RESEARCH ARTICLE OPEN ACCESS

Development Of Method To Derive Variation Pattern In Neuraminidase Enzyme Of Influenza-A Virus And Predict The Most Probable Upcoming Subtype.

Karishma Agarwal[1], Arun Malik[1], Nishtha Pandey[2], Ravi Kant Pathak[2*]

[1]*(Department of Computer Science, Lovely Professional University, Phagwara, India)*
[2]*(Department of Biotechnology, Lovely Professional University, Phagwara, India)*

ABSTRACT
The influenza A virus has proven to be lethal over the history of time. Every season the virus is usually formed from a new combination of various subtypes of hemagglutinin and neuraminidase. It is impossible to determine in what combination an outburst of the virus will occur and thus presents the challenge of developing efficient, multi-effective drug/vaccine. In this study, the variation pattern followed by the neuraminidase enzyme of the pathogen has been derived using the concept of substitution mutation. The transition score matrix has been calculated to derive the most preferred substitution mutation by an amino acid using multiple sequence alignment and un-gapped block identification. This score matrix has been used to predict the most probable mutations in the present subtype of neuraminidase and propose the next in line subtype. The prediction of the upcoming subtype has been achieved with an average accuracy of more than 60% which can further be improved and the same methodology can be applied to other such highly varying pathogenic viral proteins.

***Keywords* -** Neuraminidase, Influenza A virus, Transition score, CD-HIT, sequence alignment, variation pattern.

I. INTRODUCTION

Influenza has been recognized as one of the deadliest infectious diseases in the recent times. It has affected as large as 40% of the population in some countries. Avian flu and swine flu are some of the examples of the pandemics occurred. The Influenza A virus is responsible for causing the flu pandemics. It can cross species barrier and can affect human as well as animals (Bao et. al., 2008).The seasonal pathogenic strain exhibit different subtypes depending on the proteins that are expressed on the surface of the influenza virus. Neuraminidase (NA) and Hemagglutinin (HA) are the two large glycoprotein molecules that lie on the surface of the influenza virus (Ruigrok et. al., 1998). Envelope glycoprotein NA has an enzymatic activity. It helps the release of newly formed virus particles by cleaving the attachment of the pathogen from the surface of infected cells(Hirst, 1942).Because of its pivotal role in the spread of the infection, NA has been used as a potential target for the antiviral drugs.

Several strategies have been developed till date taking NA as target, however for each infection season the subtype of the NA changes, which makes it difficult to devise a specific vaccine. Hence the vaccine is updated every year (Colacino et. al., 1999). Similarly, the drugs that are used to target NA such as oseltamivir (Tamiflu) and zanamivir (Relenza) (Palese et. al, 1976) have also been proven to be somewhat ineffective due to emerging drug resistance (Russell et. al., 2006).Therefore there has always been a pressing need to engineer new treatment strategy for influenza virus (Barik, 2012). To solve this challenge it becomes very important to understand the pattern of variation (if any) followed by the antigenic protein (NA). In this work, it has been shown that there is an amino acid biasness followed during the transition from one subtype to another posed through substitution mutation. A method has thus been designed to predict the upcoming subtype by looking at the previous outbreak based on a transition score matrix derived through sequence analysis.

II. MATERIAL AND METHODS

2.1 Data Collection

To make a data set, protein sequences of different subtypes of Neuraminidase were collected from the RCSB Protein Data Bank (Berman et. al., 2000). The query made was using the keyword Neuraminidase and was further refined using taxonomy as Influenza A Virus and experimental method as X-Ray and Date of release from 01-01-2010 up to 31-07-2015.

2.2 Redundancy Check

It is critical that the collected data should be accurate, random and non-redundant in order to ensure that biasness of sequences that are in higher number is eliminated. For checking the redundancy of the data a cluster analysis has been performed using the tool CD-HIT (Li and Godzik, 2006) and the repetitions have been eliminated to make sure that the data is accurate and non-redundant. Representative sequence for each cluster has been derived.

2.3 Multiple Sequence Alignment

MSA has been performed with intent to determine an ungapped block of sequences. The alignment of the conserved regions in the input sequences is clearly visualized using the tool Jalview (Waterhouse et. al., 2009). A consensus sequence is also obtained from the multiple sequence alignment of representative protein sequences. The concept here is that if any change (mutation) occurs at a particular position in the consensus sequence then the effects of this mutation can be mapped to all the representative sequences which were used to attain the consensus sequence (Schneider, 2002).

