ISSN NO. 2320-5407

*Journal homepage: http://www.journalijar.com*
*Journal DOI: 10.21474/IJAR01*

*INTERNATIONAL JOURNAL OF ADVANCED RESEARCH*

**RESEARCH ARTICLE**

# Prediction of the next mutation in Hemagglutinin protein of Influenza-A virus using the variation pattern analysis.

**Simranpal Singh[1], Arun Malik[1], Nishtha Pandey[2], Ravi Kant Pathak[2*].**

1. Department of Computer Science Engineering. Lovely Professional University, Phagwara, Punjab, 144002, India.
2. Department of Biotechnology and Biosciences, Lovely Professional University, Phagwara, Punjab, 144002, India.

## *Manuscript Info*

***Manuscript History:***

Received: 15 March 2016
Final Accepted: 12 May 2016
Published Online: May 2016

***Key words:***
Hemagglutinin, Influenza, MSA, variation, probability score, CD-HIT, clustering.

***Corresponding Author***

**Ravi Kant Pathak.**

## *Abstract*

Every year certain treatment strategies are developed to combat influenza that too only after the epidemic hits the population. The major setback in designing the treatment strategy is due to the variation in the surface antigenic determinants (Hemagglutinin and Neuraminidase) of the virus. In this work, the position specific contribution of an amino acid in the variation of the Hemagglutinin protein has been derived. Multiple sequence alignment of non-redundant sequences of Hemagglutinin from different strains has been used to derive a position specific weighted probability score matrix. The next-in-line variation in the subtype of the Hemagglutinin protein of the influenza A virus has been predicted using the calculated score matrix. Although the prediction has been accomplished with an average accuracy of 60%, the accuracy can still be improved. This strategy may be proven to be useful to design a drug before the outburst of the disease.



## Introduction:-

Influenza A virus has been found to cause the most severe disease in human and have been reported to cause pandemics when crossed the species barrier (Klenk et. al., 2008). Since its first appearance in Spain during 1918-19, it has challenged human population in USA during 1957-58, Hong kong in 1968 (NIAID, NIH, 2011)and over 74 countries in 2009 (WHO, 2013). This virus is empowered with a unique structure and variation in the expression of surface proteins, which imparts a perfect self-defense mechanism. The viral envelop, mainly constituted of 2 types of glycoproteins, Hemagglutinin (HA) and Neuraminidase (NA), protect the core RNA genome which is segmented in nature (Bouvier et. al., 2008). There are 18 different types of HA reported till date (Tong et. al., 2013). HA has been reported to function in recognition of target host cells, and to facilitate the entry of the viral genome into the target cells (Whiteet al.,1997). This makes it the most important and primary target of neutralizing antibodies(Throsby et. al., 2008, Ekiert et. al., 2009, Sui et. al., 2009 and Corti et. al., 2011).

Drift from one strain to another probably depends on point mutation, which might change antigenic determinants, while the region which does not play significant role in triggering the immune response remains highly conserved between different subtypes. These single base substitutions give rise to the diversity in the pathogen (Willy et. al., 1980). Among the various strategies developed till date to solve this problem, tert-butylhydroquinone (TBHQ)(Russell et. al., 2008), neutralizing human antibodies (nABs) (Ekiert et. al., 2009), peptides (Xintian et. al., 2013) and vaccines (Chen et. al., 2011, Anne et. al., 2013) are some. However due to the high level of variability from one season to another season in the strain, these strategies have gained limited success (Ekiert et. al., 2009). The major challenge is to know which strain is going to be prevailing in the coming season, to be prepared with the defense strategy. Current work intends to address this problem and to device a method to predict the next variation in the protein.

## Material and Methods:-

**Data collection:-**

Available protein sequences of all the types of HA were taken from PDB (Berman et. al. 2000) with the keyword Hemagglutinin. Further refinement has been done with experimental method-X-RAY AND taxonomy-VIRUS AND release date between 01-01-2010 up to 31-07-2015. The results were downloaded in the form of fasta files.

**Cluster Analysis and Redundancy Check:-**

Redundancy in the data obtained from PDB has been removed using CD-HIT (Li et. al., 2006). CD-HIT clustered the input protein sequences based on the identity of the characters. The value for identity percentage has been kept as 100 to cluster the duplicate entries. One representative sequence from each cluster has been derived for further analysis. This has been done to reduce the possibility of biasness of the analysis towards any specific type.

