Information Gain based Methodology to Predict the Effect of Conformational Change on the Existence of f10 Epitope on the Surface of Human H5N1 Virus Hemagglutinin Protein

Ahmed Sharaf-eldin
Faculty of computers
Helwan University

Aliaa Youssif
Faculty of computers
Helwan University

Samar Kassim
faculty of medicine
Ainshams University

Doaa Khalil
faculty of computers
Helwan University

ABSTRACT
Bio-informatics tools are widely used to obtain results that are hard to be obtained by physical experiments alone. In this study, we predicted the 3D structure of all human H5N1 hemagglutinin proteins with estimated precision 100%. We tested the existence of the F10 antibody epitope at their surfaces. The information gain (IG) is applied to calculate the degree of association between each position mutation and the absence of F10 antibody epitope on the protein surface. We identified amino acid positions that are responsible for the protein escape from neutralization by f10 antibody.

General Terms
Bioinformatics.

Keywords
Antibodies, H5N1, Hemagglutinin, Influenza virus, Mutation.

1. INTRODUCTION

Influenza A viruses belong to the virus family Ortho-myxoviridae and are the causative pathogens of respiratory infection, leading to seasonal epidemics and devastating pandemics [1]. The virus is divided into different subtypes based on the fourth segment, Hemagglutinin (HA), and the sixth segment, Neuraminidase (NA) [4],[5]. The outbreaks of avian influenza A(H5N1) in South-east Asia, the increasing geographic distribution of this epizootic virus and its ability to cause severe infections (i.e. pneumonia) in humans have led to serious concern about the control measures necessary to curb a potential pandemic [2],[3]. HA forms the spikes at the surface of virions [6],[4]. Furthermore, this protein is responsible for absorption onto the sialoside receptors, which allows the virus to enter the cell by cell membrane fusion. It is generally accepted that antibodies directed against the viral envelope proteins hemagglutinin (HA) and neuraminidase (NA) are the major correlates of protection against infection with the virus. Thus, the induction of sufficiently high titers of specific serum antibodies through vaccination or infection will protect subjects from a subsequent infection. The hemagglutinin (HA) and neuraminidase (NA) of the virus accumulate amino acid changes that can confer resistance to the neutralizing effect of antibodies. This process, called antigenic drift, has been extensively studied. The emergence of drift variants of (H5N1) strains, vaccine efficacy drops as a result of a poor antigenic match with the vaccine strains. Jianhua Sui et al [18] determined the crystal structure of F10 in complex with the H5 (H5-VN04) ectodomain. They determined the amino acids at the HA surface that participate in the interaction between HA protein and F10. They attempted to select neutralization escape mutants and isolated a mutant VN04 virus (K193E) that was resistant to 22F. In contrast, they failed to identify any viruses resistant to any of D8, F10, or A66. While these experiments cannot prove that escape mutants with unimpaired viral fitness will never arise. They examined all of the available HA sequences (total 6360). The sequences of the F10 epitope are nearly always conserved within the H5 subtype. They built their results on examination of the primary HA sequences in the public influenza sequence database. In this study, we aimed to identify H5N1 hemagglutinin amino acids positions that have mutations responsible for the protein escape from the neutralization by F10 antibody.

2. MATERIALS AND METHODS

The proposed method is to predict the structure of all human H5N1 hemagglutinin proteins to test the existence of f10 antibody epitope at their surfaces rather than testing only the primary sequences to get more accurate results and obtain critical amino acid positions that have mutations highly associated to the protein escape from neutralization by F10 antibody.

2.1 Hemagglutinin sequence data

We downloaded all human H5N1 virus hemagglutinin protein sequences, from National Center for Biotechnology Information (NCBI) site [7] and aligned them.

2.2 Hemagglutinin protein structure prediction

The 3D structures of the HA sequences were predicted using PHYRE [8], I-TASSER [20] and ROBETTA[32]servers with estimated precision 100%.