2.4 Threshold Value

In the consensus sequence each position is represented with a value called as Percent Identity. A threshold value of 30% was set because the protein sequences are considered homologous if the percentage identity in the consensus sequence is more than or equal to 30% (Pearson, 2013). Only those positions from the consensus sequences having a percent identity equal to or higher than 30% were selected.

2.5 Phylogenetic Analysis

A phylogenetic tree was calculated by using the representative sequences obtained from CD-HIT as input. The tree was calculated based on the neighbor joining method using BLOSUM 62 distance matrix (Saitou and Nei,1987) Based on the phylogenetic tree derived from the Jalview, an evolutionary path of NA was derived. From the tree, the evolutionary path of the virus in the form of clusters of sequences was obtained. These clusters of sequences are termed as sister sequences (Martin et. al., 2005). Each sister consists of a set of NA sequences. It signifies that the sequences included in particular sister occurred at a same time period in the evolution of the virus. A representative sequence was derived for each sister. This was done by selecting a representative amino acid for each position. The representative amino acid was chosen based on the occurrence of amino acid in all the NA protein sequences of a particular sister. The amino acid with maximum occurrence within the sister at a position was selected as a representative amino acid for that position.

2.6 Mutational Analysis

All the positions in the consensus that satisfied the threshold value of 30% identity were extracted along with the corresponding positions of all the sisters.

Based on the observed statistical data, a 20x20 transition matrix was calculated. In every cell of this transition matrix, a score value is stored which is calculated on the basis of relative pair change frequency. Every score value can be considered as A(i,j) where A is referred as the transition matrix and A(i,j) is the score of transition of a particular amino acid with index ‘i’ to a particular amino acid with index ‘j’. Here, ‘i’ represent the index values for every row of the matrix and similarly ‘j’ for every column of the matrix. Every time such transition is met, the score value is incremented by 1. Hence the transition matrix will consist of transition scores and it will be used while making the prediction.

2.7 Determining the position where prediction is to be made

Pairwise sequence alignment of the input sequence with the consensus sequence is performed using EMBOSS-NEEDLE (Needleman and Wunsch, 1970).Those amino acids in input sequence have been identified which are aligned with the consensus sequence considering them to be the critical positions in terms of structure and function.

2.8 Prediction

Each of these critical positions is filtered based on the threshold PID of 30% and above. Prediction process is then performed on the resulting amino acids. The predicted amino acids are then stored in the same position of the input sequence.

2.9 Transition Matrix Lookup

The process of looking up the transition matrix occurs in the following manner:

1. Result returned by pairwise alignment of consensus and input sequence i.e. the aligned amino acids and their respective positions are stored in the database.
2. For every aligned amino acid: The corresponding i index of the amino acid is identified. The scores at position i in the transition matrix are looked up to find a j index such that A[i,j] has the maximum transition value. The amino acids indexed with j’ is the predicted amino acid for the specific position.
3. The amino acids other than the critical amino acids do not undergo any change.

III. RESULTS AND DISCUSSION

3.1 Collection of data

The search in PDB using the keyword "Neuraminidase" resulted in 338 hits which when refined with organism name as "Influenza A Virus" gave 159 hits. Further refinement with experimental method as "X-Ray" resulted in 159 hits. Final refinement by selecting the Date of Release in the range of 01-01-2010 to 31-07-2015, returned 49 hits.

3.2 Redundancy check

For performing redundancy check using CD-HIT, the value for the parameter "Sequence Identity cut-off" was set to 1 to ensure the complete removal of any redundant sequence. The 49 sequences have been clustered into 20 unique and non-redundant clusters. For each of the 20 clusters, one representative sequence is assigned. In the further processing of the data, the 20 representative sequences are used.

3.3 Multiple sequence alignment

An ungapped block of positions 1 to 369 has been observed after MSA of the 20 representative sequences. It is shown in the figure 1.

3.4 Consensus sequence

After performing multiple sequence alignment on the protein sequences following consensus sequence was obtained:

```
>Consensus/1-466 Percentage Identity Consensus
GSPSNLPKPLCTIPGCSIFGKDNAIRLGSSGDVLVTRE
PYSSCDPDSCDFFACGQGALLRGKHSNGTIKDRTPY
RALISWPLGSPPLLGNSKVECIAVSSSSSHDGKGLGS
ACISGNDNDAAAVIYYGRRALTIIKDSAAIILTTQSSE
CCCICTCCSVVVTDGPAAGSADTRIYIIEGGIIHKKK
EKTSTGIGEEEECSYCYCIVRCCCCRDNNKGNNRPV
RIIDEDANIETGYVCSGIVTDTPRPDDPSTNDKCNNP
NEGGGNGGVGGGGDKGGANTWGGRTISSESSSGY
EIYKVEGAKTKPNSKKLENKQIIVNNDWSGYSGSSG
DYSIESCCCRCCFIEEIGIGGGDVDKEWTSNSIVSFSG
TSNEGGSGGWGDGSNIDGMPLADMDADMALGVM
VSMKEPGWYSFGFEIKDKECDVPCIGIEMVHDGGK
ETWHSAATAIYCLMGSGQLLWDTVTGVDMAL
```

A threshold of 30% was applied on the consensus sequence such that all the amino acids whose score is below than 30% in the consensus sequence are filtered out.

