

Identification of Cross-Neutralization Determinants by GAP Analysis: A Mutational Behavior Approach

Fusheng Li*, Peter B. Gilbert and Steve G. Self

Statistical Center for HIV/AIDS Research and Prevention, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, Seattle, WA 98109, USA

Abstract: Antigenic variation, which is the result of amino acid substitution and genetic evolution, poses a major challenge in the development of vaccines against pathogens with unstable genomes, such as HIV-1 and Influenza. Thus, it is highly important to characterize the relationship of genetic evolution, antigenic variation and positional mutation (GAP) from the perspective of vaccine development. For this purpose, we introduce an automatic and simple GAP analysis approach, which is based on the fundamental concept of "mutational behavior" in genomic research, to predict the specificity-determining positions in protein families. This approach identifies cross-neutralization determinants by mutational behavior correlation to antigenic and genetic changes. Correlated mutation analysis and structure mapping further refine the cross-neutralization determinants. The usability of this approach is confirmed by analysis of an Influenza H3N2 cross-reactive dataset.

Keywords: Neutralizing antibody, cross-neutralization, CDPs, GAP, vaccine, Influenza, HIV-1.

1. INTRODUCTION

Most infectious diseases caused by slowly mutating DNA viruses can now be controlled effectively through the use of vaccines. However, viruses with unstable genomes (mostly RNA viruses) are still a major threat to human life. The antigenic variation of these pathogens either necessitates frequent updating of vaccine strains (Influenza) or makes the vaccine a total failure (HIV-1) [9]. Antigenic variation is the consequence of amino acid substitution and genetic evolution. Therefore, GAP analysis is considered highly important from the perspective of vaccine development. In the past, different approaches have been proposed for this analysis, most of them focused on the Influenza virus [2, 3, 8, 26]. These analyses fall into two general categories: (1) comparing the antigenic and genetic evolution pattern, by phylogenetic tree, hierarchical clustering or proximity map; (2) identifying the residues with positive selection or the residues responsible for cluster shift. While these analyses greatly enhanced our understanding of antigenic and genetic evolution patterns, questions still remain. In particular, what genetic factors determine the cross-neutralization patterns among variant viruses? To answer this question, an automatic and quantitative approach is required.

Identification of cross-neutralization determinants is, in a way, similar to the practice in genomic research of predicting the specificity-determining positions (or tree-determining positions, TDPs) in protein families. Usually, protein families evolved from a common ancestor contain multiple subfamilies with similar yet distinct functions. For example, the human GPCR family with endogenous ligands alone consists of 367 receptors. These receptors can be further divided into subfamilies according to their genetic relationship [28].

GPCRs in different subfamilies recognize distinct ligand types (i.e., peptide, neurotransmitter) and hence, different functionalities. Thus, a common challenge of sequence analysis for protein families such as GPCR is to identify the residues that determine the functional specificity of proteins with a common general function (i.e., binding ligand type). Three basic methods have been proposed for this purpose and tested in various protein families. The first method takes a phylogenetic representation of a protein family as a starting point and, following the principle of Information Theory, automatically searches for the optimal division of the family into subfamilies [12, 13, 19, 20]. The second method looks for positions whose mutational behaviors are reminiscent of the mutational behavior of the full-length proteins by directly comparing the corresponding distance matrices. The third method, called "sequence space analysis," is an automation of the analysis of distribution of sequences and amino acid positions in the corresponding multidimensional spaces using a vector-based principal component analysis [4]. Among these three methods, the second ("mutational behavior analysis") has been demonstrated to be simple, flexible and powerful [5]. This method assumes that the variation pattern of the specificity-determining positions should reflect that of the entire protein family. Although this method does not use any explicit representation of the family phylogenetic tree, as does the first method, the phylogenetic information is contained in the protein distance matrix. Therefore, it is flexible and independent of any tree-construction algorithm.

Usually, mutational behavior analysis compares the positional mutation pattern to genetic change. Mutational behaviors of two different positions can also be compared in order to detect correlated mutations [10]. Due to structural and functional constraints, mutations at different positions must coordinate. One escaping mutation may be accompanied by one or more compensatory mutations at other positions. It is important to differentiate these two kinds of mutations. Positions with correlated mutation are believed to be in close contact. Indeed, a discernible trend was observed when

*Address correspondence to this author at the Statistical Center for HIV/AIDS Research and Prevention, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, Seattle, WA 98109, USA; Tel: (206) 667 6470; Fax: (206) 667-4812; E-mail: Fusheng@scharp.org

comparing the correlated mutation with distance from the crystal structure: a residual pair with higher correlation has shorter distance [10]. Correlated mutations were characterized by the correlation coefficient between two amino acid similarity matrices [10]. Alternatively, correlated mutations can also be calculated by treating amino acids as discrete variables [15]. Although sequence space analysis was proposed as a way to predict functional residues in proteins, the patterns defined by the principal components also contain the information of correlated mutations [4].

Based on the fundamental concept of mutational behavior analysis, we propose an automatic approach for identifying the cross-neutralization determinants. For a collection of related viruses, cross-reactivity among these viruses can be considered as "specificity." The mutational behavior of cross-neutralization determining positions (CDPs) should reflect the pattern of cross-neutralization and therefore can be identified by the high correlation between positional amino acid dissimilarity and antigenic distance. TDPs are also identified and jointly interpreted with CDPs. Furthermore; mutational behavior analysis is applied to find the potential correlated mutations and the high level correlation patterns. This whole approach is illustrated by identification of cross-neutralization determinants from an Influenza H3N2 dataset. CDPs known to be involved with neutralizing antibody are successfully identified by this method. Moreover, correlated mutation analysis reveals more interesting patterns, which may be worth further investigation.

2. MATERIALS AND METHODS

2.1. Cross-Reactive Data Set

To demonstrate the mutational behavior approach, we selected a small Influenza H3N2 cross-reactive antibody dataset. This dataset contains 11 isolates from season 1971-1979 [1, 18]. Cross-reactivity is based on a hemagglutination inhibition (HI) assay, which is a binding assay testing the ability of influenza viruses to agglutinate red blood cells and the ability of animal antisera raised against the same or related strains to block this agglutination. HA1 polypeptide sequences were obtained by searching the Los Alamos Influenza Sequence Database (<http://www.flu.lanl.gov>) or they were entered from the original publication if they were not available in the database. For easy comparison, some ambiguous codons caused by low sequence quality were changed to the consensus sequence.