2.3 Detection of f10 antibody conformational epitope on HA proteins

Surendra S Negi and Werner Braun [11] presented a search method, EpiSearch that predicts the possible location of conformational epitopes on the surface of an antigen. The input of EpiSearch is the 3D structure of the antigen and the set of M peptide sequences that simulate the antigen-antibody binding sites residues. The frequency distributions of the residues in these peptide sequences are saved in a matrix. The protein surface of the antigen is divided into surface patches. A surface patch is drawn around each solvent accessible residue. The frequency distribution of the residues in each

patch is saved into another matrix. The property distance PD (A, B) of residues A in the peptide sequences and residues B in each patch is calculated as:

$$PD(A,B) = \sqrt{\sum_i \lambda_i (E_i(A) - E_i(B))^2}$$

where the E_i (i = 1,….,5) are five descriptors of the amino acids physicochemical properties and λ_i are the eigen values of the ith component of E. The number of matching residues in each peptide j and patch k is saved in a matrix X_j^k. for a peptide j, the total number of predicted residues in a patch k is normalized using

$$SIM_j^k = \frac{X_j^k - X_{\min_j}}{X_{\max_j} - X_{\min_j}} \qquad \text{for k= 1,2,…,n}$$

where, X_{min} and X_{max} are the minimum and maximum number of matching residues that exist in all patches for a given peptide sequence j. The final score, $Score^k$, for a patch k is calculated as the average over all M peptides,

$$Score^k = \frac{1}{M}\sum_{j=1}^{M} SIM_j^k$$

The patch that has the highest score is the predicted patch. The method was validated using six test cases where peptide sequences from phage display experiments and the antigen-antibody X-ray structures were available. The method correctly predicted the location of conformational epitopes.
We used the (EpiSearch) server to test the existence of F10 antibody epitope at the surfaces of the HA protein predicted structures. We used as an input peptide sequence, the peptide sequence that was discovered to be in the interaction site between H5N1 HA and F10 antibody by Jianhua Sui et al [18].

2.4 Determination of escape and non escape cases for the f10 antibody epitope

We performed pair-wise comparison between each pair of the downloaded H5N1 HA protein sequences. The pair-wise comparison between two sequences is represented as a string, a position i is denoted as 1 if the residues at this position in the two sequences are not identical (mutation), otherwise the position i is denoted as 0.
To each compared pair of HA protein sequences, we assigned a character j which is denoted as 'e' (abbreviation for escape) if the epitope is predicted to be at the surface of only one of the two compared HA protein sequences predicted structures. The character j is denoted as 'n' (abbreviation for non-escape) if the epitope is predicted to be at the surface of both of them. The character j is denoted as '' (null) if the epitope doesn't exist at the surface of any of the two HA protein sequences predicted structures.

2.5 Identifying critical positions that are responsible for escape from neutralization by f10 antibody on HA

In this study, the effect of one amino acid position on the absence of F10 epitope is defined as the degree of association between mutation on this position and the absence of the epitope. The information gain (IG) [19] is used to calculate this degree of association. The IG is calculated for the epitope Y and each amino acid position X_i separately. An Amino acid with high IG at a specific position implies that a mutation on this position is highly correlated to the absence of F10 epitope. The IG is defined as:

$$I(X_i, Y) = H(Y) - H(Y|X_i)$$

H(Y) evaluates the entropy of the epitope Y and is given as:

$$H(Y) = -\sum_{j=e,n} P(Y = y_j)\log(P(Y = y_i))$$

Two states of y_j are the escape (e) and the non-escape (n) as explained in the previous section. So $p(Y=y_j)$ where j='e' is the probability that the epitope case is escape and is calculated by the division of the epitope escape cases (number of pairs that have associated character j assigned the value 'e') by the number of pairs that contain the epitope at the surface of at least one sequence predicted structure (the pairs that do not contain the epitope at any of the two compared sequences predicted structures surfaces are ignored). $H(Y|X_i)$ evaluates the conditional entropy of Y when given a state of position X_i. Two states of position X_i are the mutated state represented as 1 and the non-mutated state represented as 0. $H(Y|X_i)$ is defined as:

$$H(Y|X_i) = -\sum_{j=1,0} P(X_i = x_j) H(Y|X_i = x_j)$$

$P(X_i = x_j)$ where j=1 is the number of comparisons (pairs of sequences) that have position i is not the same in the two compared sequences (mutation case). And $P(X_i = x_j)$ where j=0 is the number of comparisons (pairs of sequences) that have position i is the same in the two compared sequences (non-mutation case). So for each position X_i we have 6 associated values: the total number of mutation cases, the total number of non-mutation cases, the number of epitope escape cases while the mutation occurs in the position, the number of epitope non-escape cases while the mutation occurs in the position, the number of epitope escape cases while the mutation does not occur in the position, the number of epitope non-escape cases while the mutation does not occur in the position.