3.5 Phylogenetic analysis

The phylogenetic tree was used to derive various groups/sisters of sequences which signified major chronological mutations. The sequences in each sister signify that those sequences have occurred in same time period during the evolution of NA. A total of 13 sisters were identified with one or multiple sequences as shown in table 1.

Table 1: 13 Sisters and the corresponding sequences that constitute them.

Sister	Sequence PDB_IDs in each sister	Sister	Sequence PDB_IDs in each sister
Sister 1	4CPL:A	Sister 7	4DGR:A
	4CPO:A	Sister 8	4QN3:A
Sister 2	4QN4:A	Sister 9	4H52:A
Sister 3	4K3Y:A		4H53:A
Sister 4	4GDI:A	Sister 10	4MWJ:A
	4GDJ:A		4MWL:A
Sister 5	4MC7:A	Sister 11	4HZV:A
Sister 6	4GZO:A		4HZY:A
	4GZS:A	Sister 12	3SAL:A
Sister 7	4DGR:A	Sister 13	3K36:A
			3K38:A

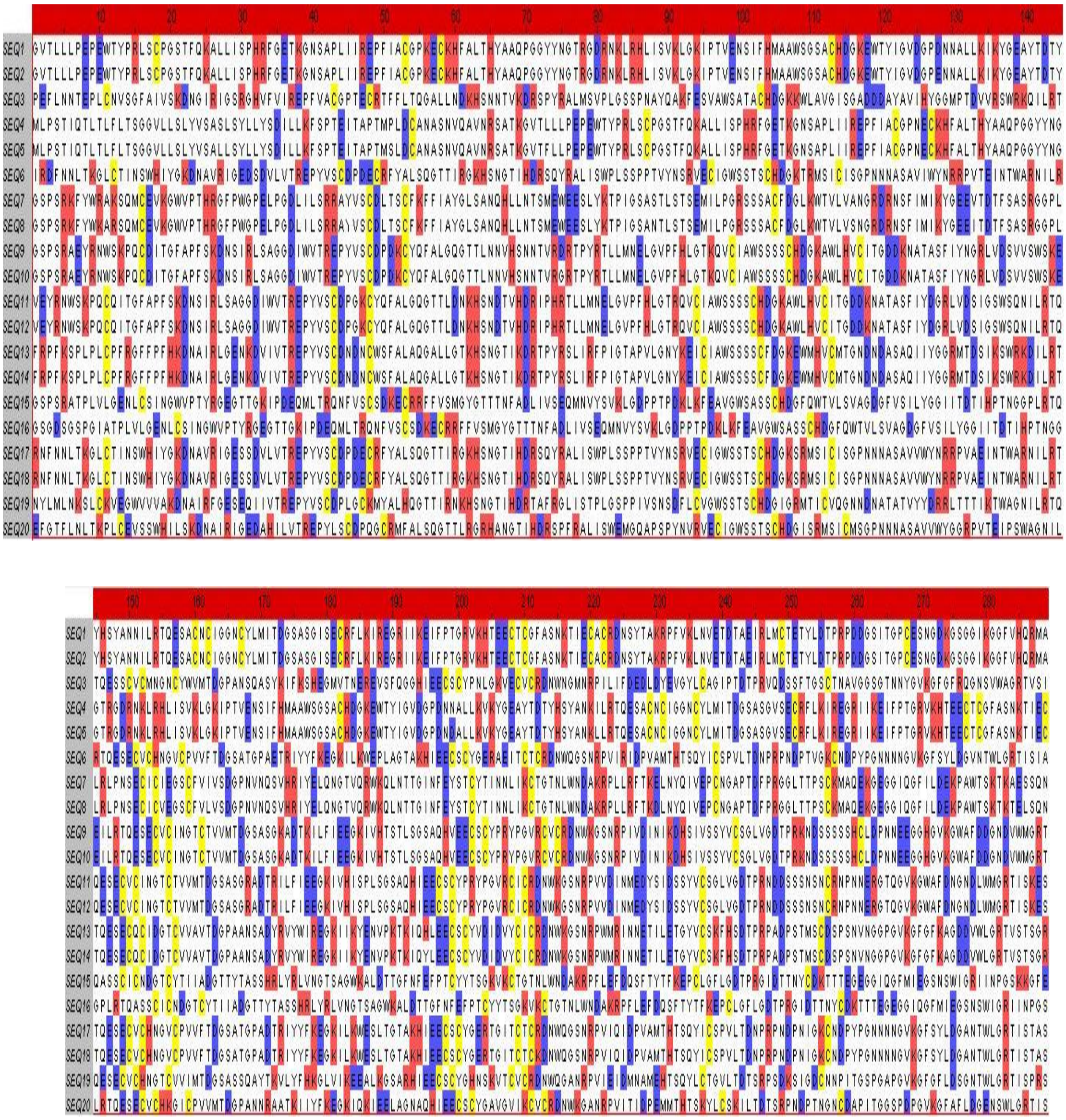

Figure 1: Ungapped block of 20 representative sequences from position 1 to 369 as obtained from

Each value in the transition matrix is calculated on the basis of relative pair exchange frequency. Every time such transition is met, the score value is incremented by 1. The matrix points towards the possible amino acid biasness followed by the virus during variation as shown in the Figure 2.

1.6 Mutational Analysis

Each value in the transition matrix is calculated on the basis of relative pair exchange frequency. Every time such transition is met, the score value is incremented by 1. The matrix points towards the possible amino acid biasness followed by the virus during variation as shown in the Figure 2.

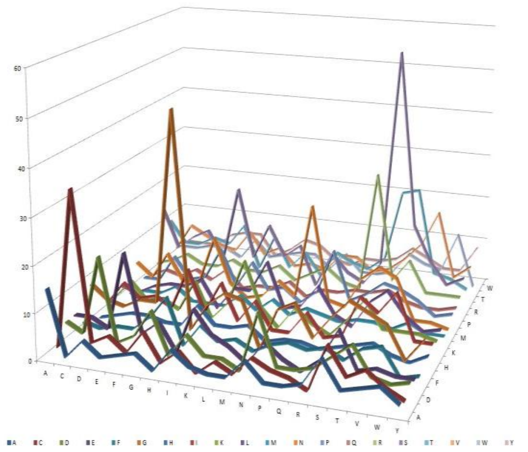

Figure 2: 3-D graph representation of the 20X20 transition matrix representing the transition frequency of one amino acid to another.

3.7 Input