**Multiple Sequence Alignment and Block Identification:-**

Multiple Sequence Alignment (MSA) (Sievers et. al., 2011) was carried out on the representative sequences derived from CD-HIT. The results were visualized in Jalview ( Waterhouse et. al., 2009). The consensus sequence observed from MSA has been stored for further reference. An un-gapped block of all the protein sequences has been identified from the MSA.

**Formulation of a Weighted Probability Score Matrix:-**

A position specific 2-D weighted probability score matrix was formulated based on a score value which is obtained by the product of probability of occurrence and weight of each amino acid at that position in the un-gapped block derived from MSA. The method for calculation of this matrix is devised based on the concept of sequence logo (Crooks et. al., 2004). A stack is calculated for each of the positions in the alignment data of proteins. The frequency of occurrence of the amino acids in that position is taken into consideration.

Procedure:

1. In every row I (or the y axis) the number of distinct amino acid, $n_j(\text{distinct}\{aa\})$ and their frequencies, $f(aa_i)$, was counted.
2. In every column J (or the x axis) there is the increasing order of positions of the block.
3. Probability is calculated as the ratio of number of occurrence of a particular amino acid over the total number of amino acids in jth column.

$$P = \frac{f(aa_i)}{n_j(aa)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

P is the probability of occurrence of amino acid

4. Weight is calculated as inverse of the number of distinct amino acids in the jth column

$$W = \frac{1}{n_j(distinct\{aa\})} \quad \ldots\ldots\ldots\ldots(2)$$

W is the weight of each position.

5. In every cell (i,j) following formula is applied:

$$Score = P * W \quad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Based on the above procedure, a 20 X (number of positions in the identified block) 2-D score matrix was calculated, it was then used a base in the prediction algorithm.

**Identification of critical positions for prediction:-**

Global pairwise alignment (Needleman and Wunsch, 1970) using EMBOSS NEEDLE (Rice et. al., 2000) was performed on the sequence for which next prediction is to be made and the consensus sequence that has been obtained during the MSA of all the non-redundant protein sequences of HA. The purpose of performing a pairwise alignment is to identify the significant positions having similarity in terms of function, structure or evolution.

**Statistically Predicted Output:-**

For each of the positions obtained after the pairwise alignment of the input sequence with the consensus sequence, the corresponding amino acid for that position is stored in the database. The values in the score matrix are updated for every position being predicted. Out of the updated values, the amino acid with maximum chance of occurrence is identified based on the calculated score. The amino acid showing a greater chance of occurrence is concluded to the amino acids that will occur next in the chronology.

## Result and Discussion:-

**Data collection:-**

155 protein sequences of HA were taken after refinement as compared to the initial 498 hits obtained through simple keyword search as Hemagglutinin.

**Table 1: Query results**

| **Keyword** | **Hits** |
|---|---|
| **Keyword** {Hemagglutinin} | 498 |
| **Keyword** {Hemagglutinin} AND **Taxonomy** {Virus} | 423 |
| **Keyword** {Hemagglutinin} AND **Taxonomy** {Virus} AND **Experimental Method** {X-Ray} | 412 |
| **Keyword** {Hemagglutinin} AND **Taxonomy** {Virus} AND **Experimental Method** {X-Ray} AND **Time Period** {2010 to 2015} | 155 |

**Cluster analysis and redundancy check:-**

155 refined sequences obtained from PDB became input to the CD-HIT which was operating on the default parameters. The identity cut off had been changed to 100% in order to eliminate the repeated sequences completely. CD-HIT has returned 79 unique clusters as shown in Figure 1. Each cluster contained one or more than one protein sequence in it. From every cluster a representative sequence has been chosen in such a way that it represents the whole cluster. A total of 79 representative sequences were retrieved.