2.6 Validation of critical positions

In order to validate our results. We predicted the interaction between H5N1 HA predicted structures and f10 antibody using computational docking. Computational docking of antibody-antigen complexes can now achieve excellent results. Computational docking solves two problems [21]: (1) Finding the correct solution, by altering the relative position of the partners then repeating the calculations; (2) discriminating the accurate solution from the incorrect ones using a scoring function. To search for the binding orientation, the two partners are moved and the score is evaluated. Minimization protocols are used to retrieve the conformation that has the lowest energy. The movement stops when the score does not further improve. There are three docking algorithms: (1) the relative positions of the docking molecules are changed; (2) the relative positions and the side-chain conformations are changed; (3) the backbone conformation is changed, the relative positions are changed and side-chain conformations are also changed. The first algorithm is called rigid body docking and depends on the fact that biological interfaces have complementary shapes. the conformations of the starting structures remains unchanged through the docking process and the scoring function is counted for only the intermolecular interactions [22,23]. ZDock [24] depends on rigid body docking and achieves the best results in the CAPRI experiment [25]. Rosetta-Dock [26] depends on the second algorithm. Since the side-chain conformation has a limited number of allowed torsion angles, the docking job can be completed with great success

and low computational requirements. Rosetta-Dock server performs a local docking search around the starting conformations; the user uploaded coordinate files should reserve a sensible estimate for the starting positions. The protein partners should be placed near contact with the relevant pieces of the proteins facing each other. Rosetta-Dock is highly successful in the blind prediction challenge of the Critical Assessment of CAPRI. HADDOCK [27] depends on the third algorithm which includes backbone conformation changes. To evaluate the performance of the three servers, we used them to predict the interaction between (H5-VN04) and f10 antibody (the same protein sequences used by Jianhua Sui et al).We predicted F10 antibody structure using PIGS (Prediction of Immunoglobulin Structure [28]) server and Rosetta Antibody [29] server.

3. RESULS

3.1 Identification of critical positions responsible for escape from neutralization by f10 antibody

We calculated the information gain for 568 positions in the aligned HA protein sequences. The starting 144 positions have the highest information gain values. Similarly the ending 237 positions have high information gain values (the information gain values for these positions are between 1.0 and 0.81). All proteins that do not have the f10 epitope on their surfaces have deletions in these positions which mean that the deletion of these positions highly induces an escape from neutralization by F10 antibody.

3.2 Evaluation of the performance of the docking servers

We used three servers to predict the interaction between F10 and (H5-VN04): ZDock, Rosetta-Dock and HADDOCK. The antibody starting structure was predicted by the Rosetta-Antibody and PIGS servers and also obtained from the experimentally determined structure of the complex (bound conformation)[pdb entry: 3FKU]. Ten models were produced by Rosetta-Antibody server (R1, R2, R3,……..R10) and one by PIGS. We docked each model of F10 separately with HA to know which one gives the most accurate results in docking. We predicted the structure of the HA protein sequence (H5-VN04) using three servers PHYRE [8], I-TASSER [20] and ROBETTA[32]. This predicted structure was used as a starting structure for docking. ZDock server produced decoys with RMSD to the X-ray structure greater than or equal 20 Å for all starting structures of F10. Rosetta-Dock produced highly accurate results. Complete results are summarized in Table 1. Docking the PIGS and the Rosetta antibody models of F10 using HADDOCK produced decoys with an RMSD equal 7.9 Å or greater. HADDOCK produced decoy with RMSD of 5 Å when docking the experimentally determined conformation as starting structure for antibody [pdb entry: 3FKU]. HADDOCK finds the right position of the heavy chain , but the light chain is moved away from the heavy, probably to reduce inter-chain steric clashes.

