Phylogenetic Analysis Towards Structure Prediction: Influenza a Virus (A/India/m777/2007 (H5N1))

Manish Kumar

[1]Department of Computer Science, Shri Venkateshwara University, Uttar Pradesh, India

Abstract: ***The importance of influenza viruses as worldwide infectious agents is well recognized. The different subtypes of influenza A virus though they all are closely related, they have distinctly different pathogenic behavior which plays an important role in survival in different species. We concluded from phylogenetic analysis that from the common ancestor these strains are diverged more in the course of evolution. So as to adopt a far better survival strategy, this drift is a lot of outstanding. The purpose of modeling is to help the Drug developers and Biotechnologists to develop the drug more efficiently and with more effectiveness in future by analyzing the modeled structures of the protein.***

Keywords: influenza, phylogeny, hemagglutinin, homology modeling.

1. Introduction

Influenza, commonly called "the flu," is an illness caused by viruses that infect the respiratory tract. Members of the Orthomyxoviridae family of RNA viruses cause influenza. Transmission to humans in close contact with poultry or other birds occurs rarely and only with some strains of avian influenza. Potential for transformation of avian influenza into a form that both causes severe disease in humans and spreads easily from person to person is a great concern for world health. There are 16 different HA subtypes and 9 different NA subtypes together forming different combinations. Among them the highly pathogenic are avian H5N1 viruses that caused 18 confirmed infections and six deaths in Hong Kong during 1997 and 2 cases and 1 death in 2003. Thus, the H5N1 avian influenza A virus is a known danger to human across the globe. Surface-exposed or secreted proteins are of primary interest due to their potential as *vaccine candidates, diagnostic agents* and therefore the ease with that they'll be accessible to drugs (Allan and Wren 2003; Mora et al 2003; Flower2002).

Hemagglutinin protein is the receptor-binding and membrane fusion glycoprotein of influenza virus and the target for infectivity-neutralizing antibodies. The entire hemagglutinin protein (HA) from the H5N1 is consists of 568 amino acids, with a mass of 56 kDa. The HA molecule composed of HA1 and HA2 subunits, with the HA1 monetary unit mediating initial contact with the cell membrane and HA2 being responsible for membrane fusion. However, H5N1 viruses have also been found in dead migratory birds, which can recommend a job of untamed birds within the maintenance and unfold of H5N1 virus within the region (Chen et. al 2005).

Recently, several motifs within the three proteins like nucleoprotein, neuramidinase, hemagglutinin, and if influenza virus were identified. Theses motifs were PKC, amidation, kinase 2, tyrosin kinase, glycosylation, ATP/GTP binding site, myristoylation, (Tamanna et al, 2006).

2. Objective

The objective of this paper is to construct tree based on Phylogenetic Analysis of the Influenza A Virus Genomes, obtained from different subtypes in order to predict-Similarity and Relationship between different subtypes, Conservation level of individual gene in individual subtype and which strain is least prevalent for pandemic occurrences. Further, sequence analysis and comparative structure prediction is carried out on the most distantly related/highly mutation susceptible subtype of HA gene (on novel sequence obtained from GenBank) in order to predict-protein family, superfamily etc from the sequence, protein sequence to function prediction and protein secondary & tertiary structure prediction.

3. Methodology

Gene sequences of five different subtypes (i.e. H5N1, H2N2, H1N1, H9N2, and H3N2) of Influenza A virus available in different 8 segments are collected from NCBI- GenBank. Thus we got 40 nucleotide sequences and these sequences were named according to ‘gi’ number for further use. Also hemagglutinin protein sequence of Influenza A virus (A/India/m777/2007(H5N1)) was used for further analysis/predictions.

(i) Phylogenetic Analysis

Multiple sequence alignment using CLUSTALX was done. The obtained aligned sequences were submitted to phylogenetic analysis tools. Three different Tools used are Phylip, FastDNAml and MrByes. Phylip tool with Bootstrap analysis was used to build a tree by neighbor joining (NJ) method. To compliment the result obtained from Phylip, tree based on FastDNAml based on Maximum Liklihood substitution model and MrByes method based on Bayesian analysis were also constructed [Figure 1].

(ii) Sequence Analysis

The HA protein sequence of Influenza A virus (A/India/m777/2007(H5N1)) was selected from GenBank as stated in objective above. Database search for Family, Superfamily and Domain was done using InterProScan at http://www.ebi.ac.uk/Tools/InterProScan/ and BLOCKS database was searched with query sequence for already available Blocks [Table 2, 3].