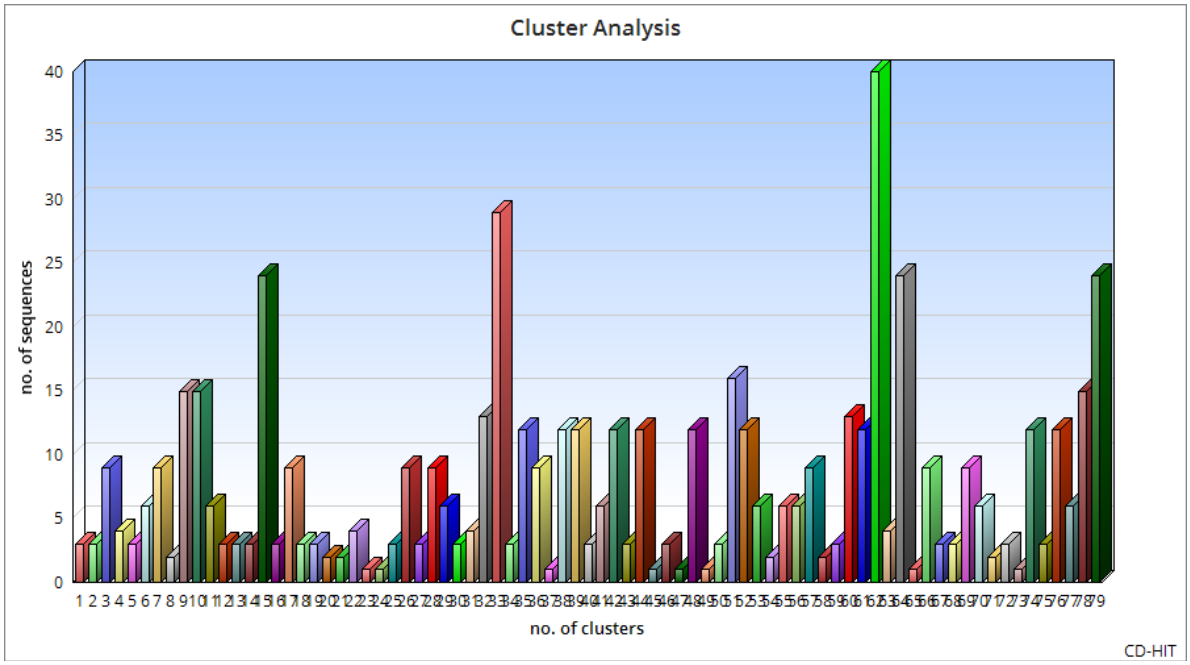


**Figure 1: Cluster Analysis**

**Multiple sequence alignment:-**

MSA of 79 representative sequences was performed. An un-gapped block of 164 positions was observed using Jalview as shown in Figure 2 and Figure 3:

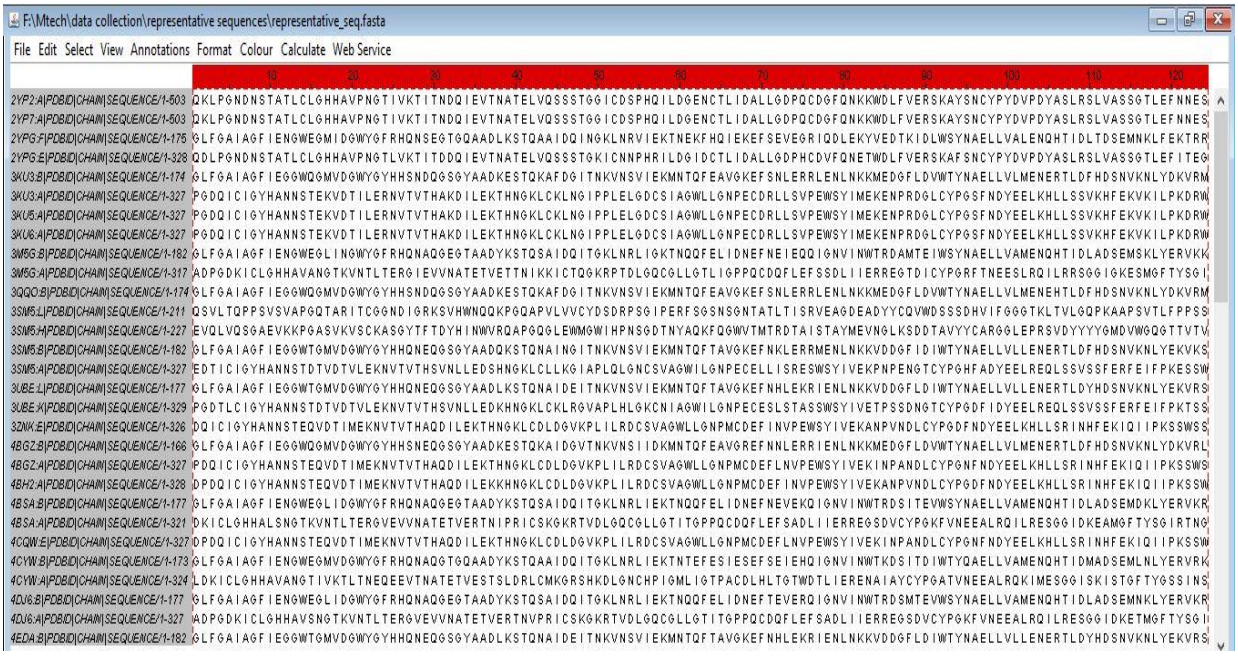


**Figure 2: Ungapped block of positions from 1-124**

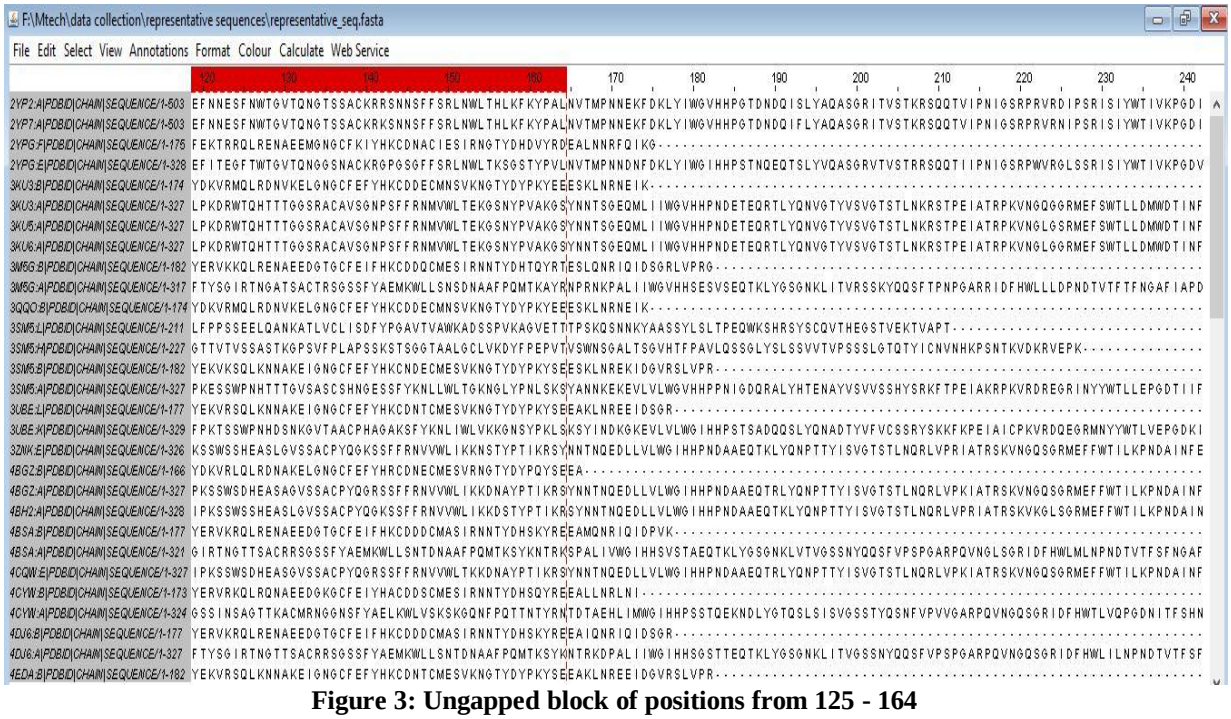


**Figure 3: Ungapped block of positions from 125 - 164**

**The consensus sequence retrieved from MSA was:-**

GDFGAIAGFIENGWEGMVDGWYGYEHQNEQGSGTAADKKSTQGAIDGITGKLNSLIEKTNTQFELIGNEFN
ELEKRIENLIKKVEDGFDDVWSYNAELLVLLENELTLDSHDSEVKKLYEKVRSQLRENAKESGNGCFEFYH
KCDNFCMESIRNGTYDYTKYREEANLNREEIDGLRGIHHPHDEAEQTTLYQNYTTYSSVGTSSTSQRNQPEI
PTRSKVNGVRGRMEFFWTILTILDPIDFESNGNNIAPEAPYKIKKKGSSGIMKSEGSEGNCGTKCQTPTGAIN
SSNPFHNIHPLTIGECPKYVKSKKLVLATGLRNLPNIEKRERRIFGRIAGFIEAGWEEGGDGWYGFGHQNSE
GIGEAADGIATQKAINQIAGKENRLIGKNNEEFHNGEKEFIEGEFRIQDLEINVEDDKIDDWSYNAELLVALE
NQHTEDDRDLDLNDLFERNKHQLIENAEDMGNGCFKIGHKCDNACCGDICNGTYDHDTYRDEALKEEFQI
KRQEIEGIRLVPR

Since the range of positions of the observed un-gapped block is from 1-164, same range of positions have been marked significant in the consensus sequence as well.

**Weighted probability score matrix:-**

For every position of the un-gapped block a weighted probability score has been calculated with respect to every amino acid. The same procedure is followed for each of the 164 positions and the complete 20 X 164 score matrix is obtained. Example of the score matrix for 1st position of the conserved block has been shown in the following table:

**Table 2: Weighted Probability Score for position 1**

| Amino acid | Weighted Probability score at Position 1 | Amino acid | Weighted Probability score at Position 1 |
|---|---|---|---|
| A | 0 | M | 0 |
| C | 0.005 | N | 0 |
| D | 0 | P | 0.005 |
| E | 0 | Q | 0.006 |
| F | 0 | R | 0.001 |
| G | 0.057 | S | 0.001 |
| H | 0.003 | T | 0.001 |
| I | 0.013 | V | 0 |