3.3 Validation of critical positions

In order to validate the critical positions that are responsible for the HA protein escape from neutralization by F10 antibody, we predicted the interaction between all human H5N1 hemagglutinin protein predicted structures and F10 antibody using Rosetta-Dock server (which is the most accurate). We used the bound structure of F10 antibody [pdb entry: 3FKU] as a starting structure for the antibody. To place the protein partners near contact with the relevant pieces of the proteins facing each other, the two partners (the predicted structure of HA and the bound structure of F10) were aligned with the X-ray complex structure determined by Jianhua Sui et al using the align command in Pymol (http://pymol.org) and separated by 25 Å.

Table 1. BACKBONE RMSD VALUES (IN Å) BETWEEN THE PREDICTED DECOYS AND THE CORRESPONDING X-RAY STRUCTURE FOR F10 COMPLEXED WITH H5N1 HA. EACH ROW REPRESENTS AN ANTIBODY STRUCTURE.

Antibody starting structure	RMSD values
R1	0.5
R2	0.3
R3	0.3
R4	0.2
R5	0.4
R6	0.4
R7	0.7
R8	0.2
R9	0.3
R10	0.6
BIGS	0.1
Bound	0.1

The docking resultant structures were then analysed using HBPLUS3.06 [30] and ligplot4.22 [31] program to identify specific contacts between F10 and HA. The HA proteins that have deletions in the starting 144 positions or the ending 237 positions show weak interactions with F10 antibody which agrees with the results obtained in section A. Fig 1 shows the interaction between (H5-VN04 represented by chain "C') and F10 antibody (represented by chain "Z") as determined by Jianhua Sui et al. there are five strong hydrogen bonds with five amino acids (lys38, His32, Gln34, Asp19, Thr49) and six hydrophobic contacts with six amino acids (Trp21, Thr41, Gly20, His13, Ile45, Ile56). Fig 2 shows the interaction between a human H5N1 hemagglutinin protein (Accession Number: ABW74713 represented by chain A) that has deletions in the ending 237 positions (compared to the sequence used by Jianhua Sui et al) and F10 antibody (represented by chain Z). In the optimized structure, there are only one hydrogen bond with one amino acid (Gln31) and three hydrophobic contacts (weaker interaction). Fig 3 shows the interaction between a human H5N1 hemagglutinin protein (Accession Number: ADM95462 represented by chain A) that has deletions in the starting 144 positions (compared to the sequence used by Jianhua Sui et al) and F10 antibody (represented by chain Z). In the optimized structure, there are only one weak hydrogen bond and three hydrophobic contacts. Other proteins that have fewer deletions (in the ending 219 positions) show similar behaviour (weak interactions) while much fewer deletions (in the ending 78 positions) have no effect on the binding between H5N1 HA and F10.