2.2. Genetic Distance, Antigenic Distance and Amino Acid Dissimilarity

Mutational behavior analysis relies on the comparison of distance (and dissimilarity) matrices. The genetic distance matrix was calculated using the PROTDIST program from PHYLIP package (<http://evolution.genetics.washington.edu/phylip.html>). The antigenic distance matrix was generated as described earlier [18]. For two different isolates (i.e., a and b), if homologous titers of two viruses are T_a and T_b and two heterologous titers against each other are $T_{a,b}$ and $T_{b,a}$, the antigenic distance $D_{a,b}$ (or $D_{b,a}$) is defined as

$$D_{a,b} = \sqrt{(T_a \times T_b) / (T_{a,b} \times T_{b,a})} \quad (1)$$

Positional amino acid dissimilarity is based on the BLO-SUM62 amino acid substitution matrix [11]. The values in the matrix are the log odds ratio between the observed amino acid pair frequencies and those expected by chance, which roughly reflect the amino acid similarity of physico-chemical properties. Since these values are the amino acid similarity measure, they are simply negated to dissimilarity for mutational behavior analysis. For a multiple alignment with n positions, n positional amino acid dissimilarity matrices can be generated.

2.3. Genetic and Antigenic Maps

Usually, genetic and antigenic relationships are visualized by a phylogenetic tree (for genetic distance) or a hierarchical cluster (for genetic and antigenic distances). Recently, an "antigenic map" was introduced with a geometric interpretation of binding assay data, which is based on the theoretical concept of "shape space" [7, 17, 23]. Multidimensional scaling (MDS) is used to position the isolates in the genetic or antigenic map. This algorithm is performed so that the distances from the proximity genetic or antigenic maps match those from the distance matrix. This feature of antigenic mapping leads researchers to believe that an antigenic map can provide greater resolution than an assay in cross-reactivity prediction [26].

2.4. Correlation of Genetic and Antigenic Evolutions

Correlation analysis of genetic and antigenic changes is an essential part of our proposed framework for GAP analysis (Fig. 1). It can provide insight beyond the genetic and antigenic maps. For example, antigenic change may only be correlated with genetic change within a narrow range. Thus, significant correlation may be observed only within this range. The Spearman rank correlation coefficient was used to quantify the correlation level. The higher the correlation, the more accurately cross-reactivity can be predicted by genetic change.

2.5. Identification of CDPs and TDPs

Mutational behavior analysis identifies TDPs by comparing the mutational behavior of each individual alignment positions with that of the whole-length proteins [5]. The idea is that, in these positions, amino acids would be only conserved between proteins genetically similar, and the other way around. The mutational behavior of the proteins can be represented by a matrix whose elements are the protein distances of all protein pairs. On the other hand, the mutational behavior of each individual alignment positions is represented by a matrix whose elements are amino acid dissimilarities of all the residual pairs at that position. These two matrices are compared with a Spearman rank-order correlation coefficient. Positions with a high value of this correlation coefficient are the ones for which dissimilarities between amino acids are correlated with the genetic distances between the corresponding proteins, and hence are predicted as TDPs.

Similarly, CDPs are identified by comparing the amino acid dissimilarity matrix with antigenic distance matrix, whose elements are the antigenic distances of all virus pairs. In this case, antigenic distance is considered the "functional classification" of the proteins [22]. Obviously, CDPs may coincide with TDPs when genetic and antigenic evolutions

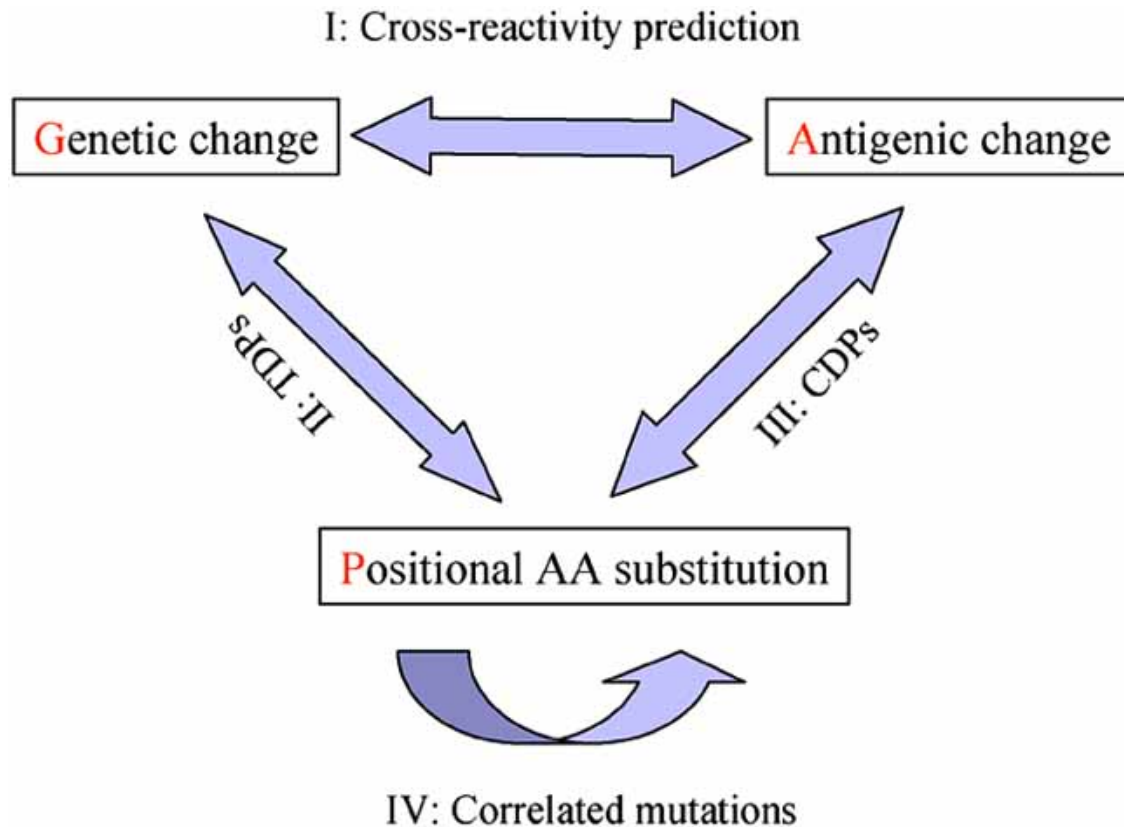


Fig. (1). Schematic representation of mutational behavior approach for GAP analysis. (I) Correlation of genetic and antigenic variation. Correlation level indicates the accuracy of cross-reactivity prediction by genetic change. (II) Mutational behavior correlation to genetic change identifies TDPs. (III) Mutational behavior correlation to antigenic change identifies CDPs. (IV) Correlated mutation between two positions.

are highly correlated. Nevertheless, there may be a disagreement between them, considering the punctuated antigenic evolution of Influenza [26].

