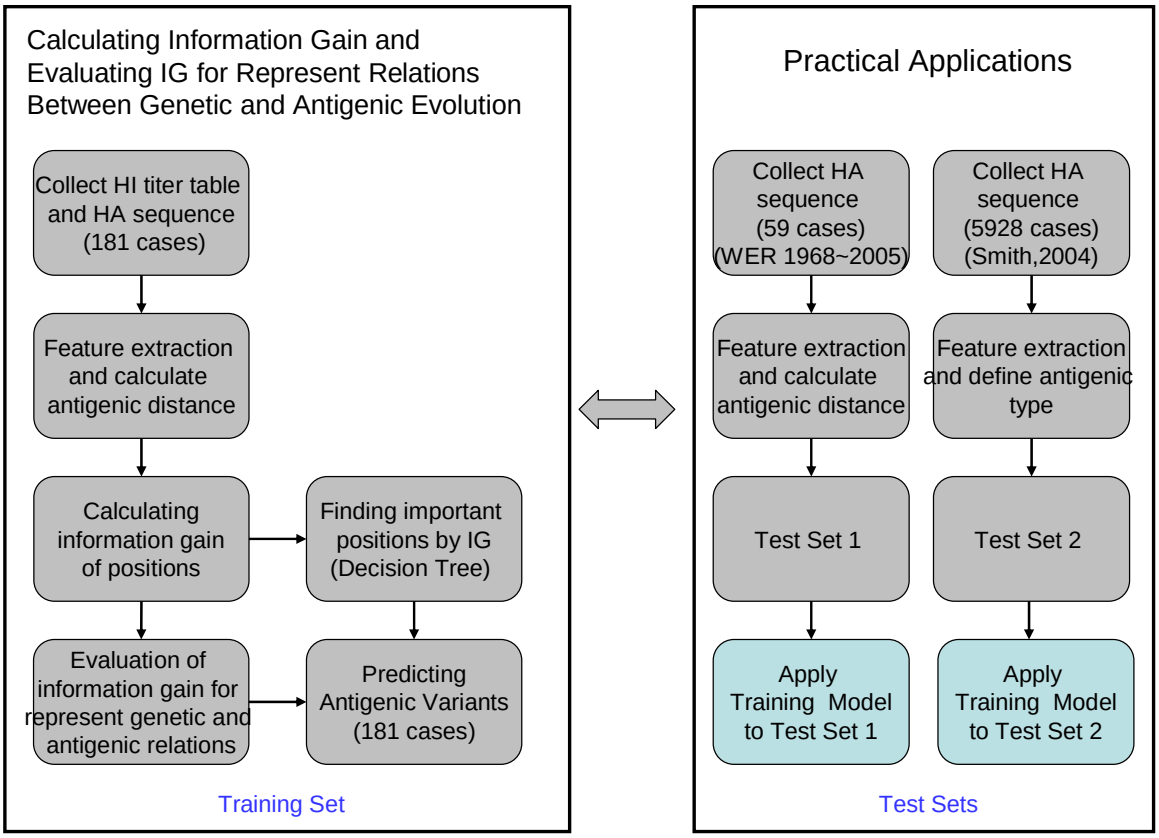


Figure 4. The flowchart of this research.

The research flowchart could be divided into two parts. In the left part we calculate the information gain of 329 HA positions from on a representative training dataset and evaluate the suitability for information gain to represent the relations between genetic and antigenic evolutions. Then in the right part we apply the important positions selected by information gain to predict antigenic variants on two unseen and meaningful application sets (test sets).

In the first part we first select representative influenza strains from the WHO influenza vaccine strains. Then we extract the genetic data from sequence and antigenic data from HI titer value. The performance of predicting antigenic variants is compared to related works. In the second part the model is applied to practical applications and the result is analyzed.

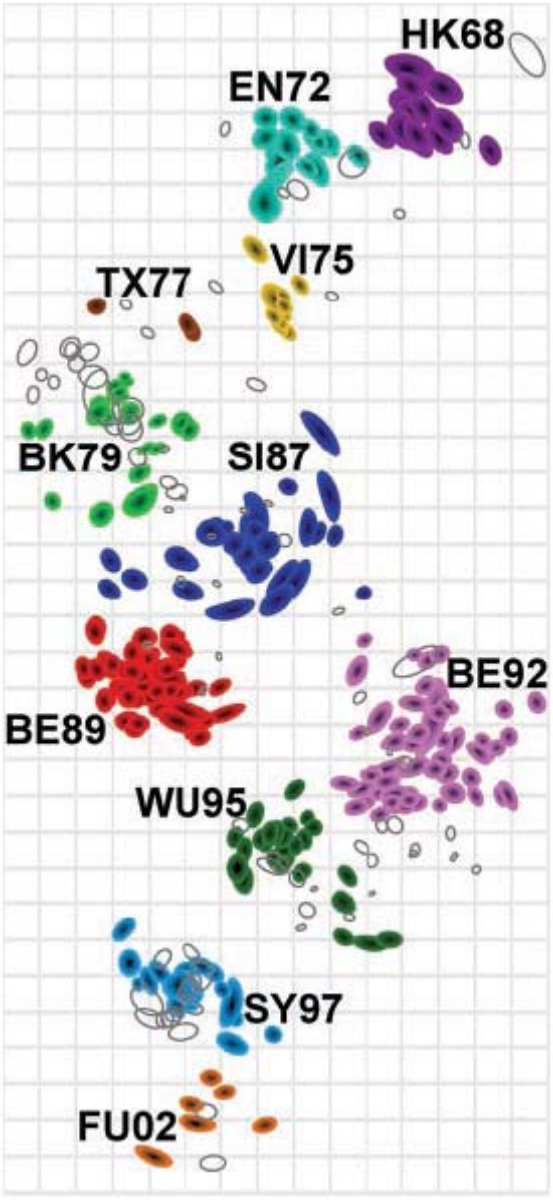
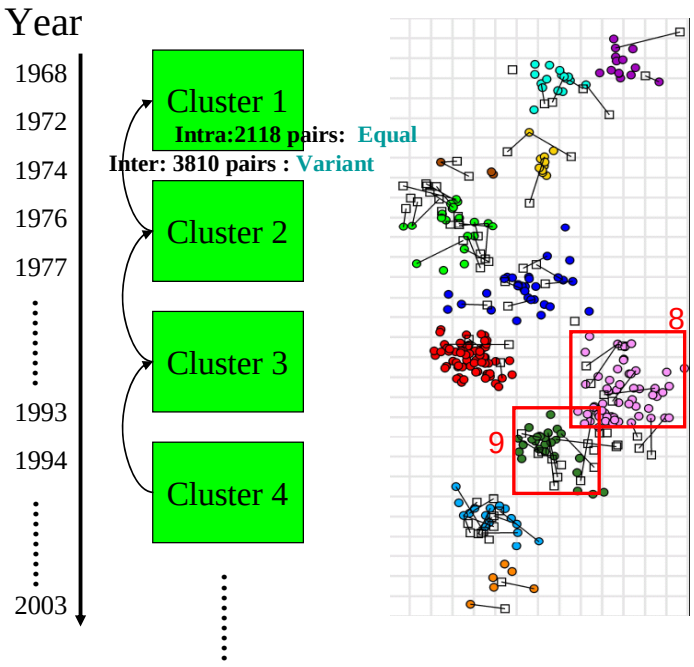


Figure 5. The antigenic map of influenza A(H3N2) virus from 1968 to 2003.

The cluster result is based on K-mean and which using the antigenic distance transformed from HI titer. There are 11 clusters with different colors and different cluster means different antigenic properties. Every cluster is named after the first vaccine strain in the cluster—two letters refer to the location of isolation [11].

(A) **(B)** **(C)**



Group	Name	Seq #
1	HK68	12
2	EN72	10
3	VI75	8
4	TX77	3
5	BA79	14
6	SI87	18
7	BE89	29
8	BE92	43
9	WU95	25
10	SY97	16
11	FU02	3
total		181

Figure 6. This figure show how the antigenic type in application set is defined. (B) (C) There are 11 clusters as a result by K-mean method which using antigenic distance from HI titer values. (A) The equal type is defined within a cluster and the variant type is defined when the cluster transitions are occur [11].

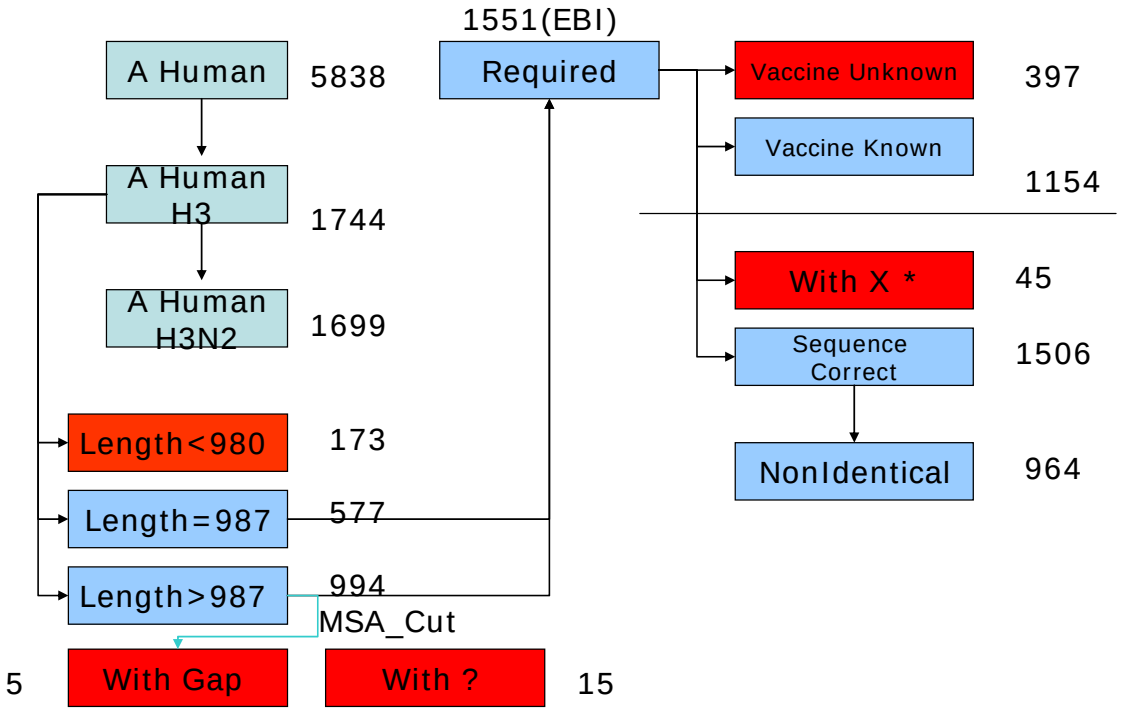


Figure 7. The flowchart of processing query sequences from ISD. The sequences deposit in the ISD are in the nucleotide format and the length of HA not always long enough (> 987 nucleotides). If the sequence have “?” nucleotides or could not be translated into protein sequence then that sequence is removed.