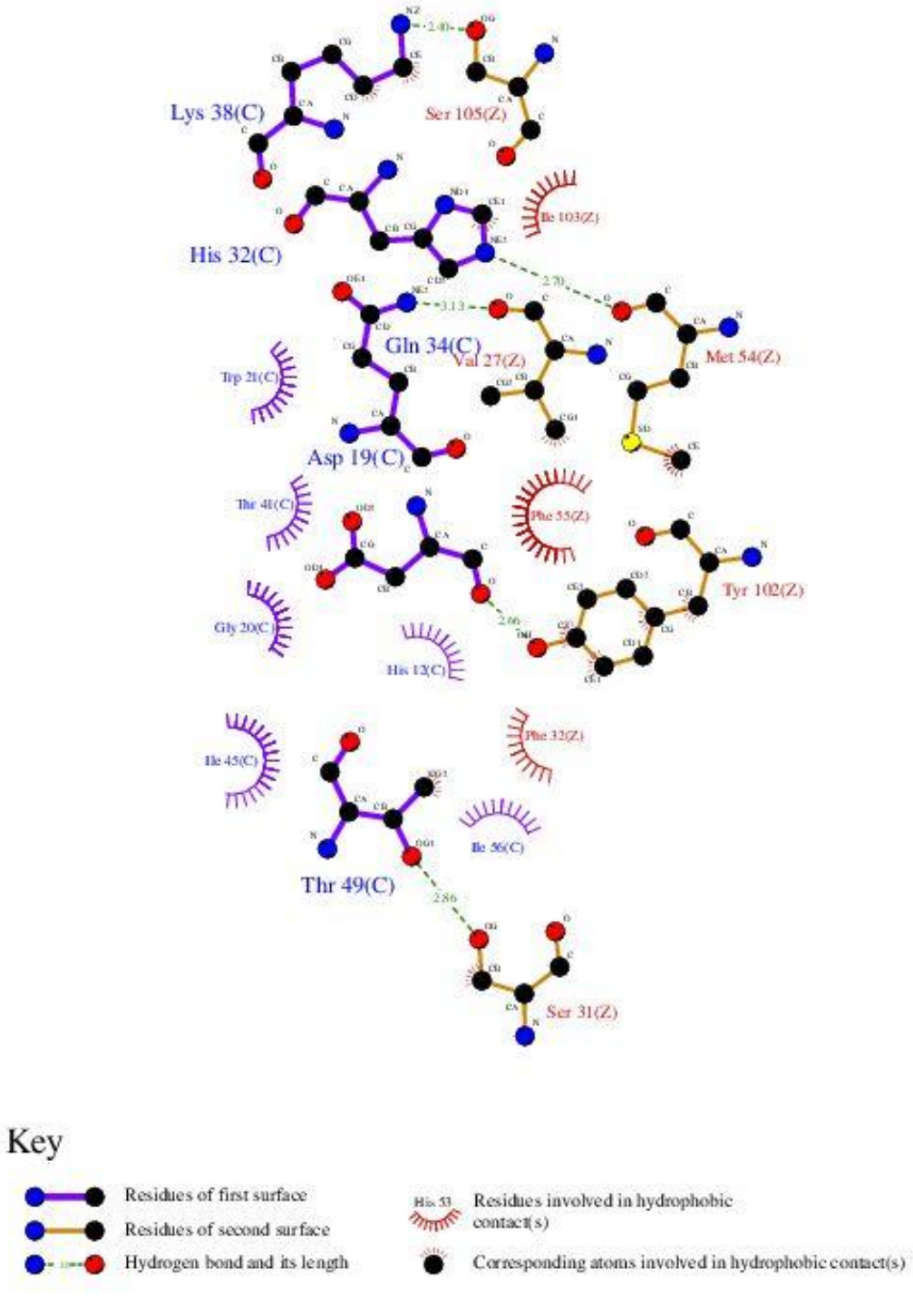

Fig 1: The interaction between (H5-VN04) and F10 antibody as determined by Jianhua Sui et al