2.6. Correlated Mutation and Correlation Map

Correlated mutation between two positions can also be similarly detected by the Spearman rank correlation coefficient of their amino acid dissimilarity matrices. High correlation between a residue pair indicates that either they are in close contact or that they are compensatory mutations [10]. Therefore, correlated mutation analysis is critical in defining the neutralizing antibody-binding surface and in differentiating the escaping mutations (true CDPs) from compensatory mutations. A correlation matrix can be obtained by calculating all pair-wise mutational correlations, followed by MDS analysis to generate a correlation map. Genetic and antigenic changes can also be positioned on this map. We expect that a higher level complex pattern may be revealed from this correlation map.

2.7. Position Space Analysis

Casari *et al.* introduced a method used to predict functional residues in protein families [4]. This method represents each sequence as a vector point in a multi-dimensional space (sequence space), with residue positions and residue types as the basic dimensions. Principal component analysis (PCA) is then applied to determine the specific sequence patterns (profiles), a combination of specific residue types at specific sequence positions. PCA is a statistical method used

to reduce a set of observed variables into a relatively small number of components that account for most of the observed variances. This is accomplished by mathematical linear transformations of the observed variables under two conditions. The first condition is that the first component (the principal component) accounts for the maximum amount of variance possible, the second component accounts for the next maximum amount of variance, and so on. The second condition is that all components are uncorrelated with each other. We adopt this approach in our mutational behavior analysis for similar purposes. PCA is applied to the amino acid dissimilarity matrix rather than the vector representation of residues. The pattern observed from this “position space” will be compared to that from the correlation map.

2.8. Structure Mapping

The crystal structure of haemagglutinin was solved more than 20 years ago [33]. The HA glycoprotein of the influenza virus is a trimer comprising two structurally distinct regions: a triple-stranded coiled-coil of α -helices and a globular region of antiparallel β -sheet, which contains the receptor-binding site. The variable antigenic determinants are positioned on top of this stem. Five antigenic sites were deduced from this structure and sequence data [30, 31, 33]. At least one amino acid substitution in each site seems to have been required for the production of new epidemic strains between 1968 and 1975. A recent complete study of viruses from the 1968-2003 seasons confirmed the importance of mutations in these five antigenic regions [26]. The defined TDPs, CDPs

and other positions of interest will be mapped to this structure using the UCSF Chimera package [24]. Interpretation of mutational behavior patterns in the context of antigen structure further helps us define the cross-neutralization determinants.

2.9. Statistical Analysis

All statistical analysis and plotting were performed using R statistical package [27].

3. RESULTS

3.1. Correlation of Genetic and Antigenic Evolution

The first step in GAP analysis is to compare the genetic and antigenic evolution patterns. In our analysis, comparison of the genetic and antigenic maps showed a remarkable difference (Fig. 2A, B). While the viruses seem to be clustered

together according to seasons on the genetic map, this pattern is not obvious on the antigenic map. This observation is consistent with the analysis of a complete H3N2 virus set isolated from seasons 1968 - 2003 [26]. Not surprisingly, the correlation of genetic and antigenic changes is not very strong (Fig. 2C).

HA1 sequence alignment shows that, among the 329 total positions, only about 50 have variations. The rest of the positions are completely conserved across the 11 isolates (Fig. 3). The variation pattern varies from position to position. Which variable position(s) determine the cross-neutralization pattern is the focus of the following analysis.

3.2. Identified TDPs and CDPs

We first identified TDPs whose mutational behaviors are highly correlated with genetic changes. Among the 50 vari-