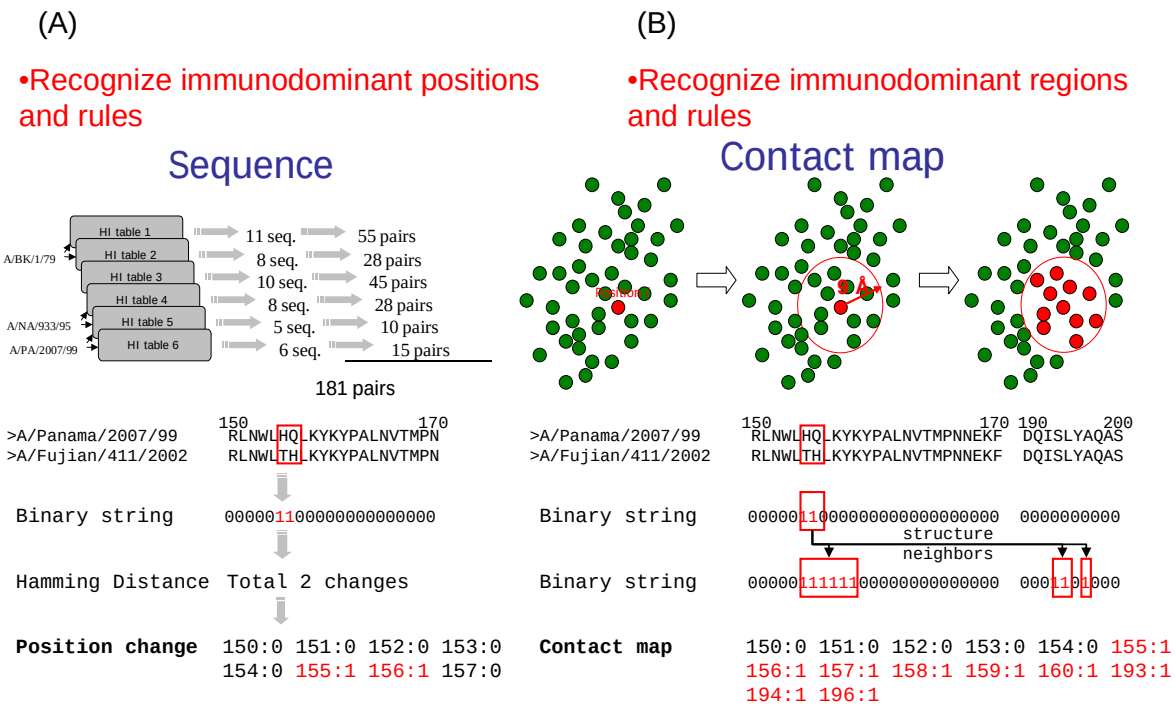


Figure 8. This figure shows how the position-specific changes and contact map coding works. (A) If the position occurs change then the binary string would denote “1”. Each position change is treated as individual feature. (B) In the contact map coding, if any single position in a region occurs change then all member in this region are consider as changed.

STRAIN DESIGNATION / REFERENCE ANTIGENS	REFERENCE FERRET ANTISERA A SYD/5	B MOS/10	C PN/2007	D FUJ/140	E CI/6416	F HK/1550	G NY/55	H GUA/405	I FUJ/411		Date collected	Passage
1 A/SYDNEY/05/97	320	80	80	80	160	160	320	160	20		06/23/97	CK2E2/E2
2 A/MOSCOW/10/99	320	320	160	40	320	160	320	160	20	J_I	01/15/99	E2/E2
3 A/PANAMA/2007/99*	160	160	320	320	320	320	640	320	40		07/12/99	E5
4 A/FUJIAN/140/2000	40	20	160	640	160	320	160	320	20		7/30/00	E2/E4
5 A/CHILE/6416/2001	80	40	80	80	320	80	320	160	20		6/14/01	E7
6 A/HONG KONG/1550/2002	640	320	320	320	640	640	1280	640	160		01/19/02	E5
7 A/NEW YORK/55/2001	320	160	320	160	640	320	1280	320	40		11/22/01	E4
8 A/GUANGZHOU/405/2002	40	40	160	80	160	160	80	320	20	J_J	21/2002	C2/C4
9 A/FUJIAN/411/2002	80	80	160	320	320	320	160	640	1280		08/11/02	X/C2

I_I

I_J

$$\text{Equation 1}: \sqrt{\frac{(\text{homologous I\_I})(\text{homologous J\_J})}{(\text{heterologous J\_I})(\text{heterologous I\_J})}}$$

$$\text{Distance}(\text{antigen 3}, \text{antigen 9}) = \sqrt{\frac{(320)(1280)}{(160)(40)}} = 8$$

Figure 9. This figure demonstrate that how the antigenic distance is calculated. Currently there are two kinds of transformation equations in related works[12, 23], we adapt equation one in this work.

Table A

Antigenic	Virus1	Virus2	Genetic Sequence (Different Positions on HA)
Variant	A/England/42/72	A/Mayo Clinic1/75	S9N N31D D63N T83K T126N S145I G158E N188D S193K G208R I217V I278S
Variant	A/Port Chalmers/1/73	A/Mayo Clinic1/75	F3L S9N N31D T83K T126N S145I G158E A160T N193K T198A I217V I278S
Variant	A/Mayo Clinic1/75	A/England/864/75	K50R E82K N137Y I145N Q189K K193N M260I Y308N
Equal	A/Philippines/2/82	A/Mississippi/1/85	N2K T138A N144V E156K V163A I182V L226Q
Variant	A/Mississippi/1/85	A/Victoria/7/87	F94Y G124D Y155H K156E S159Y K189R Q226L
Variant	A/Beijing/353/89	A/Beijing/32/92	S133D E135G K145N E156K I186S E190D N193S I214T L226Q T262N
Equal	A/Hong Kong/23/92	A/Hong Kong/1/94	S47P D124N G135K T214I N216D S219Y Q226L
Variant	A/Hong Kong/23/92	A/Madrid/252/93	G135K N145K R208K T214I

Table B

Position	Epitope	Change[a]	Variant[b]	Equal[c]
145	A	62/181	61/181	1/181
189	B	72/181	66/181	6/181
278	C	37/181	37/181	0/181
226	D	61/181	34/181	27/181
135	A	41/181	33/181	8/181
124	A	42/181	30/181	12/181

Table C

Entropy	InfoGain	Receptor [d]	Positive [e]	Plotkin,2003 [f]	Smith 2004 [g]
1.107	0.1969		P	11	+
1.194	0.1286			13	+
0.9471	0.1255			10	+
1.274	0.0297	R	P	2	
1.062	0	R	P	1	
1.192	0		P	3	+

Figure 10. This diagram shows how to find immunodominant positions via calculation of information gain. Table A is the source data. We want to find which position change would affect antigenic type. Table B is some simple statistic result. For example, position 145 have total 62 changes in the full 181 cases and 61 of them occurred in the variant type. Table C is the comparison between entropy and information gain of selected positions. Specially note that position 135 with high entropy but zero information gain.

a. The number of total mutation times in the 181 training set.

b. The number of total mutation times in the variant type.

d. If the residue is in the receptor binding sites.

e. If the residue is under positive selection [8].

f. The rank of codon diversity of each residues[7].

g. If this residue leads to cluster transition [11].