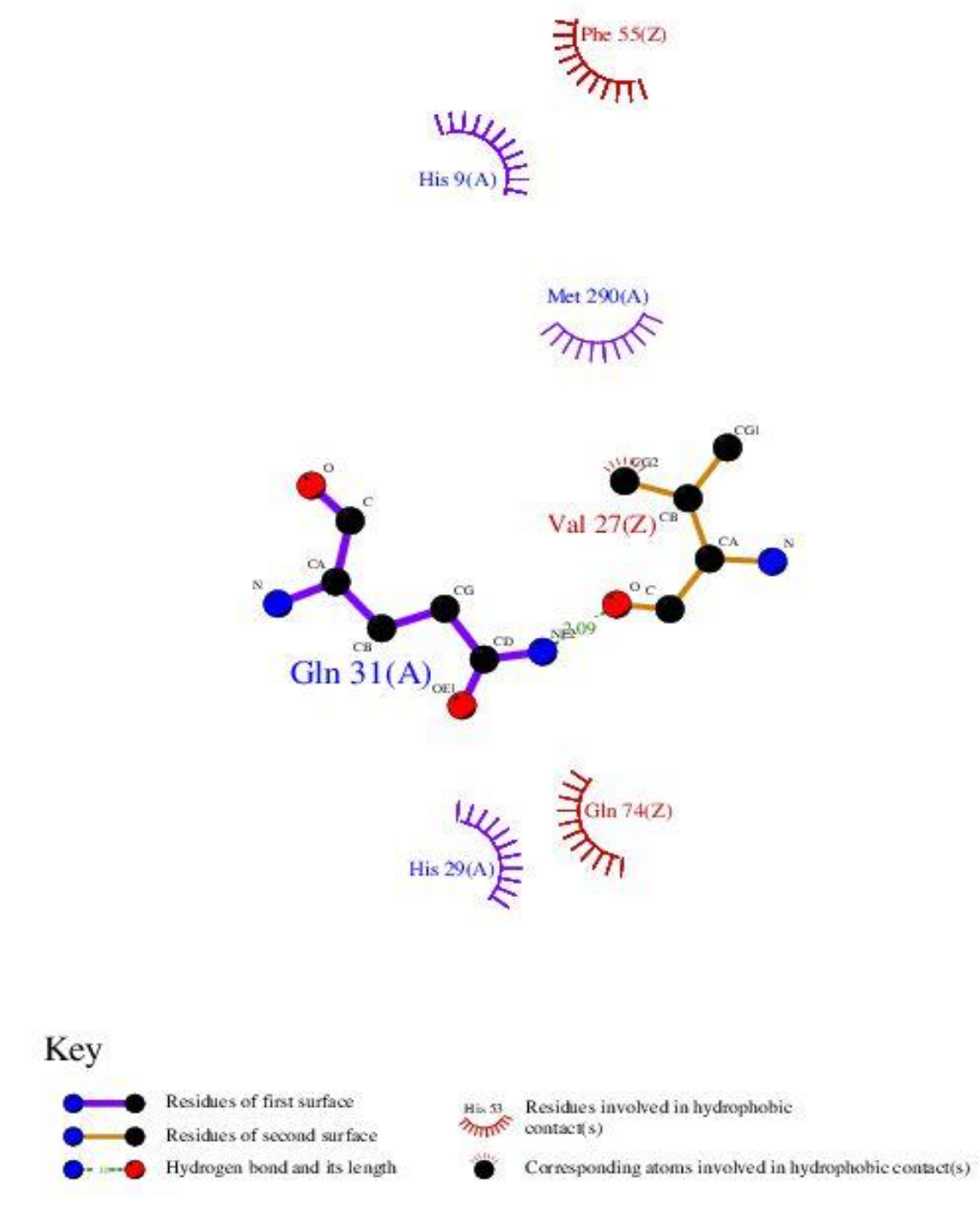

Fig 2: The interaction between H5N1 hemagglutinin protein (Accession Number: ABW74713) and F10 antibody.