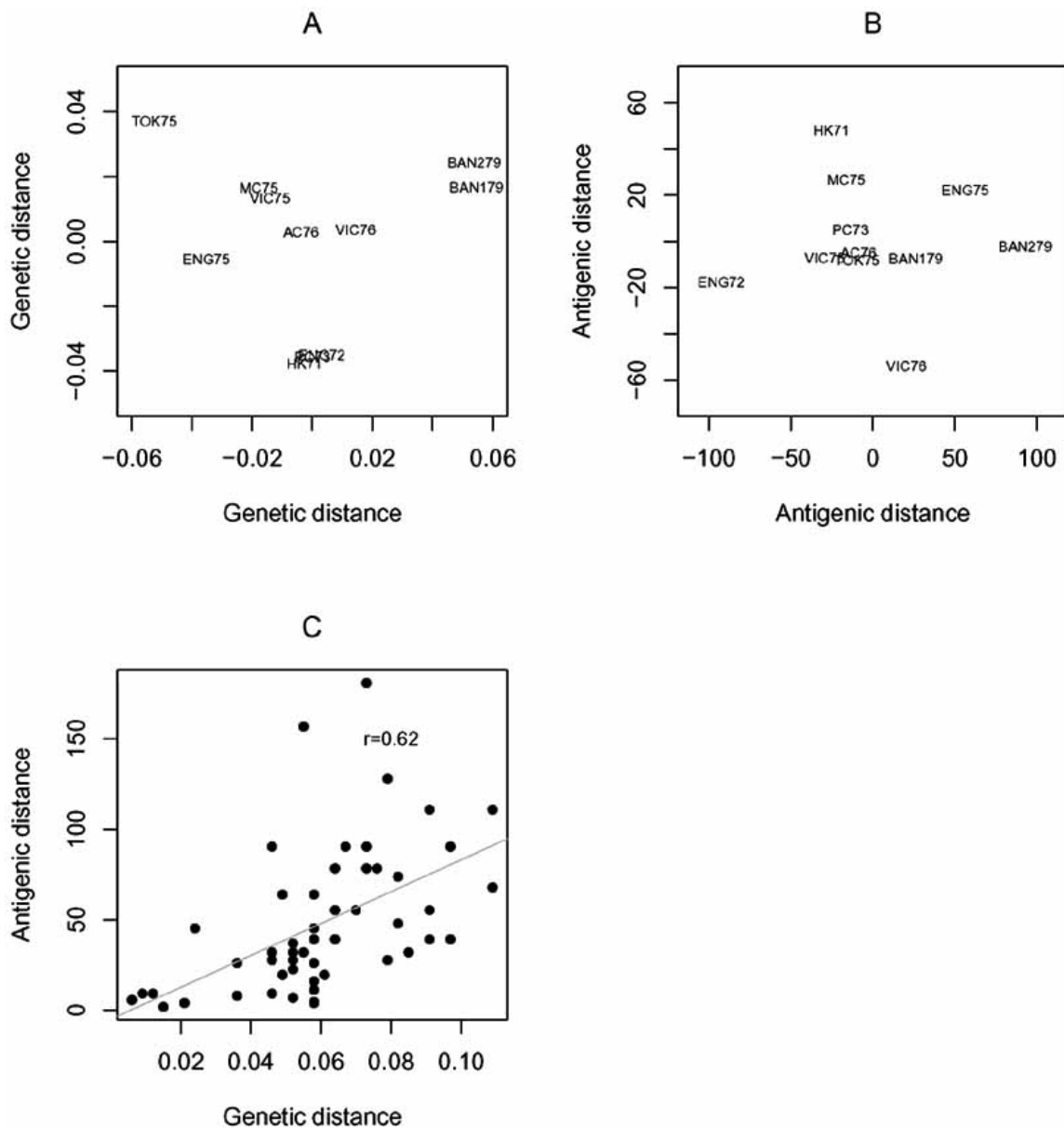


Fig. (2). Genetic and antigenic evolution pattern. (A) Genetic map. (B) Antigenic map. (C) Correlation of genetic and antigenic distances. Each point represents an unique pair of viruses.

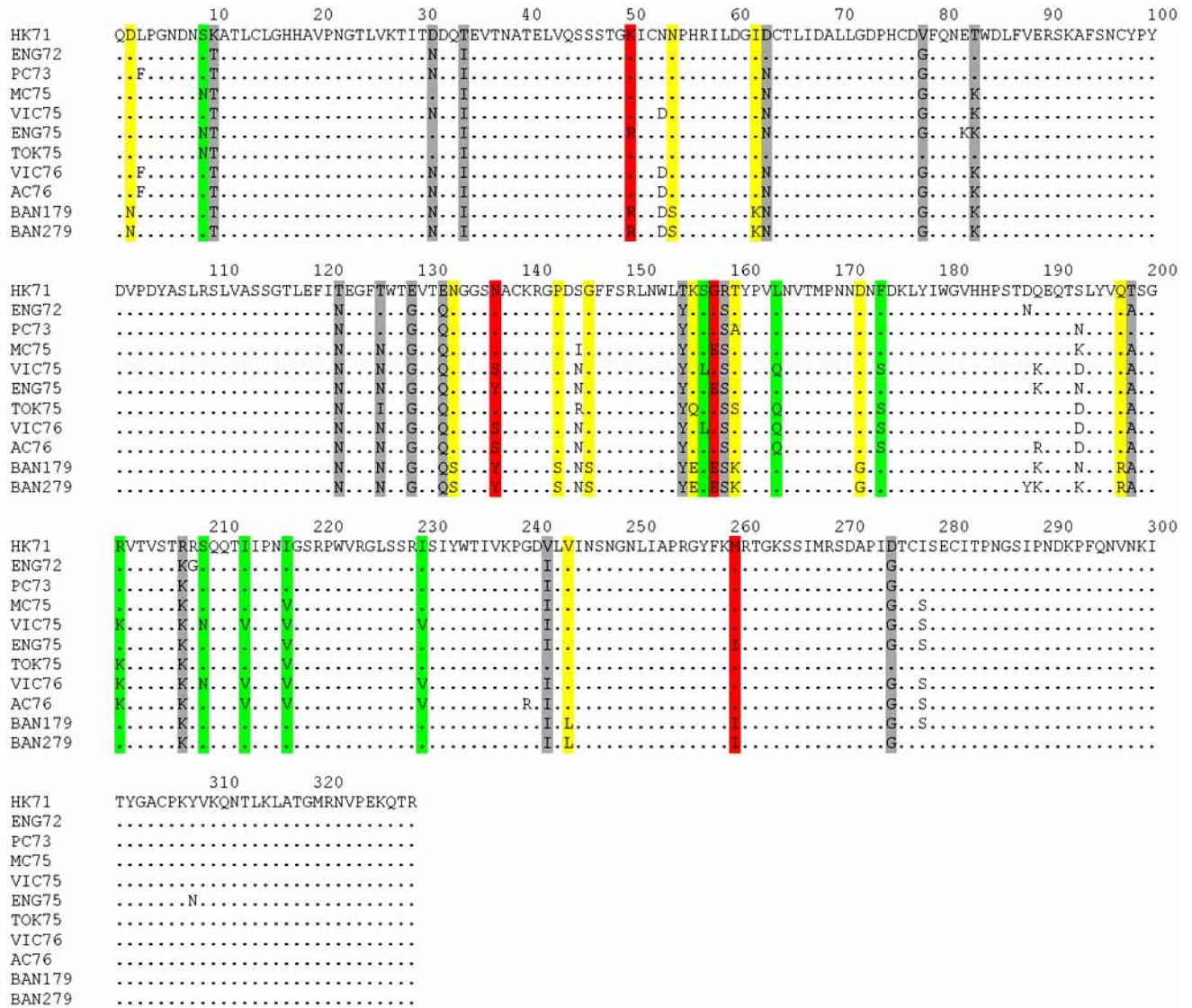


Fig. (3). Amino acid sequence alignment of the 11 HA1 proteins. Red color-shaded positions are the identified CDPs with comparable or higher antigenic correlation. Other CDPs are in yellow. Gray and green color-shaded positions indicate the non-CDPs from the two clusters respectively.

able positions, over a dozen of them showed strong correlation with genetic change and were identified as TDPs (Fig. 4A). Considering the observed genetic and antigenic evolution pattern, we anticipated that most of the TDPs would be CDPs as well. Indeed, antigenic correlation analysis identified the same set of positions as CDPs (Fig. 4B, Table 1). Generally speaking, antigenic correlation of corresponding positions is lower than genetic correlation. However, several positions did show comparable, or even higher, correlation to antigenic changes. This finding may indicate that these positions play a more important role in driving the antigenic evolution (Fig. 5A).