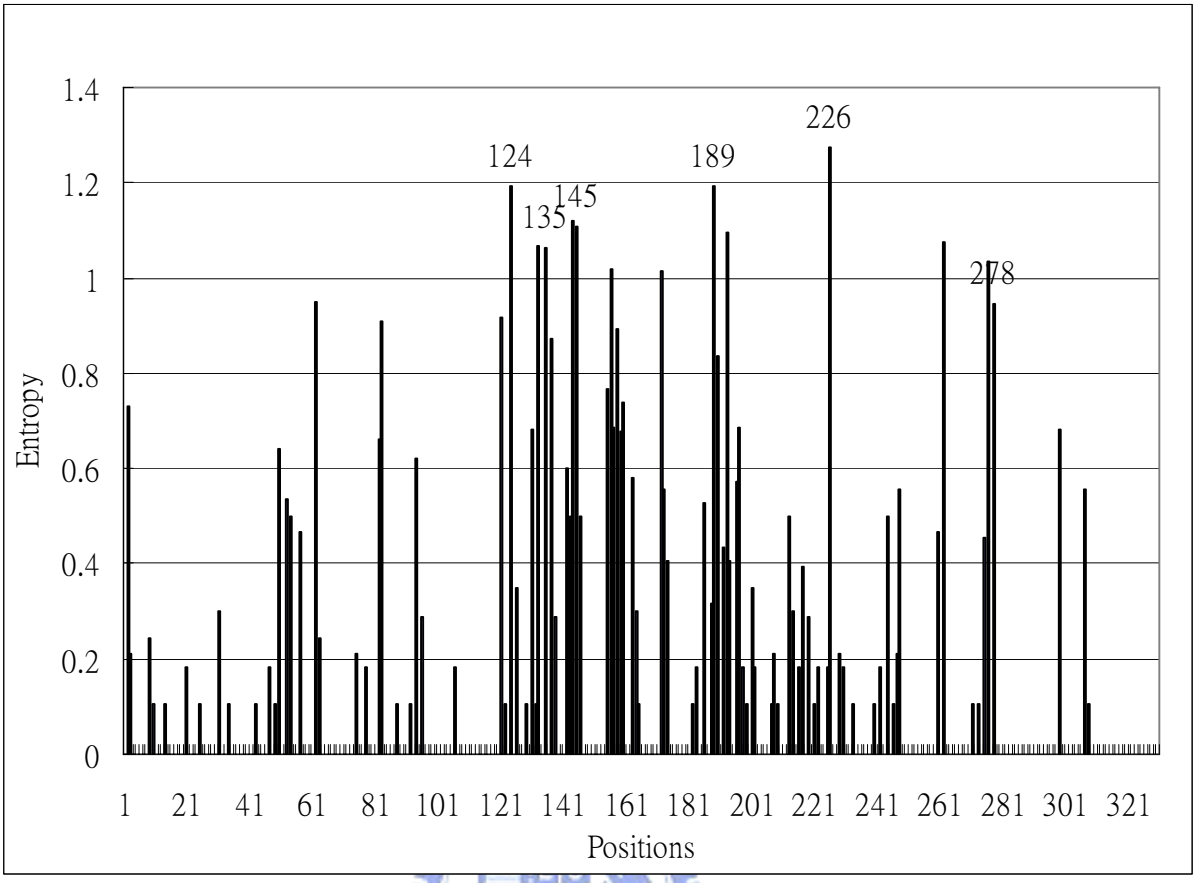


Figure 11. This figure shows the entropy of all 329 positions. The entropy is an index to evaluate the genetic evolution.

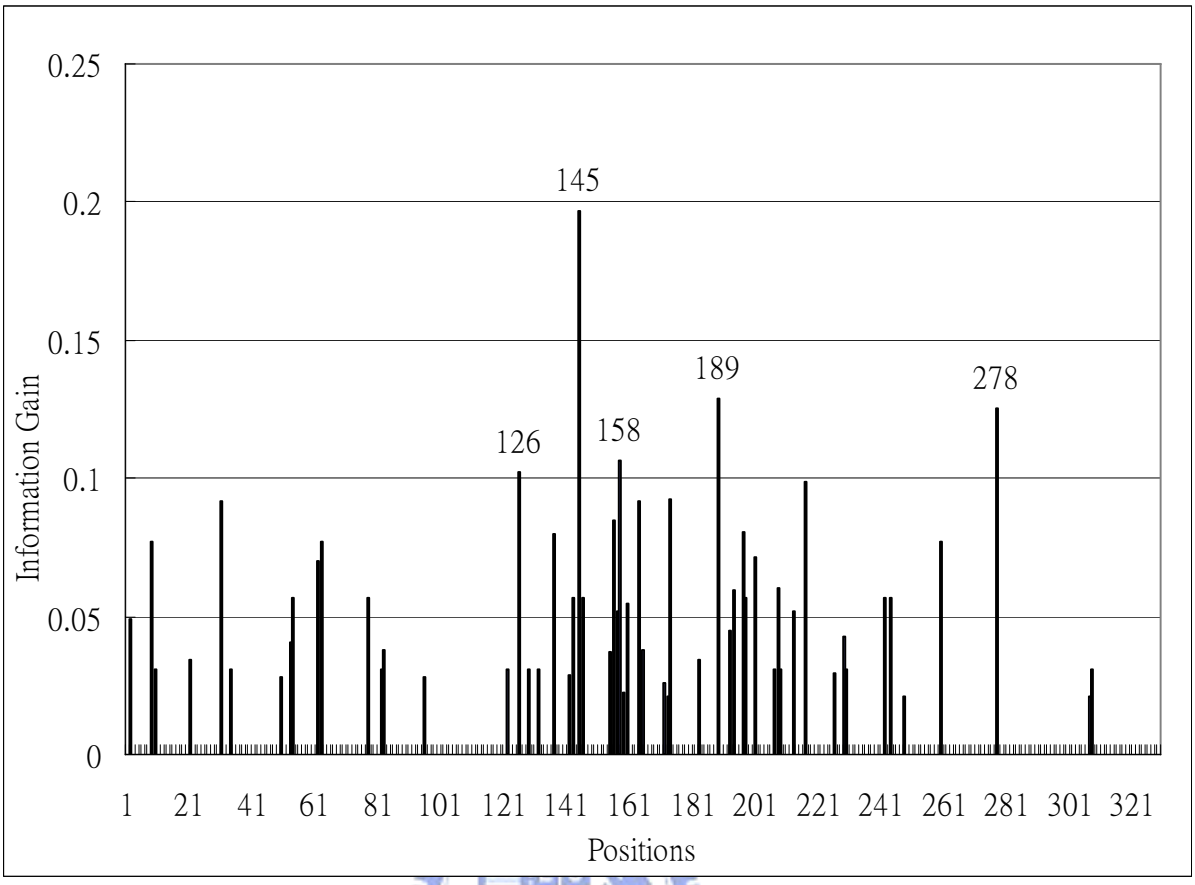


Figure 12. This figure shows the information gain of all 329 positions. The information gain may be a good index to represent the association degree between genetic and antigenic eolutions. Special note that position 145 have been verified by experiment that could lead to cluster transition [11].

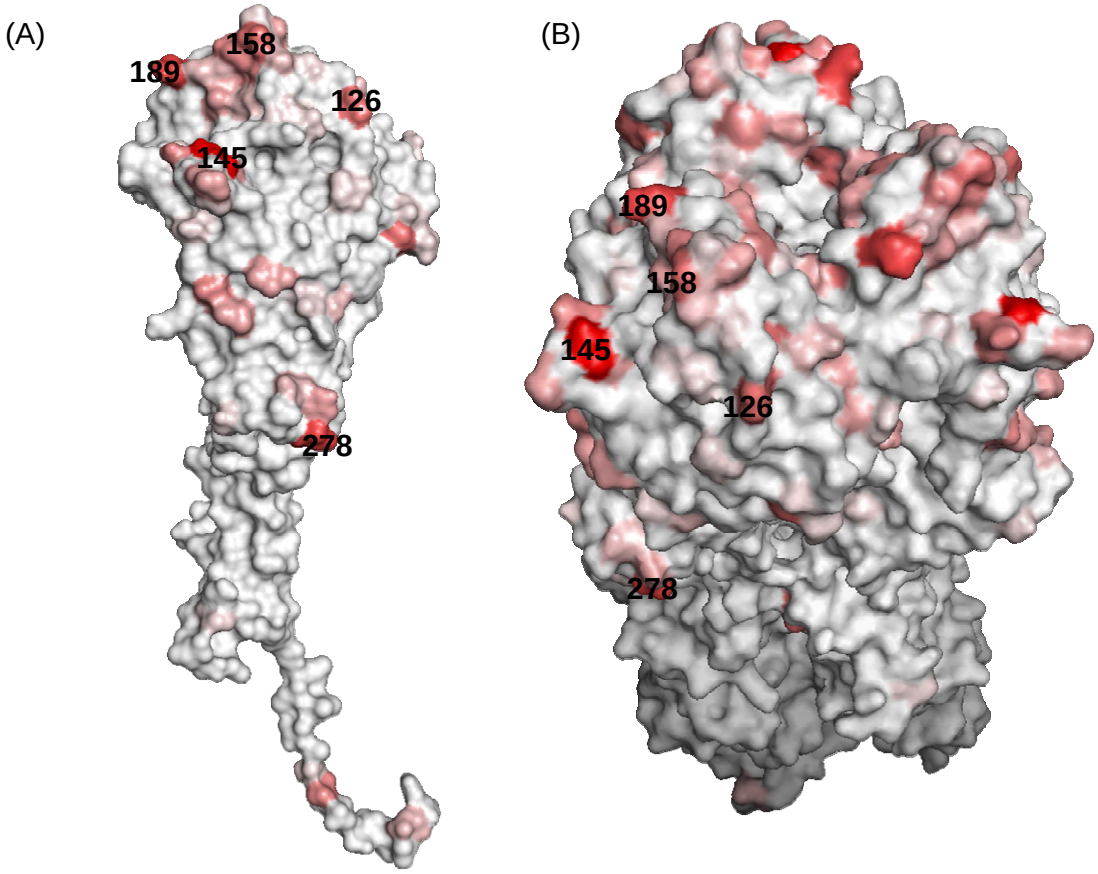


Figure 13.

This figure shows the information gain for 329 positions on the HA protein structure in the form of color. The red the color means more high the information gain and the top five information gain positions are labeled. Figure (A) is the front view of HA monomer. Figure (B) is the top view of HA trimer. Compare the red region between front view and top view shows that the top view show more high information gain positions.