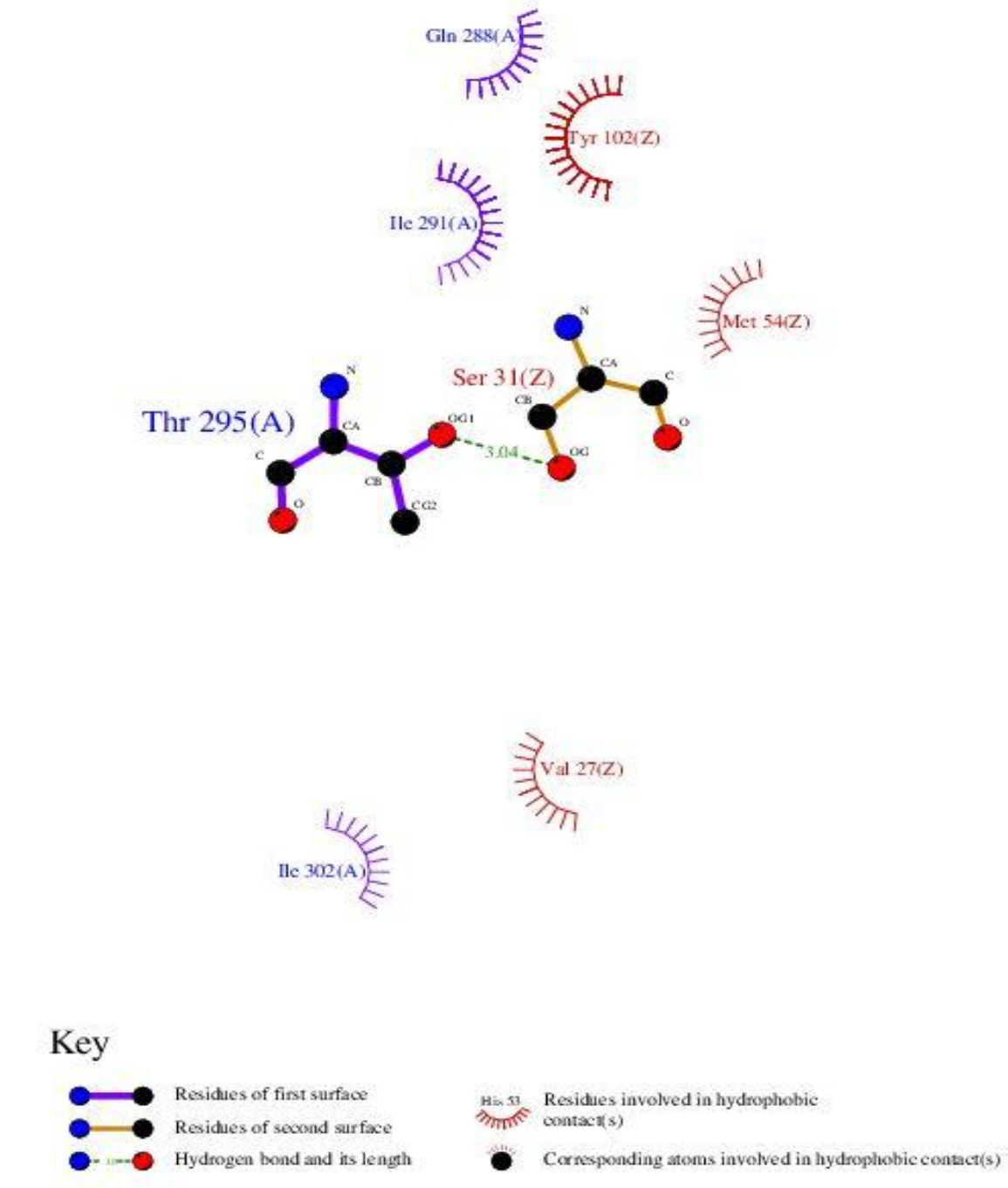

Fig 3: The interaction between H5N1 hemagglutinin protein (Accession Number: ADM95462) and F10 antibody.

4. CONCLUSION

This study demonstrates the feasibility of the information gain for identifying critical amino acid positions of HA that have mutations highly associated to escape from neutralization by F10 antibody in human influenza H5N1 viruses. We used bio-informatics tools like protein structure prediction and epitope mapping to predict the critical amino acid position highly associated to escape.

5. REFERENCES

[1] A. Ghanem, D. Mayer, G. Chase, W. Tegge,R. Frank, G. Kochs, A.Garcia-Sastre, and M.Schwemmle, "Peptide- mediated interference with influenza A virus polymerase", J. Virol., vol. 81, pp.7801–7804, 2007.

[2] E. De Clercq, "Antiviral agents active against influenza A viruses", Nat.Rev. Drug Discov., vol.5, pp.1015–1025, 2006.

[3] J.H. Beigel, J. Farrar, A.M. Han, F.G. Hayden, R.Hyer, M.D. deJong, S. Lochindarat, T.K. Nguyen,T.H. Nguyen, T.H. Tran, A. Nicoll, S. Touch, and K.Y. Yuen, "Avian influenza A (H5N1) infection in humans", N. Engl. J. Med., vol. 353, pp.1374–1385, 2005.

[4] J. B. Plotkin, J. Dushoff, and S. A. Levin, "Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus", Proc Natl Acad Sci U S A., vol.99, pp. 6263-8, 2002.

[5] E. C. Claas, A. D. Osterhaus, R. van Beek, J. C.De Jong, G. F. Rimmelzwaan, D. A. Senne, S. Krauss, K. F. Shortridge, and R. G. Webster, "Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus", Lancet, vol. 351, pp. 472-7, 1998.

[6] "Evolution of H5N1 avian influenza viruses in Asia", Emerg Infect Dis., vol.11, pp.1515-21, 2005.

[7] J. P. Jenuth, "The NCBI. Publicly available tools and resources on the Web", Methods Mol Biol., vol.132, pp. 301-12, 2000.

[8] R.M.Bennett-Lovsey, A.D.Herbert, M.J.E.Sternberg, and L.A.Kelley, "Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre",Proteins, Vol.70, pp.611–625, 2008.

[9] A.G. Murzin, S.E.Brenner, T.Hubbard, and C.Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures", J. Mol. Biol., vol. 247, pp.536–540, 1995.