Nearly all of these positions are within the five previously defined antigenic sites (Table 1). Eight of them belong to A and B antigenic sites, the major antigenic determinants. Some are the previously defined positive selected positions (133, 156, 158 and 197). One position (156) is also involved with receptor binding. All of the CDPs have a common sequence feature: mutations fixed in the last season (Fig. 3).

For positions with comparable or even higher antigenic correlation, the mutations were introduced in the middle season(s) (1975-1976) and were fixed in the last season. This pattern usually indicates a strong selective pressure.

3.3. Pattern of Correlated Mutations

Next, we characterized the correlated mutations among the identified CDPs and non-CDPs. Interestingly, all of the defined CDPs were clustered together on the correlation map (Fig. 5B). Furthermore, non-CDPs also formed two distinct clusters. These three clusters (the CDPs and the two non-CDPs) occupy three different directions and point to the center where the genetic variation is located. Nevertheless, MDS analysis positioned antigenic variation away from the center.

The two non-CDP clusters showed distinct sequence patterns: positions from one cluster only mutated in the middle season(s) (1973-1976), and positions from another cluster mutated and fixed in the earlier season (1972-1975) (Fig. 3).

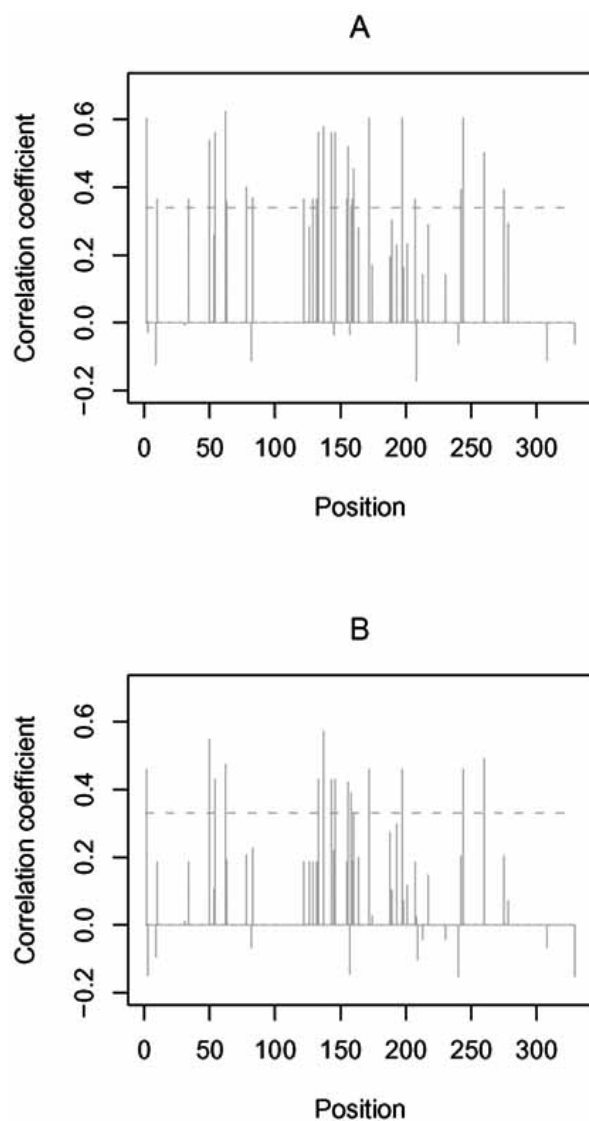


Fig. (4). Positional mutation correlations to genetic and antigenic changes. (A) Correlation level to genetic change of different positions. (B) Correlation level to antigenic change of different positions. Dashed lines indicate the correlation threshold of CDPs and TDPs.

The pattern of the three clusters from the correlation map is reproduced on the plane projected by the first and second principal components (Fig. 5C). Clusters from both methods generally agree. The three clusters point to the center, at which the conserved positions are laid. However, still some subtle differences can be observed. For example, in the CDP cluster, positions with comparable correlations to genetic and antigenic changes tend to be located more towards the conserved center.

3.4. Structure Mapping

Interpretation of CDPs in the structural context can further help us understand the above pattern and refine the cross-neutralization determinants. While over half of the defined CDPs are located on the rim of the receptor-binding pocket (A and B antigenic sites), no positions from non-CDP clusters can be mapped to this region (Fig. 6). On the contrary, other antigenic sites contain both CDPs and non-CDPs.

We also find that positions from one non-CDP cluster seem to cluster in the HA1 interface region. However, the implication of this pattern is not clear.

4. DISCUSSION

Identification of cross-neutralization determinants is an important challenge in the development of vaccines against variable pathogens such as Influenza and HIV-1. In this analysis, we introduced an automatic approach to identify the cross-neutralization determinants, which is based on the fundamental concept of “mutational behavior” to identify the specificity-determining positions in a protein family. The power of this approach relies on the comparisons of positional mutational behavior to both genetic evolution and antigenic variation, as well as the correlated mutations. An initial application of this approach to a small Influenza cross-reactive data set confirmed its usability.

While Influenza evolves genetically at a constant rate, the antigenic evolution pattern seems punctuated [26], making it very difficult to accurately predict cross-reactivity by genetic change. Indeed, prediction accuracy is low by multiple regression, SVM and ANN, only slightly better than that by simple regression using genetic distance as a predictor (data not shown). Multiple factors contribute to this difficulty with antigenic interpretation of genetic data. First, not all mutations will result in antigenic change. While some mutations are true escaping events selected by immune pressure, others may merely be compensatory mutations to maintain function and structure integrity. The remaining may simply be neutral mutations. Second, the particular amino acid substitutions and the location of the substitution are also important. Usually, several mutations are required for antigenic cluster transitions. However, sometimes only one amino acid mutation is sufficient [26]. Finally, interaction of multiple substitutions makes the interpretation even more difficult. We introduced the GAP method to answer these complicated questions. Using this method, strong correlation to antigenic change indicates a position’s importance in cross-reactivity. Neutral mutations can be easily identified by their lack of correlation to either genetic or antigenic changes. Actually, mutational behavior analysis may be more sensitive than synonymous-nonsynonymous ratios in detecting positive selection.