| K | 0 | W | 0 |
| L | 0.008 | Y | 0 |

**Prediction:-**

Chronologically two latest outbreaks occurred as H7N9 in China, April 2016 and H5N6 again in china in March 2016 (WHO, Disease Outbreak News (DONs), 2016). Since the H5 subtype has been expressed earlier than H7, sequence of H5 has been chosen to be the input to the methodology and the predicted output is expected to be of strain H7.

**Predicted output:-**

Pairwise alignment of input sequence and consensus sequence was carried out using EMBOSS-NEEDLE with default parameters. The aligned positions were then processed through the prediction algorithm and the predicted sequence was obtained.

**Input sequence:-**

H5 (PDB ID-4KWM)

ADPGDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKTHNGKLCDLDGVKPLILRDCSVAGWLLGNPM
CDEFINVPEWSYIVEKANPANDLCYPGNFNDYEELKHLLSRINHFEKIQIIPKSSWSDHEASSGVSSACPYQG
TPSFFRNVVWLIKKNNTYPTIKRSYNNTNQEDLLILWGIHHSNDAAEQTKLYQNPTTYISVGTSTLNQRLVP
KIATRSKVNGQSGRMDFFWTILKPNDAINFESNGNFIAPEYAYKIVKKGDSAIVKSEVEYGNCNTKCQTPIG
AINSSMPFHNIHPLTIGECPKYVKSNKLVLATGLRNSPLRER

**Predicted sequence:-**

ADP**GA**QIC**FI**YHA**W**NSTEQ**GWW**IME**H**NV**E**VTHAQD**A**LEKTHN**GAI**CD**I**DGVKPLILRDCSVAGW**LL**G**N**P
MC**EE**E**I**N**I**PE**LIE**IV**E**KANPA**V**D**S**CYPGNFNDYEE**E**KHLLSRINHF**KK**IQIIPK**S**SWSDHEAS**SG**VSSACPYQ

**K**TPS**F**FR**E**VVWLIKK**D**N**TK**PTIKRSYNNTNQEDLLILWGIHHSNDAAEQTKLYQNPTTYISVGTSTLNQRL
VPKIATRSKVNGQSGRMDFFWTILKPNDAINFESNGNFIAPEYAYKIVKKGDSAIVKSEVEYGNCNTKCQTP
IGAINSSMPFHNIHPLTIGECPKYVKSNKLVLATGLRNSPLRER

A pairwise alignment of the predicted sequence and the expected H7 sequence having PDB ID-3M5G was carried out and a similarity of 51.5% was obtained.