(iii) Secondary Structure Prediction

Various tools like DSC, GOR4, PHD, PREDATOR and SOPMA at NPSA server, JPRED3, PSIpred were used to predict consensus secondary structure in order to complement each other [Figure 2].

(iv) Homology Modelling

The templates for query sequence were selected using PSI-BLAST search with PDB database search. Further SwissPDB viewer was used to assign structurally conserved regions of template structure to query sequence and then Swiss Model project mode was used to build the complete model [Figure 3, 4]. The model assessment and validation was done using Anolea, Procheck, Verify3D and Ramachandran Plot [Table 4].

4. Results

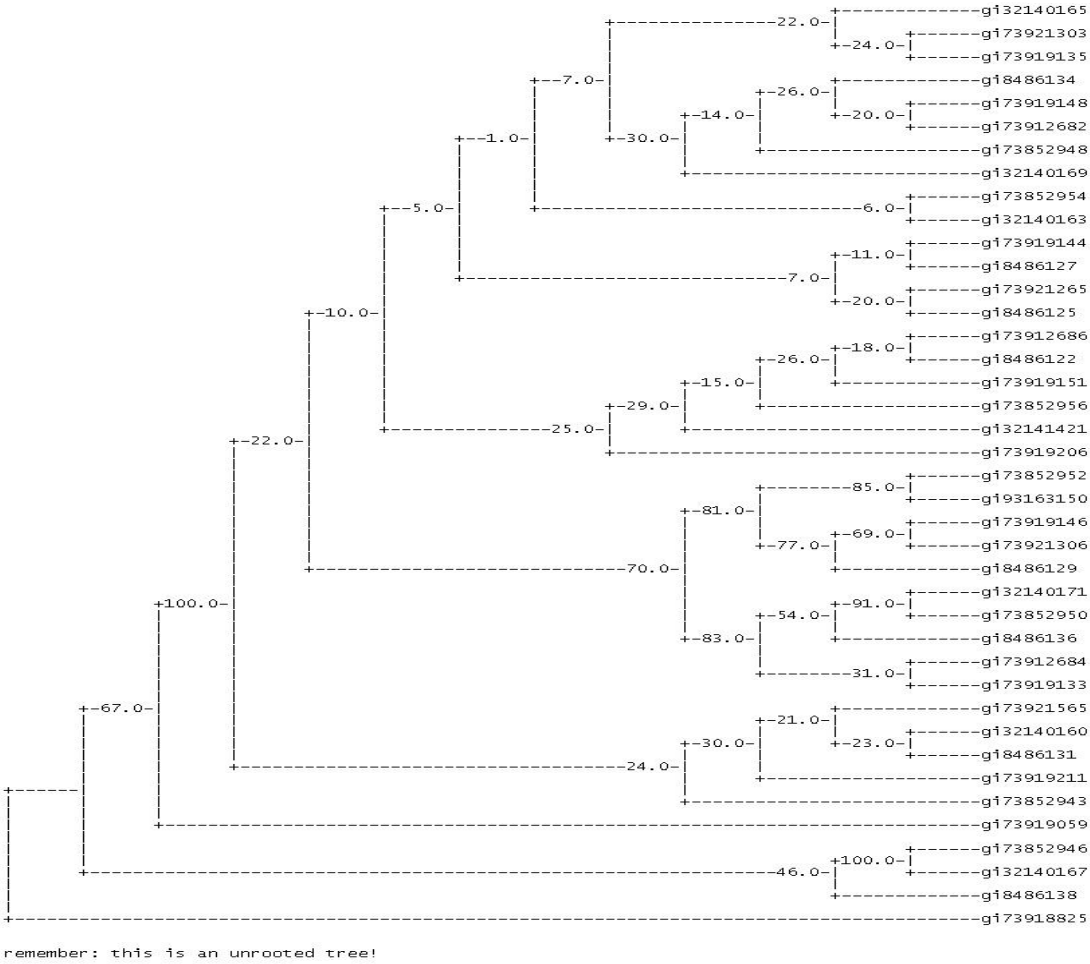


Figure 1: Phylogenetic analysis

Table 1: Analysis of H5N1, H9N2, H3N2, and H1N1

[A]

Gene showing higher branch length in H5N1	
GENE NAME	BRANCH LENGTH (approx.)
HA	3.707

[B]

Gene showing higher branch length in H9N2	
Gene Name	Branch Length (Approx.)
PB2	1.987
PA	2.487
NS2/NS1	2.436

[C]