The latest influenza outbreak has been recorded by WHO on 26th April, 2016 in which a human was tested positive for H7N9, a similar case of influenza outbreak has been observed few days earlier by WHO in china on 23rd march,2016 in which human infection with avian influenza H5N6 has been observed (WHO, "Disease Outbreak News (DONs)," 2016). This data has been used to test the validity of prediction algorithm. Therefore N6 with PDB-ID 4QN4 has been selected as the input, to which the prediction sequence must come similar to N9.

Input sequence is:

EFGTFLNLTKPLCEVSSWHILSKDNAIR
IGEDAHILVTREPYLSCDPQGCRMFALSQGTTL
RGRHANGTIHDRSPFRALISWEMGQAPSPYNV
RVECIGWSSTSCHDGISRMSICMSGPNNNASA
VVWYGGRPVTEIPSWAGNILRTQESECVCHKG
ICPVVMTDGPANNRAATKIIYFKEGKIQKIEEL
AGNAQHIEECSCYGAVGVIKCVCRDNWKGAN
RPVITIDPEMMTHTSKYLCSKILTDTSRPNDPT
NGNCDAPITGGSPDPGVKGFAFLDGENSWLGR
TISKDSRSGYEMLKVPNAETDTQSGPISHQVIV
NNQNWSGYSGAFIDYWANKECFNPCFYVELIR
GRPKESSVLWTSNSIVALCGSKERLGSWSWHD
GAEIIYFK

The predicted sequence has been observed as:

EFGTFLNLTKPLCEVSSWHILSKDNAV
RIGEDAHILVSREPSLSCDPQGCRMGALSTGTT
LRGRHANGTIHDRSPFRALISWEMGQAPSPYN
VRVECVGWSSTSCHDGISRMSICMSGPNNNAS
AVVWSGGRPVSEVPSWAGNVLRSTESECVCH
KGICPVVMSDGPANNRAASKIIYFKEGKVQKIE
ELAGNAQHIEECSCSGAVGVIKCVCRDNWKG
ANRPVITVDPEMMTHSSKSLCSKILSDSSRPND
PSNGNCDAPITGGSPDPGVKGFAFLDGENSWL
GRTISKDSRSGSEMLKVPNAETDTQSGPISHQV
IVNNQNWSGSSGAFIDSWANKECFNPCGYVEL
IRGRPKESSVLWTSNSVVALCGSKERLGSWSW
HDGAEIIYFK

3.8 Validation

In order to validate the results obtained from the prediction methodology the phylogenetic tree of the input data set was observed. In the Phylogenetic tree if I was an instance of one input sequence then P was next the observed sequence in the tree. Based on these observations, the Input sequence I was processed using the tool, and obtained a prediction sequence P’. Now in order to determine the similarity between P and P’, pairwise alignment of P and P’ was performed and the similarity percentage was noted.