**Accuracy:-**

The method to check the accuracy of the prediction has been chosen such that a global pairwise alignment of the predicted sequence is performed with the actual sequence P' in the phylogeny of the representative sequences. P' represents that sequence in the phylogeny that stands next to the input sequence. The prediction algorithm works with the accuracy as shown in the table below:

**Table 3: Validation results**

| S. No.: | Input Sequence | Next Sequence | Number of predicted positions | Identity percentage | Similarity percentage |
|---|---|---|---|---|---|
| 1 | 4BGZ:A | 4CYW:A | 29 | 33.9 | 52 |
| 2 | 3ZNK:E | 2YP2:A | 25 | 21.3 | 34 |
| 3 | 2YP2:A | 2YP7:A | 33 | 93.6 | 95 |
| 4 | 3UBE:L | 4BGZ:B | 129 | 72.9 | 86.4 |
| 5 | 4N60:D | 4NRJ:F | 115 | 38.5 | 52.2 |
| 6 | 4M40:E | 4NRJ:E | 32 | 81.2 | 86.4 |
| 7 | 4F23:C | 4FIU:C | 33 | 95.1 | 95.7 |
| 8 | 4HKX:A | 4FQR:X | 36 | 24.2 | 41.5 |
| 9 | 3KU3:A | 3KU5:A | 37 | 33.6 | 49.6 |
| 10 | 4LKI:A | 4M4Y:A | 28 | 30.7 | 49.7 |
| | | | **AVERAGE** | **53** | **60** |

Using this method, upon calculation of an average of 10 completely random protein sequences a similarity of 60% and an identity of 53% percent was observed.

## Conclusion:-

79 non redundant representative sequences have been used to perform MSA and based on the position specific weighted probability score, which represents the variation effect in the due course of evolution, a methodology has been designed to predict the next in line subtype. Although the accuracy of the method has been calculated as 60% (based on similarity), scope of improvement still lies open. The accuracy if could be increased further, it can be implemented to other viral diseases also, in which the viral pathogen adopts the same strategy of variation to bypass the immune system without getting identified. This includes HIV-AIDS, SIV and other diseases. Future endeavour would be to increase the accuracy and develop a prediction tool based on the developed methodology. The output of the prediction tool shall be helpful in designing drugs/vaccines which can be effective against any subtype.

## References:-


1. **A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton**, (2009),Jalview Version 2—a multiple sequence alignment editor and analysis workbench, *Bioinformatics*, vol. 25, pp. 1189-1191.
2. **Anne-Cécile V. Bayne, David Boltz, Carole Owen, Yelena Betz, Goncalo Maia, ParastooAzadi, Stephanie Archer-Hartmann, Ross Zirkle, J. Casey Lippmeier**, (2013), Vaccination against Influenza with Recombinant Hemagglutinin Expressed by Schizochytrium sp. Confers Protective Immunity. PLOS ONE, 8(4): e61790, doi:10.1371/journal.pone.0061790
3. **Bouvier NM, Palese P** (September 2008). The biology of influenza viruses.Vaccine. 26 Suppl 4: D49–53. doi:10.1016/j.vaccine.2008.07.039. PMC 3074182.PMID 19230160.
4. **Chen JR, Ma C, Wong CH.,**(2011),Vaccine design of hemagglutinin glycoprotein against influenza. Trends in Biotechnology 29 (9): 426–434doi:10.1016/j.tibtech.2011.04.007
5. **Corti D, Voss J, Gamblin SJ, Codoni G, Macagno A, Jarrossay D, Vachieri SG, Pinna D, Minola A, Vanzetta F, Silacci C, Fernandez-Rodriguez BM, Agatic G, Bianchi S, Giacchetto-Sasselli I, Calder L, Sallusto F, Collins P, Haire LF, Temperton N, Langedijk JP, Skehel JJ, Lanzavecchia A** (August 2011). A

neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. Science 333 (6044): 850–6. doi:10.1126/science.1205669. PMID 21798894