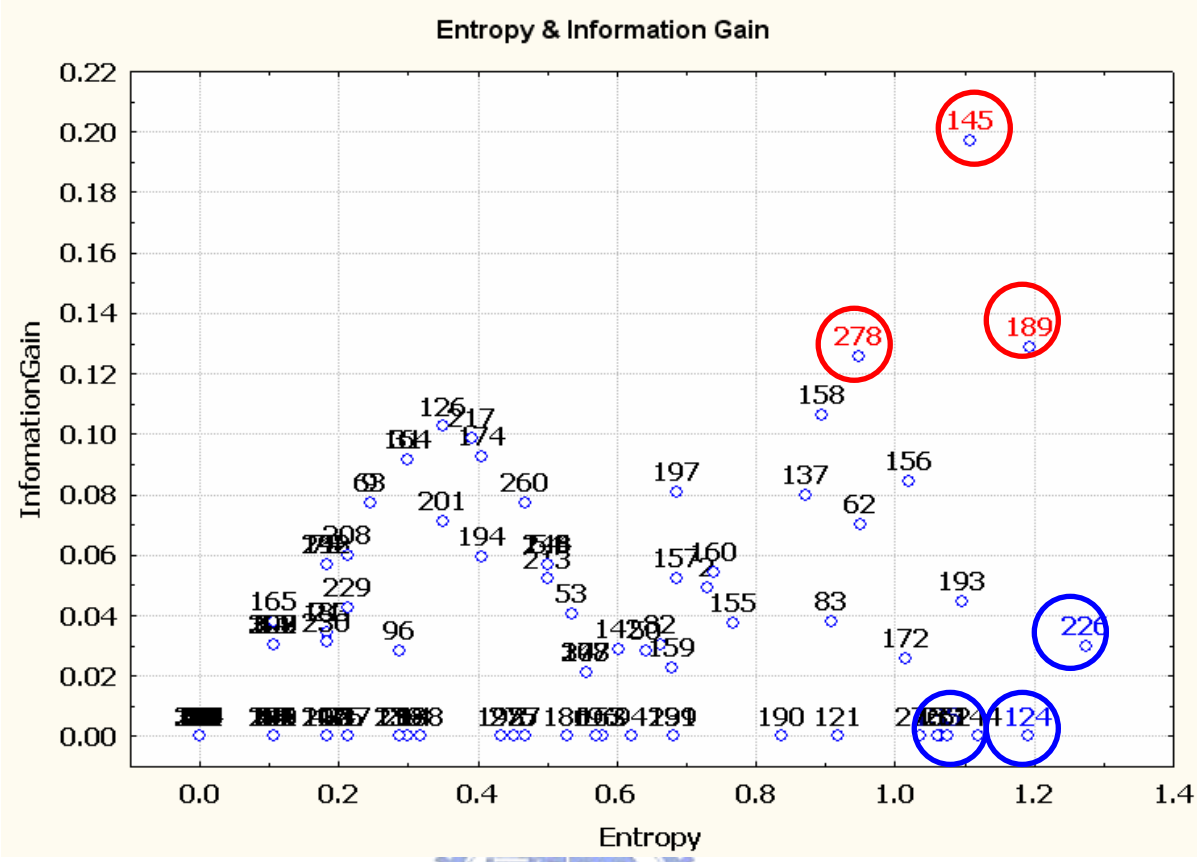


Figure 14. Evaluate each position's importance via entropy and information gain. The upper three positions (145 278 189) have both high entropy and information gain are potential immunodominant positions. The lower three positions(135 124 226) have only high entropy are consider as important sites in the genetic level but show low information gain. Special note that position 145 have been verified by experiment that could lead to cluster transition [11].

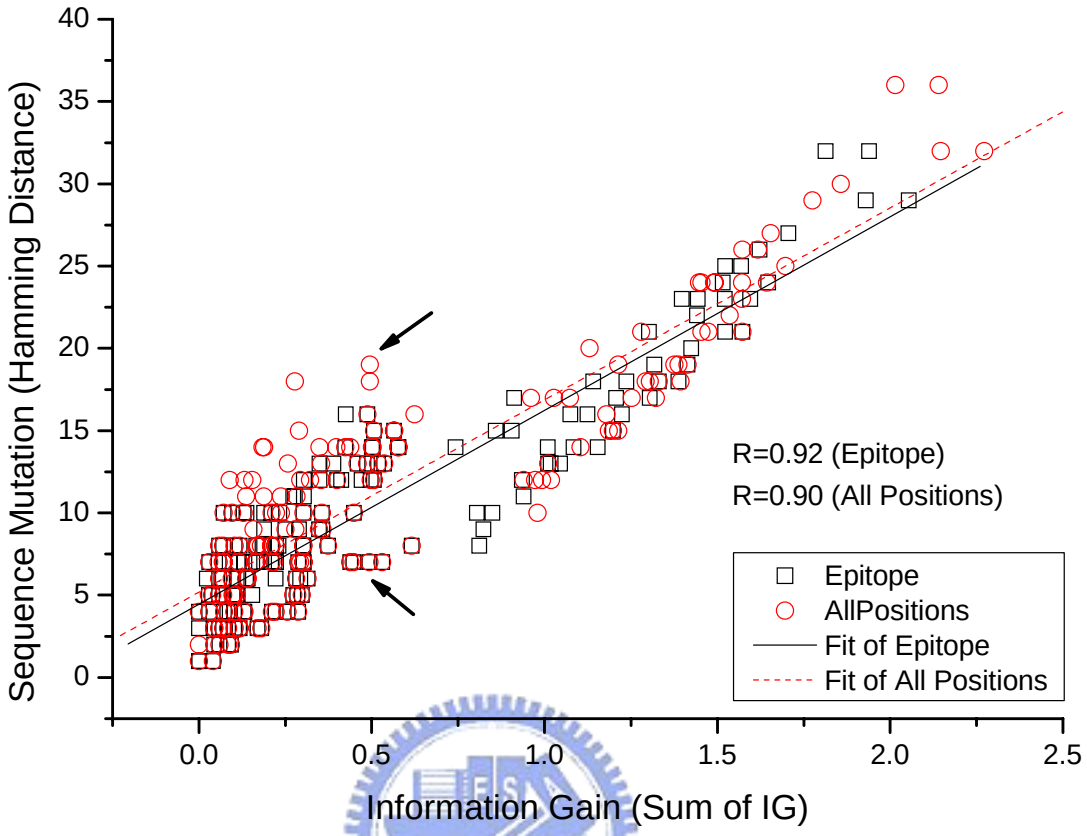


Figure 15. The relation between information gain and genetic evolution.

For each 181 cases, we compare the genetic changes and information gain for both All positions (329 positions) and epitope sites (131 positions). The linear regression R shows good relation (R >0.9) between genetic change and information gain and epitope sites could better fit the genetic change.

But for the same value of information gain, the genetic sequence may have high diversity change. For example the information gain value near 0.5, the position change number could range from 7 to 19. The result shows that information gain treats each position change with different weight, but not equal weight.

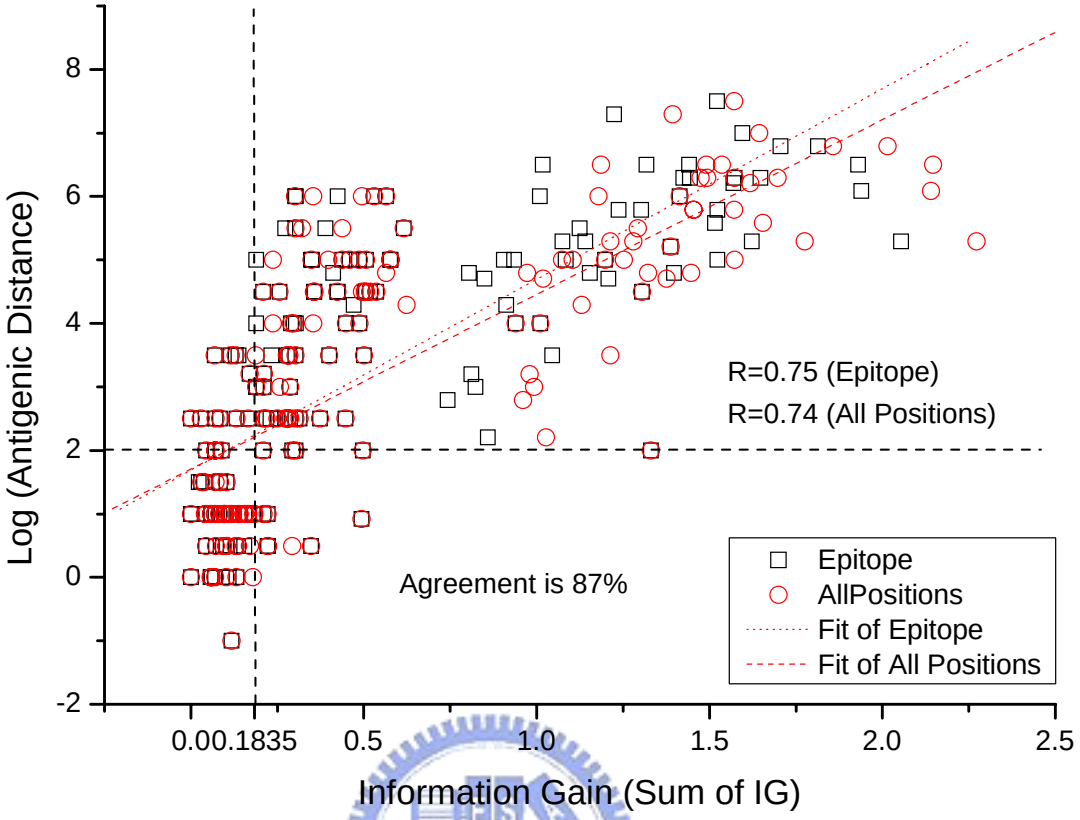


Figure 16.The relation between information gain and antigenic distance.

For each 181 cases, we compare the antigenic distance and information gain for both All positions (329 positions) and epitope sites (131 positions). The result shows that sum of information gain could fit the linear relation to antigenic distance (R >0.74). The result also shows that epitope could better fit the antigenic distance than all positions.

Antigenic variants are defined when antigenic distance ≧ 4, from this figure when sum of information gain > 0.1835 , we could get best predicting performance for predicting antigenic variant. The agreement is 87%.

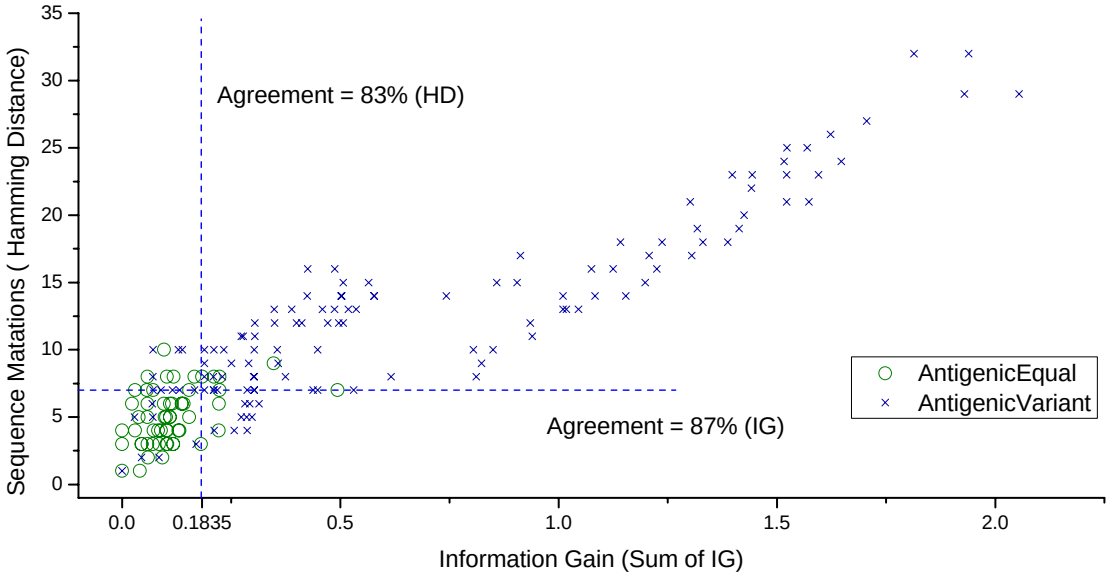


Figure 17.