[10] H.M.Berman et al, "The protein data bank", Nucleic Acids Res., vol.28, pp.235–242, 2000.

[11] Surendra.s.negi and Werner Braun, "Automated Detection of conformational epitopes Using phage Display peptide sequences", Bioinformatics and Biology Insights, vol.3, pp.71–81 , 2009.

[12] R.Fraczkiewicz and W.Braun, "Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules", J.Comp.Chem., vol.19, pp.319,1998.

[13] SS.Negi and W.Braun, "Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces", J.Mol.Model, vol.13, pp.1157–67, 2007.

[14] SS.Negi, CH.Schein, N.Oezguen, TD.Power, and W.Braun, "InterProSurf: a web server for predicting interacting sites on protein surfaces", Bioinformatics, vol.23, pp.3397–9, 2007.

[15] M.Venkatarajan and W.Braun, "New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties", J.Mol.Model, vol.7, pp.445–53, 2001.

[16] VS.Mathura, CH.Schein, W.Braun, "Identifying property based sequence motifs in protein families and superfamilies: application to DNase-1 related endonucleases", Bioinformatics, vol.19, pp.1381–90, 2003.

[17] C.Damian, Ekiert, Gira Bhabha, Marc-André Elsliger, Robert H. E. Friesen, Mandy Jongeneelen, Mark Throsby, Jaap Goudsmit, and Ian A. Wilson, "Antibody recognition of a highly conserved influenza virus epitope: implications for universal prevention and therapy", Science, vol.324, pp.246–251, 2009.

[18] Jianhua Sui et al, "Structural and Functional Bases for Broad-Spectrum Neutralization of Avian and Human Influenza A Viruses", Nat. Struct.Mol.Biol., vol.16, pp.265–273, 2009.

[19] I. H. W. a. E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", Morgan Kaufmann, 2000.

[20] Ambrish Roy, Alper Kucukural, Yang Zhang. "I-TASSER: a unified platform for automated protein structure and function prediction", Nature Protocols, vol.5, pp.725-738, 2010.

[21] Halperin I, Ma B, Wolfson H, Nussinov R, "Principles of docking: An overview of search algorithms and a guide to scoring functions", Proteins, vol.47, pp.409–443,2002.

[22] Camacho C.J, Gatchell D.W, Kimura S.R, Vajda S, "Scoring docked conformations generated by rigid-body protein-protein docking", Proteins, vol.40, pp.525–537,2000.

[23] Cheng T.M, Blundell T.L, Fernandez-Recio J, "pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking", Proteins, vol.68, pp.503–515,2007.

[24] Chen R, Li L, Weng Z, "ZDOCK: an initial-stage protein-docking algorithm". Proteins,vol.52, pp. 80–87,2003.

[25] Gray J.J, Moughon S.E, Kortemme T, Schueler-Furman O, Misura K.M, Morozov A.V, Baker D, "Protein-protein docking predictions for the CAPRI experiment". Proteins, vol.52, pp.118–122, 2003..

[26] Lyskov S, Gray J.J, "The RosettaDock server for local protein-protein docking", Nucleic Acids Res, vol.36, pp. W233–W238, 2008.

[27] Dominguez C, Boelens R, Bonvin A.M, "HADDOCK: a protein-protein docking approach based on biochemical or biophysical information", J. Am. Chem. Soc,vol.125, pp. 1731–1737,2003.

[28] PIGS, Prediction of ImmunoGlobulin Structure. Available online: http://www.biocomputing.it/pigs .

[29] Sivasubramanian A, Sircar A, Chaudhury S, Gray J.J, "Toward high-resolution homology modeling of antibody FV regions and application to antibody-antigen docking", Proteins, vol.74, pp.497–514, 2009.

[30] I.K. McDonald, J.M. Thornton, "Satisfying hydrogen bonding potential in proteins", J. Mol. Biol, vol.238, pp.777–793, 1994.

[31] A.C. Wallace, R.A. Laskowski, J.M. Thornton, "LIGPLOT: a program to generate schematic diagrams of protein-ligand interac-tions", Protein Eng, vol.8 ,pp.127–134, 1995 .

[32] Srivatsan Raman et al, "Structure prediction for CASP8 with all-atom refinement using Rosetta", Proteins, vol.77, pp.89-99, 2009.