6. **Ekiert DC, Bhabha G, Elsliger MA, Friesen RH, Jongeneelen M, Throsby M, Goudsmit J, Wilson IA** (April 2009). Antibody recognition of a highly conserved influenza virus epitope. Science 324 (5924): 246–51. doi:10.1126/science.1171491. PMC 2758658. PMID 19251591
7. **G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner**, (2004), WebLogo: a sequence logo generator *Genome research,* vol. 14, pp. 1188-1190, .
8. **H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne** (2000) The Protein Data BankNucleic Acids Research, 28: 235-242.
9. **Klenk, Hans-Dieter; Matrosovich, Mikhail; Stech, Jürgen** (2008). Avian Influenza: Molecular Mechanisms of Pathogenesis and Host Range. Animal Viruses: Molecular Biology.Caister Academic Press. ISBN 978-1-904455-22-6.
10. **Needleman, Saul B.; and Wunsch, Christian D.** (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48 (3): 443–53. doi:10.1016/0022-2836(70)90057-4. PMID 5420325
11. **Rice P , Longden I , Bleasby A,**(2000), EMBOSS: the European Molecular Biology Open Software Suite. Trends in genetics : TIG 16 (6) :276-7PMID: 10827456
12. **Russell RJ, Kerry PS, Stevens DJ, Steinhauer DA, Martin SR, Gamblin SJ, Skehel JJ. (2008).** Structure of influenza hemagglutinin in complex with an inhibitor of membrane fusion. Proc Natl AcadSci 105(46):17736-41
13. **Sievers F , Wilm A , Dineen D , Gibson TJ , Karplus K , Li W , Lopez R , McWilliam H , Remmert M , Söding J , Thompson JD , Higgins DG,**(2011), Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology 7 :539 doi: 10.1038/msb.2011.75, PMID: 21988835
14. **Sui J, Hwang WC, Perez S, Wei G, Aird D, Chen LM, Santelli E, Stec B, Cadwell G, Ali M, Wan H, Murakami A, Yammanuru A, Han T, Cox NJ, Bankston LA, Donis RO, Liddington RC, Marasco WA** (March 2009). Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. Nat. Struct. Mol. Biol. 16 (3): 265–73. doi:10.1038/nsmb.1566. PMC 2692245. PMID 19234466
15. **Throsby M, van den Brink E, Jongeneelen M, Poon LL, Alard P, Cornelissen L, Bakker A, Cox F, van Deventer E, Guan Y, Cinatl J, terMeulen J, Lasters I, Carsetti R, Peiris M, de Kruif J, Goudsmit J** (2008). Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. PLoS ONE 3 (12): e3942. doi:10.1371/journal.pone.0003942. PMC 2596486. PMID 19079604
16. **Tong S, Zhu X, Li Y, Shi M, Zhang J, Bourgeois M, Yang H, Chen X, Recuenco S, Gomez J, Chen LM, Johnson A, Tao Y, Dreyfus C, Yu W, McBride R, Carney PJ, Gilbert AT, Chang J, Guo Z, Davis CT, Paulson JC, Stevens J, Rupprecht CE, Holmes EC, Wilson IA, Donis RO** (October 2013). New World Bats Harbor Diverse Influenza A Viruses. .PLoS Pathogens 9 (10): e1003657. doi:10.1371/journal.ppat.1003657. PMC 3794996.PMID 24130481.
17. **W. Li and A. Godzik, (2006),** Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics,* vol. 22, pp. 1658-1659.
18. **White JM, Hoffman LR, Arevalo JH, et al.** (1997). Attachment and entry of influenza virus into host cells. Pivotal roles of hemagglutinin. In Chiu W, Burnett RM, Garcea RL. Structural Biology of Viruses. Oxford University Press. pp. 80–104
19. **Willy Min Jou, Martine Verhoeyen, René Devos, Eric Saman, Rongxiang Fang, Danny Huylebroeck, Walter Fiers** (1980). Complete structure of the hemagglutinin gene from the human influenza A/Victoria/3/75 (H3N2) strain as determined from cloned DNA. Cell 19 (3): 683-96. doi: 10.1016/S0092-8674(80)80045-6. PMID: 6153930