For each 181 cases, the information gain and number of sequence mutations of epitope sites are plot on the figure. When the sum of Information gain value > 0.1835, the case is predicted as antigenic variant and the agreement is 87 % (158/181). When the sum of sequence mutations ≧ 7, the case is predicted as antigenic variant and the agreement is 83% (150/181). The different predicted case are illustrated in figure 17.

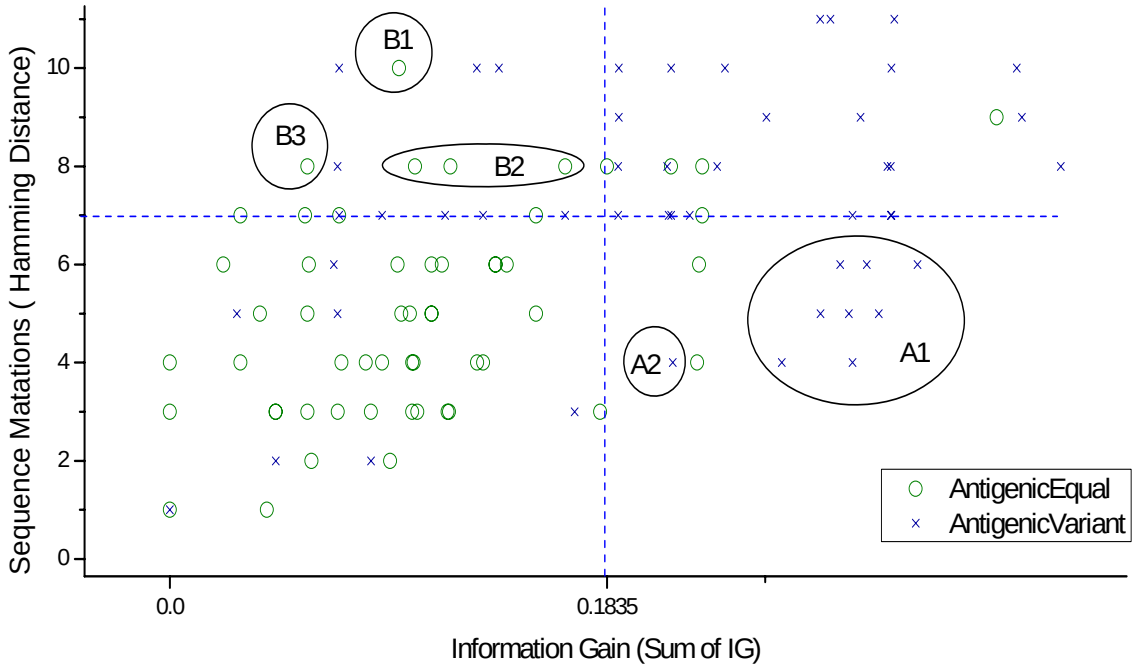


Figure 18.

The detail view of figure 16. Cases successfully predicted by information gain but false predicted by hamming distance are label with big circle. The cases in the circle A means little sequence mutations but which leads to antigenic variant pair. The cases in circle B means large sequence mutations but still antigenic equal pair. The result show that when sequence mutations are less than 11 positions, the position which actually changes would more important than the amount of total mutations. The 14 successful predicted cases are list in table 6.

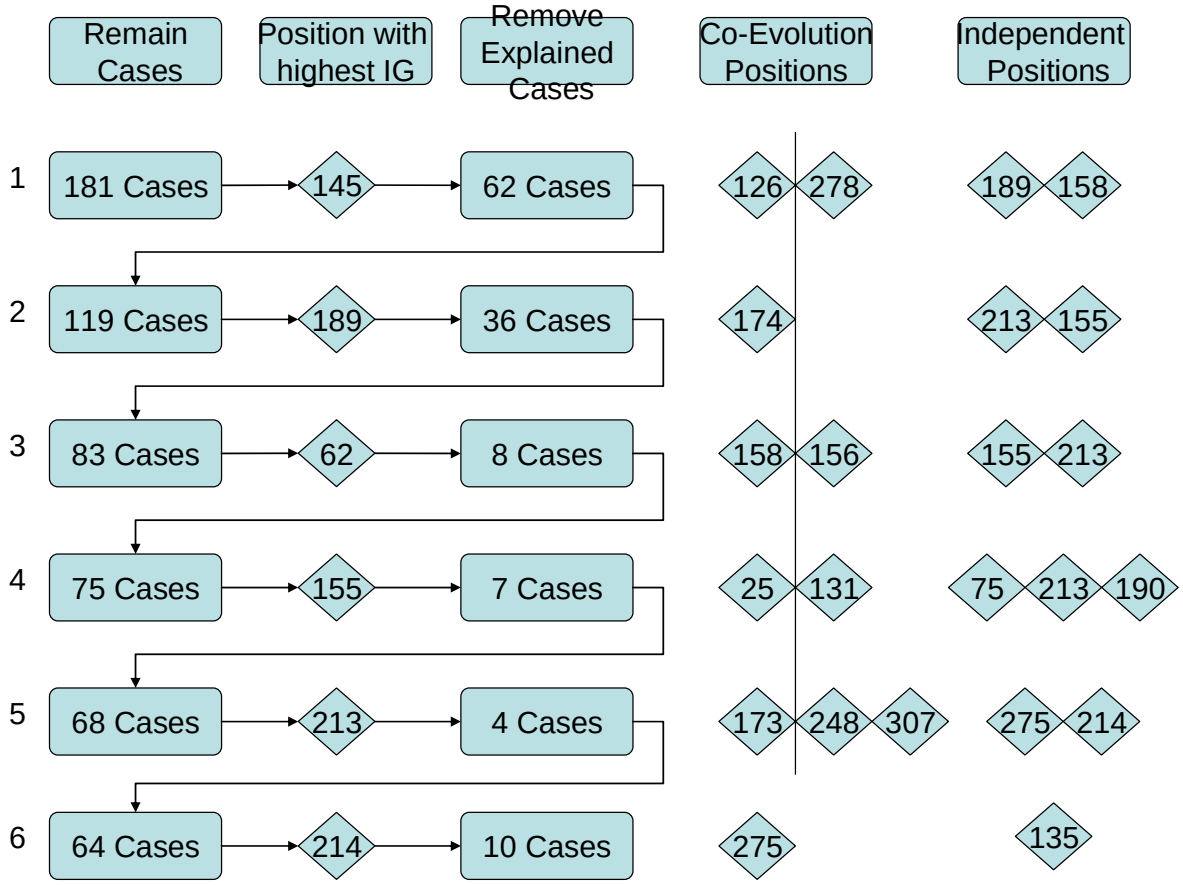


Figure 19.

We have 181 cases in the initial and then to find the position with highest information gain. Position 145 have the highest information gain in level one, so the first selected position is position 145. There are 62 cases in level one have position changes on 145, so those 62 cases are considered as explained by position 145 and removed from the original set. In each level we further consider those positions with information gain > average + 2*standard deviation. In level one there are other 4 positions satisfy the condition. The position 126 and 278 are considered as co-evolution to position 145 because when the 62 cases are removed from the dataset, their information drop significantly in the level two. The position 189 and 158 are considered as independent important positions would not drop information gain when the 62 cases are removed from the dataset. When the 62 cases are removed from the dataset, we use the same method to select the position with highest information gain in level two.