Correlated mutation analysis can reveal the complicated network of intra-molecular residue interaction. The observation that all identified CDPs are clustered together may indicate strong interactions among them. One of the major findings from this analysis is the identification of three correlated mutation clusters. This pattern can be jointly interpreted with structure mapping to differentiate the true escaping mutations from compensatory mutations. Structure mapping of positions from the three clusters indicates that receptor pocket rim region may be the only one determining the cross-reactivity. It suggests that mutations outside this region may simply be compensatory to maintain protein functionality and structure.

Antigenic distance, which is a function of accumulated genetic mutations, increases as more positions mutate. Cross-reactivity will be completely lost when genetic distance reaches some high threshold. In this case, cross-neutral-

Table 1. Identified CDPs by Mutational Behavior Analysis

Position	Antigenic region ^a	R/P ^b	Genetic correlation		Antigenic correlation	
			<i>r</i> -value ^c	<i>p</i> -value	<i>r</i> -value ^c	<i>p</i> -value
2	?		0.60	1.90E-06	0.46	0.0005
50	C		0.54	3.04E-05	0.55	2.14E-05
54	C		0.56	1.20E-05	0.43	0.0012
62	E		0.62	7.20E-07	0.47	0.0003
133	A	P	0.56	1.20E-05	0.43	0.0012
137	A	R	0.58	5.41E-06	0.57	7.40E-06
143	A		0.56	1.20E-06	0.43	0.0012
146	A		0.56	1.20E-05	0.43	0.0012
156	B	P	0.52	6.24E-05	0.42	0.0014
158	B	P	0.34	1.08E+02	0.39	0.0035
160	B		0.45	6.00E-04	0.33	0.0156
172	D		0.60	1.90E-06	0.46	0.0005
197	B	P	0.60	1.90E-06	0.46	0.0005
244	D		0.60	1.90E-06	0.46	0.0005
260	E		0.50	0.0001	0.49	0.0002

^aFive antigenic sites (from A to E) were defined previously [31, 32].

^bR indicates the positions involved with receptor binding, and P indicates positive selected positions [3].

^cSpearman rank correlation coefficient.

ization determinants cannot be identified by mutational behavior analysis to antigenic change. This makes it questionable to apply mutational behavior analysis to HIV-1, whose genetic distance is beyond this threshold. While the Influenza HA1 gene evolves at a rate of 0.8% per year at a population level, the evolution rate of HIV-1 gp120 (the HA1 counterpart of HIV-1) is as high as 1% in a single host [25]. As a consequence, the average distances from the same subtype are as high as 10-15% [14]. Cross-subtype genetic distances are even higher (30%). This extreme genetic distance results in a nearly complete lack of cross-neutralization among isolates and accounts for the failure of an HIV-1 neutralizing antibody based vaccine. This means that vaccine design strategy for Influenza cannot be copied for HIV-1 vaccine development. Therefore, researchers have turned their attention to another strategy: identifying epitopes with broad neutralizing ability [21]. Despite the fact that HIV-1 isolates generally lack cross-reactivity, "sparse" cross-neutralization is sometimes observed in the sera of HIV-1-infected patients. It would be interesting to discover the genetic factors determining this "sparse" cross-neutralization pattern. One of the major findings from this analysis is the interchangeability of CDPs and TDPs. This indicates that, in the absence of cross-neutralization data, CDPs can be identified by mutational behavior correlation to genetic change.

We identified cross-neutralization determinants at the individual residual level. Alternatively, mutational behavior analysis can also be performed at a regional level, which is especially useful in identifying the binding surface [6, 16]. Typical linear B-cell epitopes are about seven amino acids long. A 7-aa sliding window is used by most linear epitope prediction programs to find the potential epitope region [29]. However, most of the B-cell epitopes are conformational, with a binding surface of about 800 Å². Thus, a B-cell epitope region can be defined either as the continuous 7-aa window or the residues within the antibody footprint area (3-D cluster). Similar to the identification of CDPs, cross-neutralization determining regions can be identified by regional behavior analysis. By doing so, protein structure becomes a fully integrated part of the mutational behavior analysis. Our initial results showed that this approach is promising in defining the neutralizing antibody-binding surface (data not shown).

In summary, we have provided an analytical approach for identifying the cross-neutralization determinants for variable pathogens with unstable genomes. The usability of this approach was confirmed by analyzing a Influenza H3N2 cross-reactive dataset. However, due to the limitation of the sample size, the cross-neutralization determinants and the pattern found from this dataset need to be confirmed by using a more complete Influenza dataset. By doing so, we expect