Using the above mentioned validation method, when the protein sequence of N6 i.e. 4QN4 was processed as input to the prediction algorithm, the predicted protein sequence showed a 62.7% identity and 76.0% similarity with N9 having PDB-ID 4MWJ. Similarly, an average was calculated of 10 random sequences as shown in table 2. From the input data set, an average of 62.01% similarity and 44.36% identity was obtained.

Table 2: Validation result of 10 randomly selected sequences and their similarity and identity percentage with the existing next-in-line subtype as per the chronological arrangement of the sequences

S. No.	Input Sequence PDB_ID	Expected Next Sequence PDB_ID	Number of Positions predicted	Identity percentage	Similarity Percentage
1	4K3Y	4GDI	85	37.9	54.7
2	4DGR	4QN3	161	57.4	72.2
3	4QN3	4H52	138	46.8	68.5
4	4NWJ	4HZV	154	43.9	61.9
5	4HZY	3SAL	152	42.8	60.1
6	4H52	4MWJ	145	45.2	63
7	4GZS	4DGR	141	41.8	60.5
8	4H53	4MWL	144	44.9	63
9	4GDI	4MC7	79	36.3	52.9
10	4QN3	4H53	138	46.6	63.3

IV. CONCLUSION

49 protein sequences of NA were extracted from PDB and clustered into 20 unique and non redundant groups. MSA of the representative sequences from each of the clusters output a 369 positioned ungapped block which act as the basis of the variation analysis. Threshold of 30% has been used to filter the positions which might have evolutionary significance. Amino acid from all the 13 chronologically arranged sister groups at the critical positions were extracted and used to derive the transition matrix. The transition matrix thus obtained directed the focus on the possible amino acid biasness. An average accuracy of more than 60% has been achieved for the prediction algorithm based on the transition matrix. Although the accuracy can still be improved, this method proves to be a step closer to development of new treatment strategies and get prepared for any disease in which the pathogen is highly mutating.

REFERENCES

[1] 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] D. P. Martin, C. Williamson, and D. Posada, (2005), RDP2: recombination detection and analysis from sequence alignments, Bioinformatics, vol. 21, pp. 260-262.

[3] GK Hirst., (1942), Adsorption of influenza haemagglutinins and virus by red blood cells, J Exp Med, 76, 195 – 209

[4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.

[5] J. M. Colacino, K. A. Staschke, and W. G. Laver, (1999), Approaches and strategies for the treatment of influenza virus infections, Antiviral Chemistry and Chemotherapy, vol. 10, pp. 155-185.

[6] N. Saitou and M. Nei, (1987), The neighbor-joining method: a new method for reconstructing phylogenetic trees, Molecular biology and evolution, vol. 4, pp. 406-425, 1987.

[7] P Palese, RW Compans. (1976), Inhibition of influenza virus replication in tissue culture by 2-deoxy-2,3-dehydro-N-trifluoroacetylneuraminic acid (FANA): mechanism of action. J Gen Virol 33,159 - 163

[8] Rupert J. Russell, Lesley F. Haire, David J. Stevens, Patrick J. Collins, Yi Pu Lin, G. Michael Blackburn, Alan J. Hay, Steven J. Gamblin& John J. Skehel, (2006), The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design, Nature 443, 45-49

[9] RWH Ruigrok, KG Nicholson, RG Webster, AJ Hay, (1998), Structure of influenza A, B and C viruses. Textbook of Influenza, Blackwell Science, 29 – 42.

[10] S. Barik, (2012), New treatments for influenza, BMC medicine, vol. 10, p. 104.

[11] T. D. Schneider, (2002), Consensus sequence zen, Applied bioinformatics, vol. 1, p. 111.

[12] 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.

[13] W. R. Pearson, (2013), An introduction to sequence similarity (“homology”) searching, Current protocols in bioinformatics, pp. 3.1. 1-3.1. 8.

[14] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova,(2008), The influenza virus resource at the National Center for Biotechnology Information, Journal of virology, vol. 82, pp. 596-601.