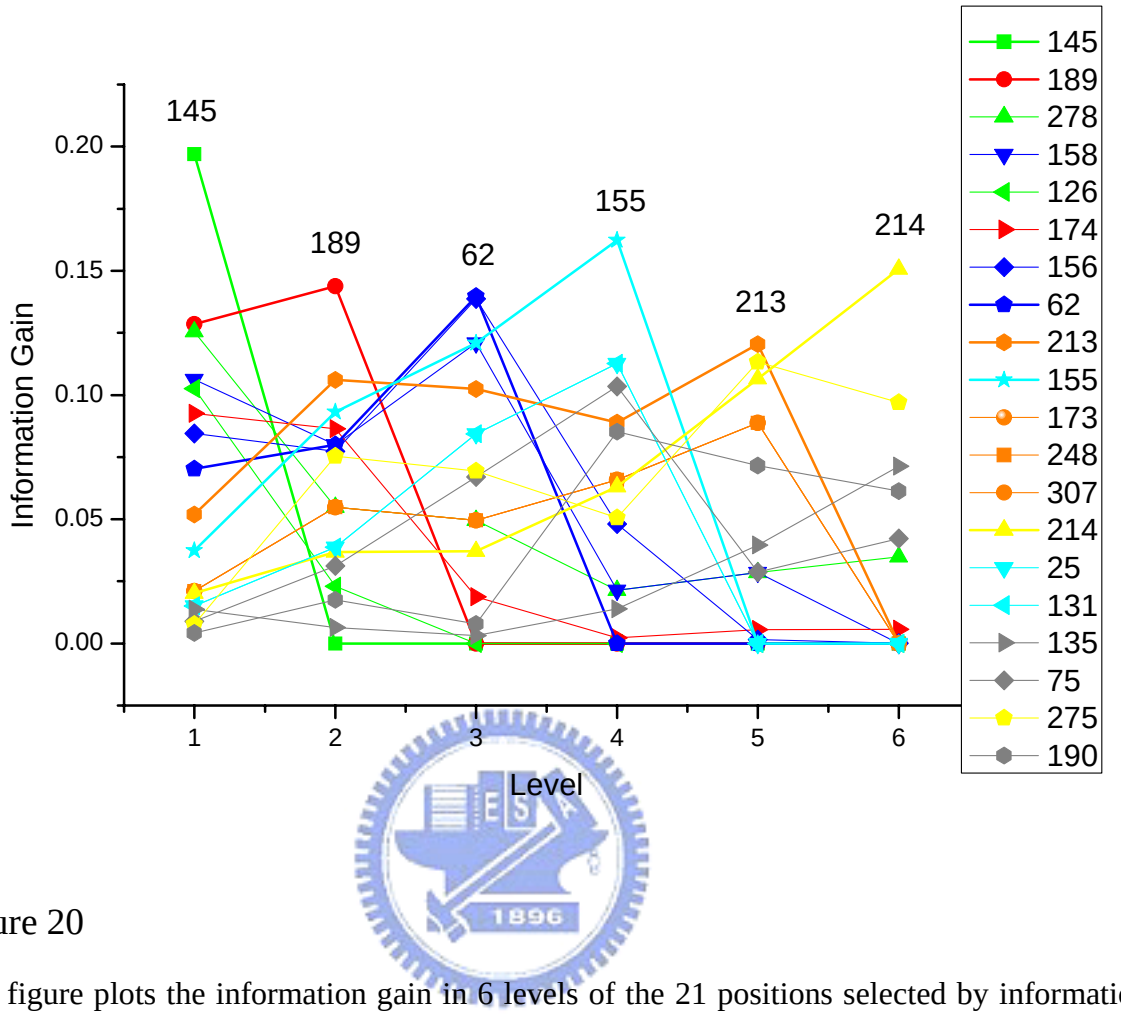


Figure 20

This figure plots the information gain in 6 levels of the 21 positions selected by information gain. In each level the selected position having the highest information gain. Positions with close information gain behavior are consider co-evolution groups are colored in the same color. For example the first group includes position 145, 278 and 126 are in green color.

Specially note that when the position with highest information gain is selected and those cases have mutation on that position is removed from the dataset. The information of that selected position would drops to zero. As a consequence, some positions (Ex: 155) do not have high information gain in the level one but it's information gain gradually increase from level one to level four.

(A)

```
145A > 0 : Variant (62/1)
145A <= 0 :
|   189B > 0 : Variant (36/5)
|   189B <= 0 :
|   |   62E > 0 : Variant (8/0)
|   |   62E <= 0 :
|   |   |   155B > 0 : Variant (7/0)
|   |   |   155B <= 0 :
|   |   |   |   213D > 0 : Variant (4/0)
|   |   |   |   213D <= 0 :
|   |   |   |   |   214D <= 0 : Equal (54/7)
|   |   |   |   |   214D > 0 : Variant (10/3)
```

All cases in this branch

Error cases in this branch

(B)

Level	Selected	Co-evolution	Independent	Epitope	Plotkin 2003	Smith ,2004
1	145A	126A,278C	189B,158B	A	11	+
2	189B	174D	213D, 155B	B	13	+
3	62E	158B,156B	155B,213D	E	21	+
4	155B	25,131A	75E,213D,190B	B		+
5	213D	173D,248D,307C	275C,214D	D		+
6	214D	275C	135A	D		

Figure 21

Figure (A) is the decision tree model of using decision tree tool C4.5 to select the positions with highest information gain. The nodes of decision tree are the positions with highest information gain in each level. The root is on the top of the tree and we should read the tree begin from root. The condition “145 > 0 : Variant (62/1)” means that when position 145 changes and the predicting type is variant. There are 62 cases have change on position 145 and 61 of them are variant type and only one cases disobey the rule. When position 145 have no change (145<=0) , then go to check the next position 189.

Figure (B) show the selected positions with highest information gain in each levels and the co-evolution positions with that selected position. The last three columns denote the biology meaning and related work’s remark of that selected position.

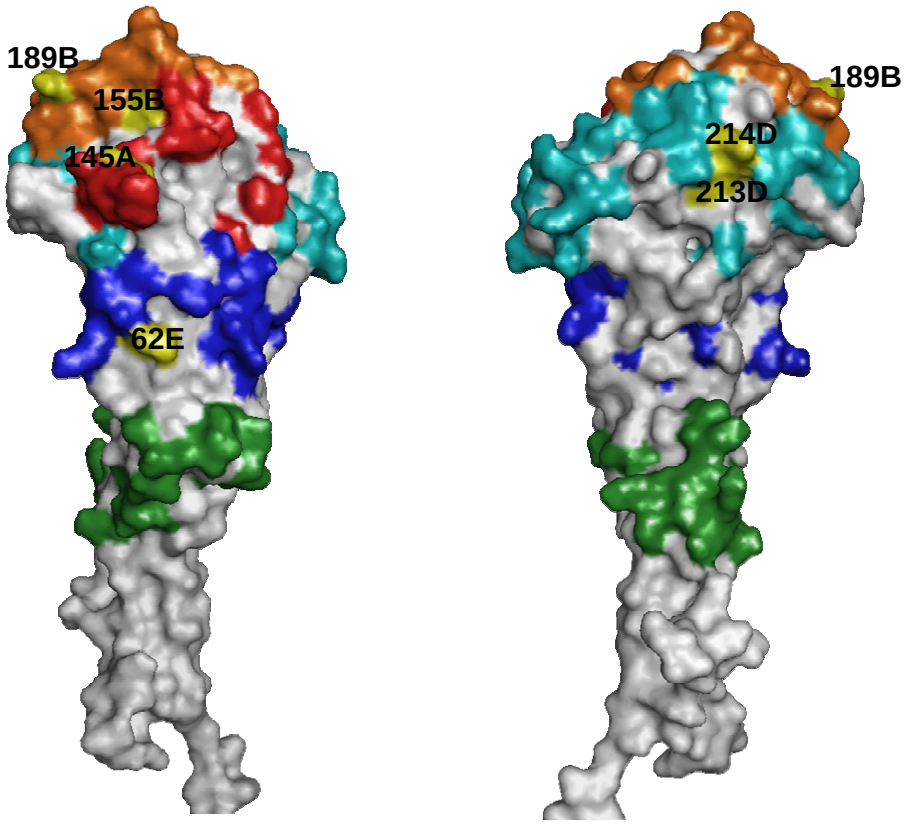


Figure 22

The positions selected by information gain in each level are label on the HA protein structure.

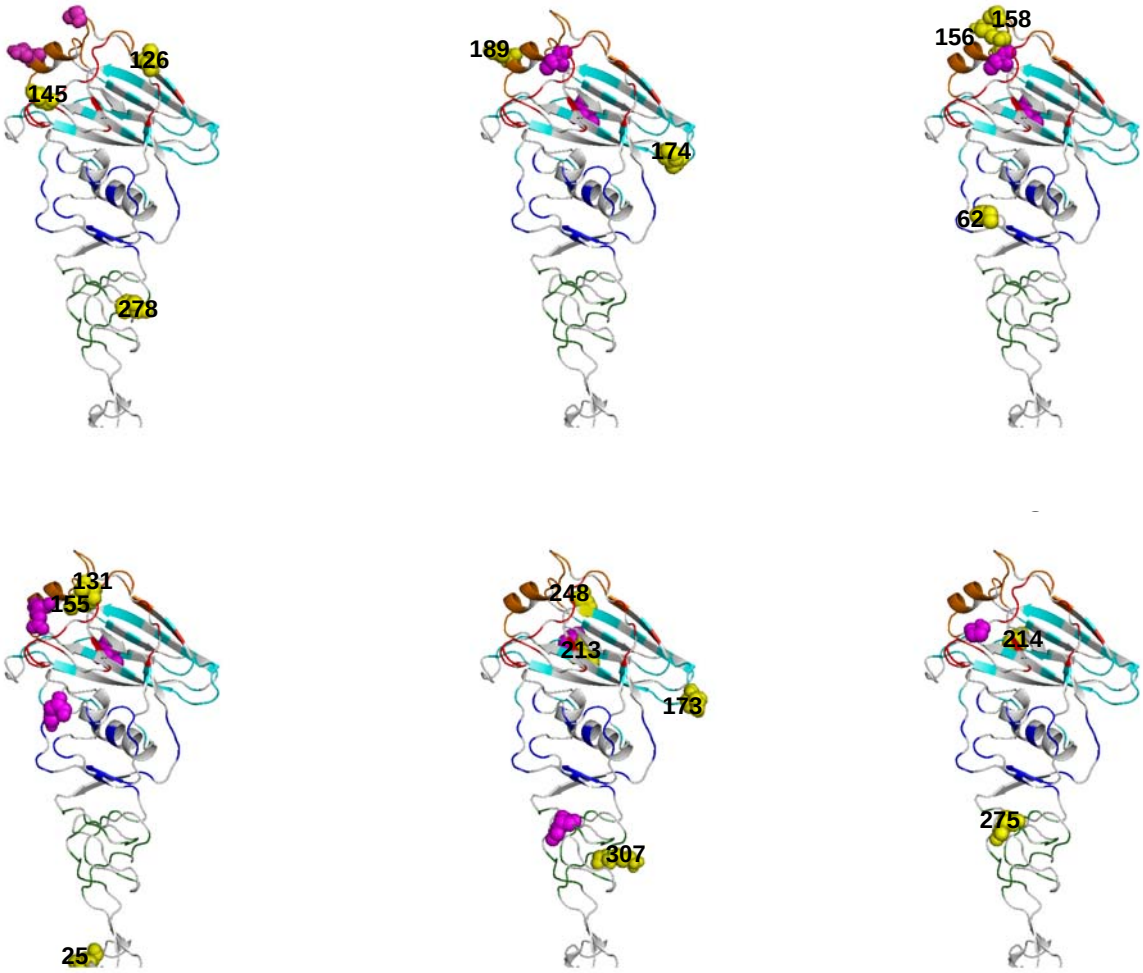


Figure 23

The important positions selected in each level.

Co-evolution sites are space filled in color yellow. From the structure view, the co-evolutions all locate on at least two epitope sites. This result match Wilson's conclusion for drift variant of epidemiologic importance.