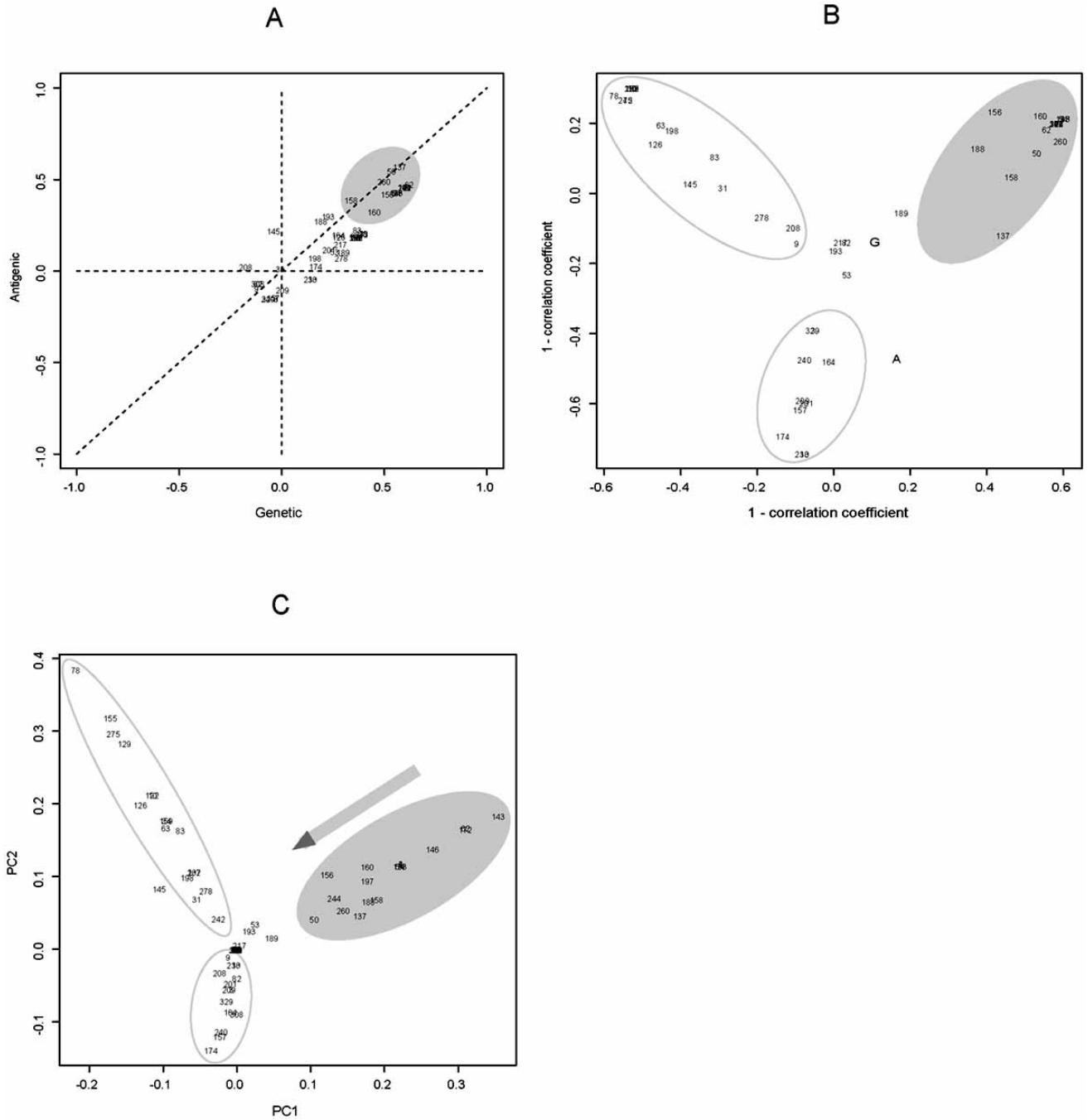


Fig. (5). GAP plots. (A) Scat plot of genetic and antigenic correlation. Shaded area contains the defined CDPs and TDPs. (B) Correlation map. Numbers indicate the positions at the HA1 protein. “G” stands for genetic change and “A” for antigenic change. Gray-shaded region is the cluster of CDPs and TDPs. Two non-CDP clusters are circled. (C) Positions on the plane projected by the first and second principal components (PC1 and PC2). Gray-shaded region is the cluster of CDPs and TDPs. Two non-CDP clusters are circled.

further details to be revealed. We also expect that this method will be an essential tool in the analysis of HIV-1 cross-reactivity data, for which the “sequence gazing” approach is less satisfactory.

ACKNOWLEDGEMENTS

We thank Emily Hemminger for editorial assistance. This work was supported by grants 5U1AI046703-05 from the National Institutes of Health.