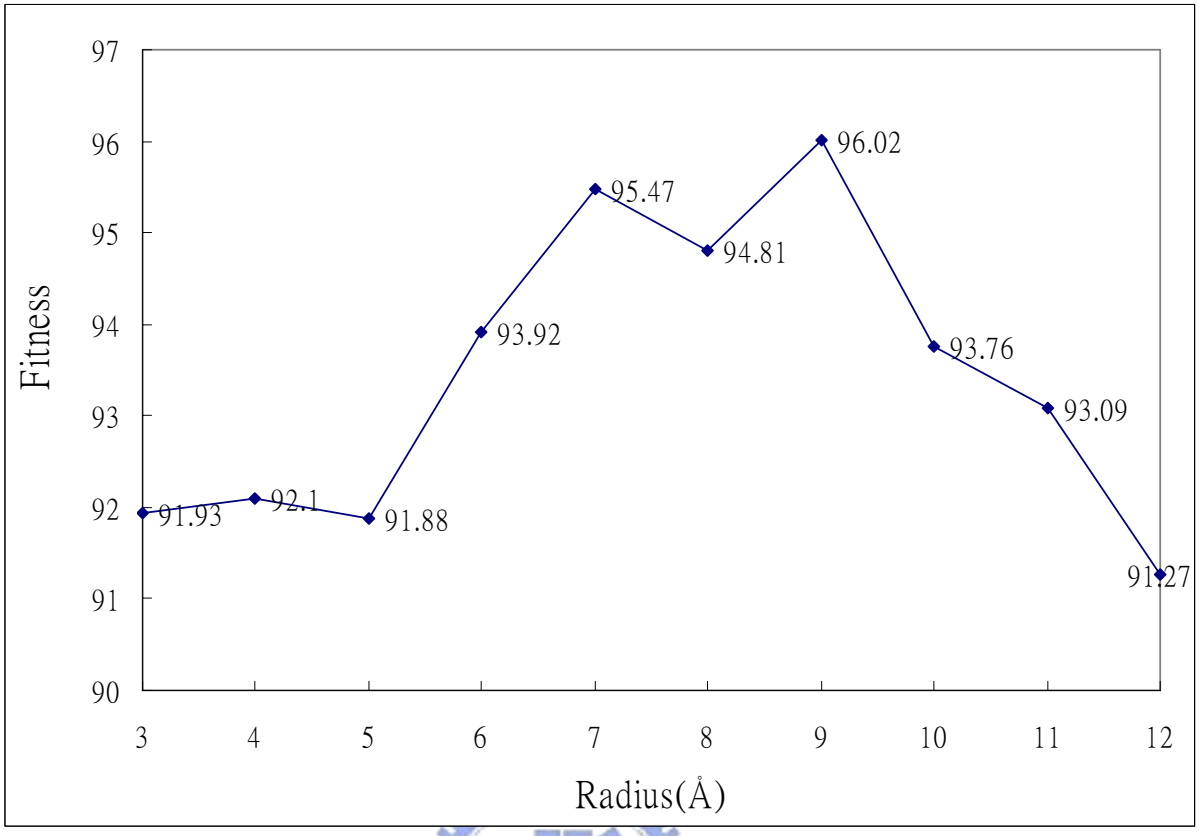


Figure 24. The predicting performance of different radius from 3 Å to 12 Å in contact map coding. As a result we choose the radius of region at 9 Å.

(A)

```
160B > 0 :
|   134- > 0 : Variant (91/1)
|   134- <= 0 :
|   |   212D > 0 : Variant (9)
|   |   212D <= 0 :
|   |   |   205- <= 0 : Equal (11/1)
|   |   |   205- > 0 : Variant (2)
160B <= 0 :
|   206- > 0 : Variant (8)
|   206- <= 0 :
|   |   199- <= 0 : Equal (35/3)
|   |   199- > 0 :
|   |   |   144A <= 0 : Equal (11/2)
|   |   |   144A > 0 :
|   |   |   |   51C > 0 : Variant (8)
|   |   |   |   51C <= 0 :
|   |   |   |   |   96D <= 0 : Equal (4)
|   |   |   |   |   96D > 0 : Variant (2)
```

```
Evaluation on training data (181 items):

           Decision Tree
         ----------------
         Size      Errors   Estimate

           10    7( 3.9%)    (11.4%)   <<


          (a)    (b)   <-Classified as
         ----   ----
          119      6   (a): Class Variant
            1     55   (b): Class Equal
```

(B)

Center	Region Covered	Changed
160B	T129B,130A,T156B,T157B,T158B,T159B,T160B, 161-,162-,195-,T196B,T197B,	8
134	T131A,T132A,T133A,134-,T135A,136- ,T145A,T146A,147-,152A,153-,154-,T155B,	7
212D	T201D,T202-,203D,204-,205-,210-,211- ,212D,T213D,T214D,	4
205	203D,204-,205-,206-,T207D,T208D,T209D,210- ,211-,212D,235-,T242D,243-,T244D,245-,T246D,	6
206	204-,205-,206-,T207D,T208D,T209D,210-,235- ,236-,237-,238D,241-,T242D,243-,T244D,	5
199	T188B,191-,T192B,195-,T196B,T197B,T198B, T199-,200-,T201D,T214D,215D,T216D,T217D, T248D,249-,250-,	11
144A	T135A,136-,T137A,139-,140A,141- ,T142A,T143A,T144A,T145A,T146A	7
51C	48C,T49-,T50C,51C,52-,T53C,T54C,58-,272- ,T273C,274-,T275C,T276C,277-,279C,282-,283- ,284-,285-,286-,287-,288-,	7
96D	68-,73-,74-,T75E,T94E,95-,T96D,97-,98-,99- ,T138A,139-,140A,224-,	4

(C)

Total	A	B	C	D	E	Epitope
102	14/19	12/22	9/27	17/41	2/22	54/131

Figure 25. This figure illustrates the decision tree bases on contact map.

(A) Each node in the tree represents a region. (B) The center of each region is list in the table. Position label with “T” means have position change on this set. (C) There are 102 residues covered by all selected regions and 54 of them are in the five epitope sites.

Specially note the first rule : when region 160 and region 134 both occur changes then 90 cases of total 91 cases lead to antigenic change, in which this rule match Wilson’s conclusion for drift variant of epidemiologic importance[5]. This rule imply am important co-evolution relation between this two regions.

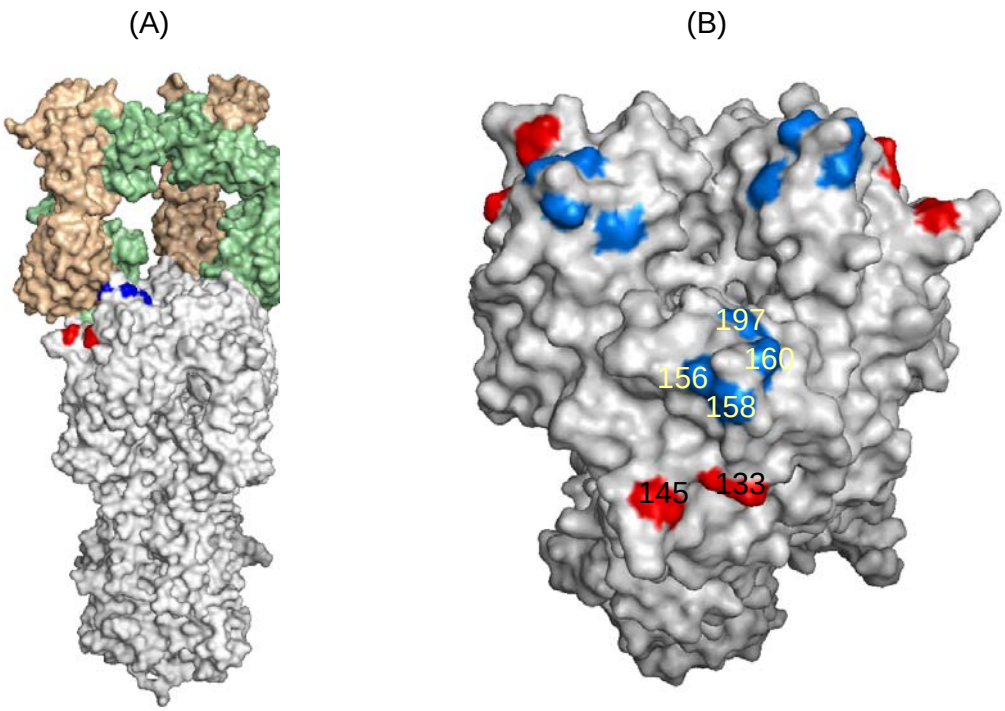


Figure 26

Figure A is the HA protein with neutralizing antibody binding (PDB 1KEN). The two regions selected by contact map are labeled in red and blue color. According to the figure, the two regions could directed block by antibody. Figure B is the top view of the HA trimer.

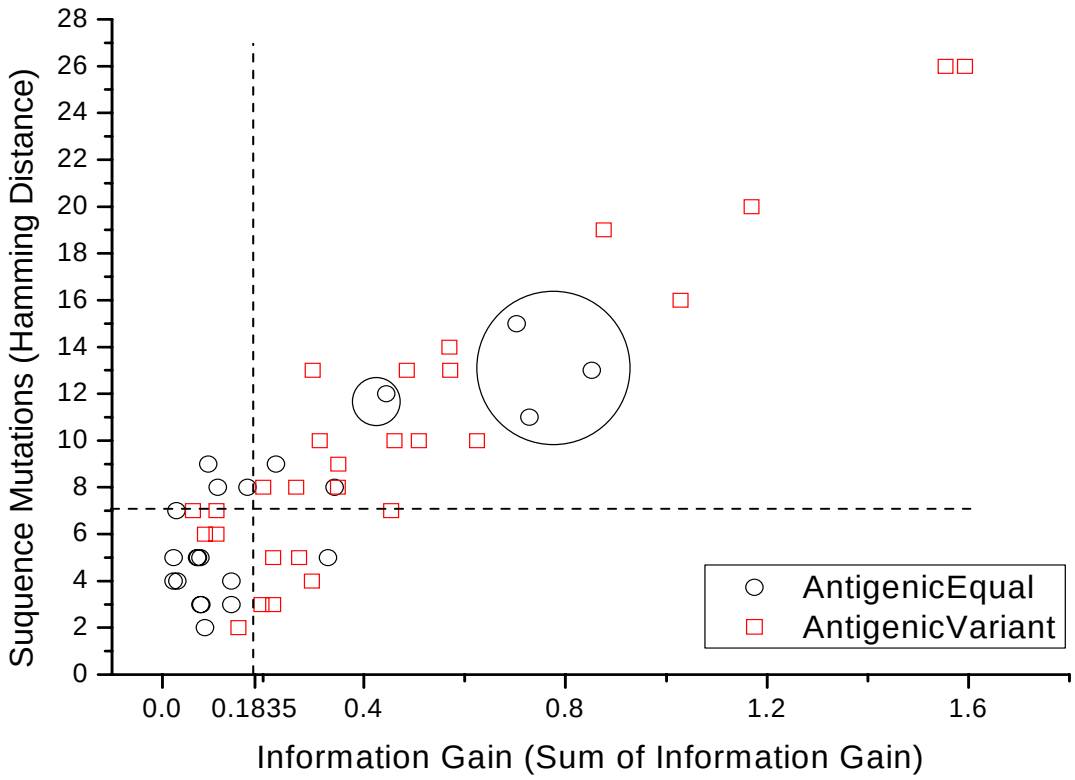


Figure 27.

This figure compare the information gain and sequence mutations of the WER 50 cases.

The four cases marked by circle have sequence mutations on epitope more than 11 positions but still antigenic equal type are false predicted by both two methods. They are

A/Hong_Kong/1/68 vs A/England/42/72:15 mutations

A/Shanghai/31/80 vs A/Bangkok/1/79:13 mutations

A/Texas/1/77 vs A/Belgium/2/81:11 mutations

A/Belgium/2/81 vs A/Philippines/2/82: 12 mutations

The detail information of these 50 cases are listed in table 11.

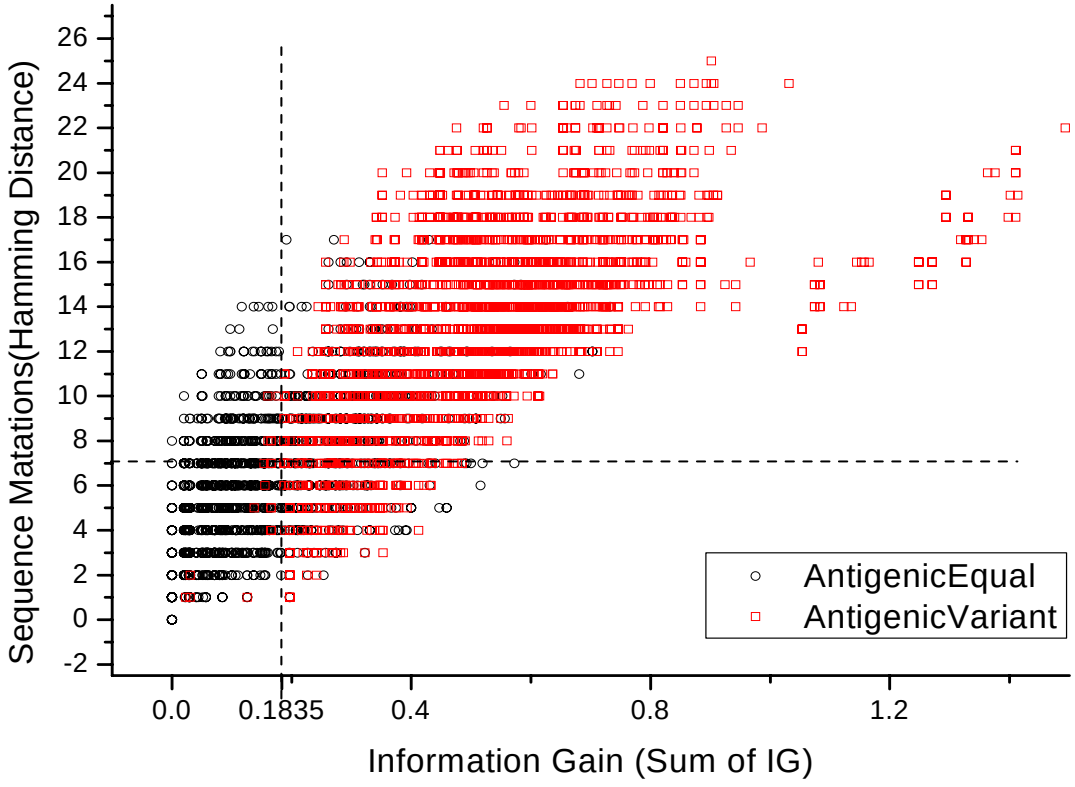


Figure 28

This figure compares the information gain and sequence mutations of the 5928 cases from Smith 2004. The result shows that information gain could predict majority antigenic variant type while the sequence mutation number could not.

(A)

```
145A > 0 : Variant (62/1) 1
145A <= 0 :
|   189B > 0 : Variant (36/5) 2
|   189B <= 0 :
|   |   62E > 0 : Variant (8/0) 3
|   |   62E <= 0 :
|   |   |   155B > 0 : Variant (7/0) 4
|   |   |   155B <= 0 :
|   |   |   |   213D > 0 : Variant (4/0) 5
|   |   |   |   213D <= 0 :
|   |   |   |   |   214D <= 0 : Equal (54/7) 6
|   |   |   |   |   214D > 0 : Variant (10/3) 7
```

(B)

Level	Position	Total Cases	Right Cases	Wrong Cases
1	145	3098	2986	112
2	189	343	261	82
3	62	517	406	111
4	155	91	91	0
5	213	60	25	35
6	214	1449	1420	29
7	214	370	12	358
Total Cases		5928	5201	727
Percentage		100.00%	87.74%	12.26%

Figure 29.

(A) is the important positions selected by information gain and each position is label with level number. (B) is the test result of test set 2 which having 5928 cases. Take level 1 for example, there are total 3098 cases having position changed on position 145 and 2986 of them are antigenic variant type which means a successfully prediction, in other wise if the antigenic type is equal then this is a false prediction. The majority false prediction were due to position 214 and the majority false predicting cases in test set 1 (WER 50 cases) also caused by position 214.

Reference

1. *World Health Organization.WHO position paper influenza vaccines.* 2002.
2. *World Health Organization. Recommendations for the use of inactivated influenza vaccines.* 2000.
3. Subbarao, K. and J.M. Katz, *Influenza vaccines generated by reverse genetics.* Current Topics in Microbiology and Immunology, 2004. **283**: p. 313-342.
4. Wiley, D.C., I.A. Wilson, and J.J. Skehel, *Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation.* Nature, 1981. **289**(5796): p. 373-378.
5. Wilson, I.A. and N.J. Cox, *Structural basis of immune recognition of influenza virus hemagglutinin.* Annual Review of Immunology, 1990. **8**: p. 737-771.
6. *World Health Organization.WHO Manual on Animal Influenza Diagnosis and Surveillance.* 2002.
7. Plotkin, J.B. and J. Dushoff, *Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(12): p. 7152-7157.
8. Bush, R.M., et al., *Positive selection on the H3 hemagglutinin gene of human influenza virus A.* Molecular Biology and Evolution, 1999. **16**(11): p. 1457-1465.
9. Bush, R.M., et al., *Predicting the evolution of human influenza A.* Science, 1999. **286**(5446): p. 1921-1925.
10. Plotkin, J.B., J. Dushoff, and S.A. Levin, *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): p. 6263-6268.
11. Smith, D.J., et al., *Mapping the antigenic and genetic evolution of influenza virus.* Science, 2004. **305**(5682): p. 371-376.
12. Lee, M.S. and J.S. Chen, *Predicting antigenic variants of influenza A/H3N2 viruses.* Emerging Infectious Diseases, 2004. **10**(8): p. 1385-1390.
13. Fitch, W.M., et al., *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): p. 7712-7718.
14. Macken CA, L.H., Goodman L, Boykin L, *The value of a database in surveillance and vaccine selection.* Options for the control of influenza IV, 2001: p. 103-106.
15. Both, G.W., et al., *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): p. 52-60.
16. *World Health Organization. Recommended composition of influenza virus vaccines for use in the 1988–1989 season.* Wkly Epidemiol Rec, 1988. **63**: p. 57-9.
17. Ellis, J.S., P. Chakraverty, and J.P. Clewley, *Genetic and antigenic variation in the*

haemagglutinin of recently circulating human influenza A (H3N2) viruses in the United Kingdom. Arch Virol, 1995. **140**(11): p. 1889-904.

18. *Centers for Disease Control and Prevention. Information for FDA vaccine advisory panel meeting. Atlanta: The Centers.* 1997: p. 30.
19. Coiras, M.T., et al., *Rapid molecular analysis of the haemagglutinin gene of human influenza A H3N2 viruses isolated in spain from 1996 to 2000.* Arch Virol, 2001. **146**(11): p. 2133-47.
20. *Centers for Disease Control and Prevention. Information for the Vaccines and Related Biological Products Advisory Committee, CBER, FDA. Atlanta: The Centers.* 2003: p. 28.
21. Bernstein F C, K.T.F., Williams G J B, Meyer E F Jr, Brice M D, Rogers J R, Kennard O, Shimanouchi T & Tasumi M, *The Protein Data Bank: a computer-based archival file for macromolecular structures.* Journal of Molecular Biology, 1977. **112**: p. 535-542.
22. Kilbourne, E.D., B.E. Johansson, and B. Grajower, *Independent and disparate evolution in nature of influenza A virus hemagglutinin and neuraminidase glycoproteins.* Proc Natl Acad Sci U S A, 1990. **87**(2): p. 786-90.
23. Archetti, I. and F.L. Horsfall, Jr., *Persistent antigenic variation of influenza A viruses after incomplete neutralization in ovo with heterologous immune serum.* J Exp Med, 1950. **92**(5): p. 441-62.
24. Quinlan, J.R., *C4.5: Programs for Machine Learning.* 1993.
25. *World Health Organization.WHO Report on Global Surveillance of Epidemic-prone Infectious Diseases.* 2000: p. 93.