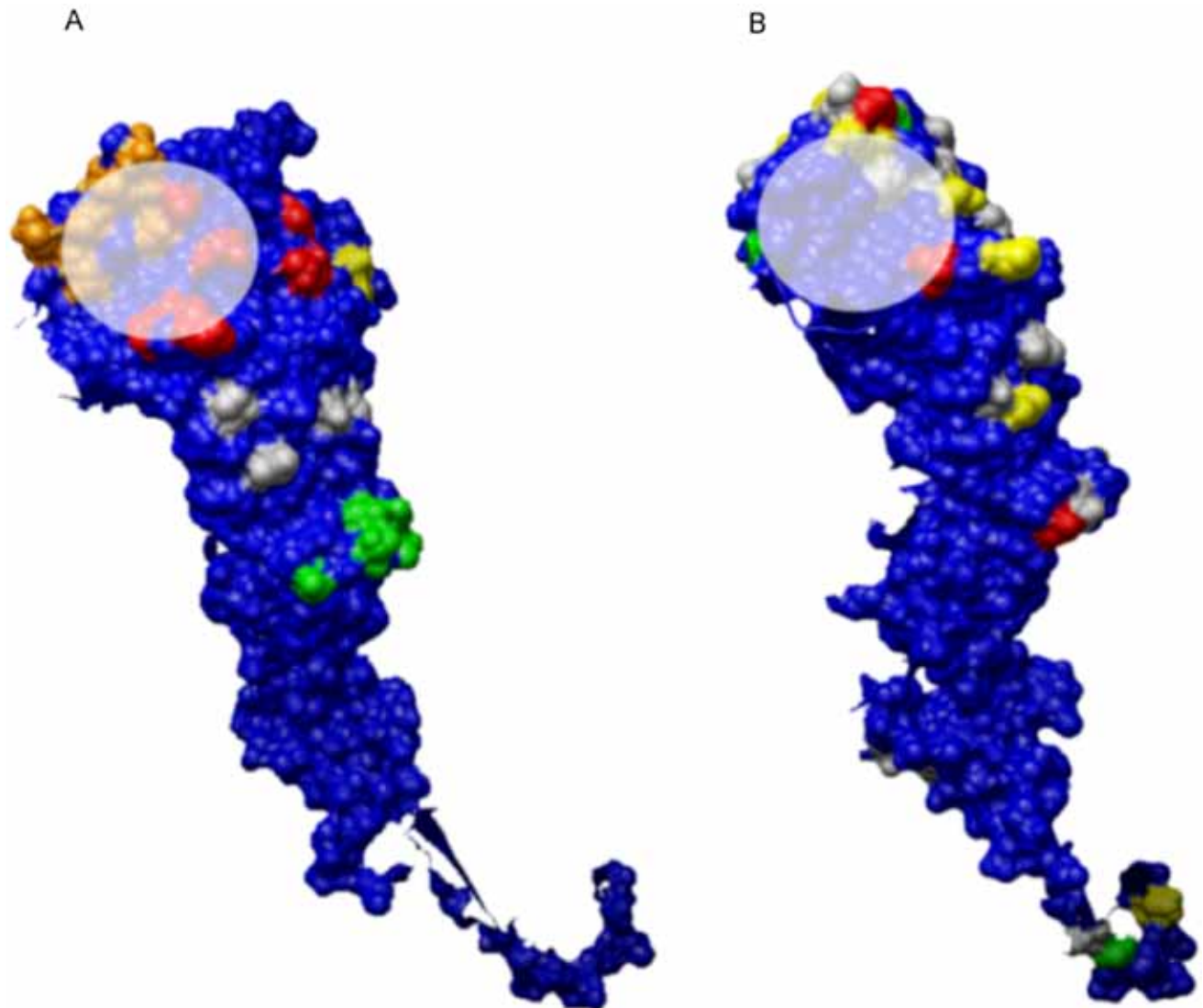


Fig. (6). Structure mapping of CDPs and non-CDPs. Only one HA1 from the trimer is shown. (A) Five antigenic sites on HA1 structure. Site A: red; Site B: orange; Site C: green; Site D: yellow; Site E: gray; (B) CDPs and non-CDPs on HA1 structure. Red residues indicate the defined CDPs with comparable or higher antigenic correlation. Other CDPs are in yellow. Gray and green residues are the non-CDPs from the two clusters. The highlighted region indicates the receptor-binding pocket.

REFERENCES

- [1] Both GW, Sleight MJ, Cox NJ, Kendal AP. (1983). Antigenic drift in influenza virus H3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential amino acid changes at key antigenic sites. *Journal of Virology*. 48: 52-60.
- [2] Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. (1999). Predicting the evolution of human influenza A. *Science*. 286: 1921-1925.
- [3] Bush RM, Fitch WM, Bender CA, Cox NJ. (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular Biology and Evolution*. 16: 1457-1465.
- [4] Casari G, Sander C, Valencia A. (1995). A method to predict functional residues in proteins. *Nature Structure Biology*. 2: 171-178.
- [5] del Sol Mesa A, Pazos F, Valencia A. (2003). Automatic methods for predicting functionally important residues. *Journal of Molecular Biology*. 326: 1289-1302.
- [6] Dopazo J. (1997). A new index to find regions showing an unexpected variability or conservation in sequence alignments. *Computer Applications in the Biosciences*. 13: 313-317.
- [7] Edelstein L, Rosen R. (1978). Enzyme-substrate recognition. *Journal of Theoretical Biology*. 73: 181-204.
- [8] Ferguson NM, Galvani AP, Bush RM. (2003). Ecological and immunological determinants of influenza evolution. *Nature*. 422: 428-433.
- [9] Flynn NM, Forthal DN, Harro CD, Judson FN, Mayer KH, Para MF. (2005). Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *Journal of Infectious Disease*. 191: 654-665.
- [10] Gobel U, Sander C, Schneider R, Valencia A. (1994). Correlated mutations and residue contacts in proteins. *Proteins*. 18: 309-317.
- [11] Henikoff S, Henikoff JG. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Science United States of America*. 89: 10915-10919.
- [12] Kalinina OV, Mironov AA, Gelfand MS, Rakhmaninova AB. (2004). Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Science*. 13: 443-456.
- [13] Kalinina OV, Novichkov PS, Mironov AA, Gelfand MS, Rakhmaninova AB. (2004). SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Research*. 32: W424-428.

- [14] Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science*. 288: 1789-1796.
- [15] Korber BT, Farber RM, Wolpert DH, Lapedes AS. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Science United States of America*. 90: 7176-7180.
- [16] Landgraf R, Xenarios I, Eisenberg D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *Journal of Molecular Biology*. 307: 1487-1502.
- [17] Lapedes A, Farber R. (2001). The geometry of shape space: application to influenza. *Journal of Theoretical Biology*. 212: 57-69.
- [18] Lee MS, Chen JS. (2004). Predicting antigenic variants of influenza A/H3N2 viruses. *Emerging Infectious Disease*. 10: 1385-1390.
- [19] Lichtarge O, Bourne HR, Cohen FE. (1996). An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*. 257: 342-358.
- [20] Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *Journal of Molecular Biology*. 316: 139-154.
- [21] McMichael AJ, Hanke T. (2003). HIV vaccines 1983-2003. *Nature Medicine*. 9: 874-880.
- [22] Pazos F, Rausell A, Valencia A. (2006). Phylogeny-independent detection of functional residues. *Bioinformatics*. 22: 1440-1448.
- [23] Perelson AS, Oster GF. (1979). Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-nonself discrimination. *Journal of Theoretical Biology*. 81: 645-670.
- [24] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*. 25: 1605-1612.
- [25] Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang XL, Mullins JI. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of Virology*. 73: 10489-10502.
- [26] Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA. (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science*. 305: 371-376.
- [27] Team R. Development Core. (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- [28] Vassilatis DK, Hohmann JG, Zeng H, Li F, Ranchalis JE, Mortrud MT, Brown A, Rodriguez SS, Weller JR, Wright AC, Bergmann JE, Gaitanaris GA. (2003). The G protein-coupled receptor repertoires of human and mouse. *Proceedings of the National Academy of Science United States of America*. 100: 4903-4908.
- [29] Welling GW, Weijer WJ, van der Zee R, Welling-Wester S. (1985). Prediction of sequential antigenic regions in proteins. *FEBS Letter*. 188: 215-218.
- [30] Wiley DC, Skehel JJ. (1987). The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annual Review of Biochemistry*. 56: 365-394.
- [31] Wiley DC, Wilson IA, Skehel JJ. (1981). Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*. 289: 373-378.
- [32] Wilson IA, Cox NJ. (1990). Structural basis of immune recognition of influenza virus hemagglutinin. *Annual Review of Immunology*. 8: 737-771.
- [33] Wilson IA, Skehel JJ, Wiley DC. (1981). Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature*. 289: 366-373.