Gene showing higher branch length in H3N2	
Gene Name	Branch Length (Approx.)
NP	2.358
M1/M2	1.573

[D]

Gene showing higher branch length in H1N1	
GENE NAME	BRANCH LENGTH (approx.)
PB1/PB1-F2	2.188
NA	2.508

5. Sequence Analysis

Table 2: InterProScan

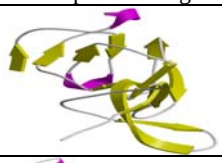
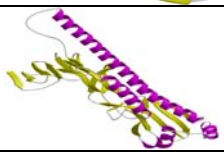
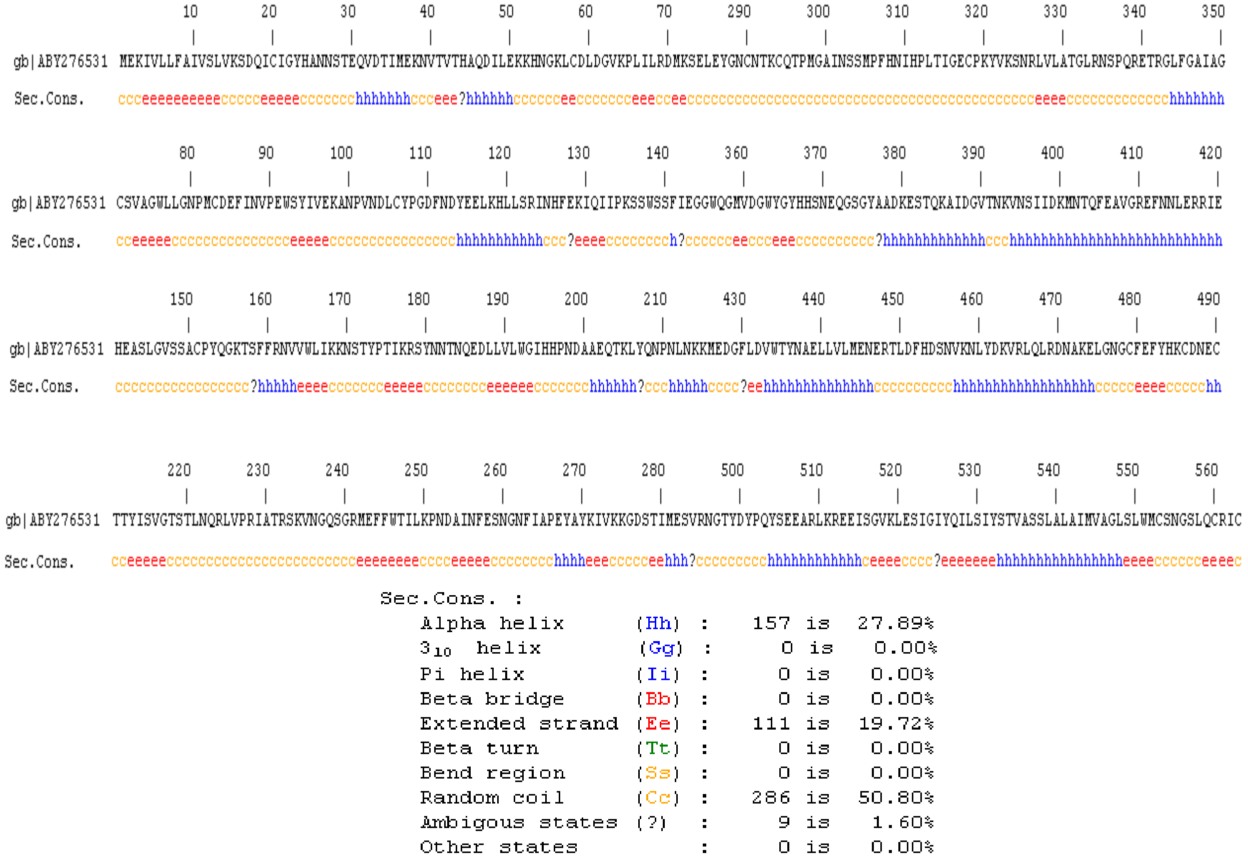
Database name	*Entry type*	*Entry name*	*ID*	*Functional Annotations/Structural Representative*
PRINTS	Domain	HEMAGGLUTININ1	PR00330	Biological Process: Heterophilic cell Adhesion Cellular Component: Viral Envelope Molecular Function: Host cell surface receptor binding
PRINTS	Domain	HEMAGGLUTININ2	PR00331	Biological Process: Viral Infectious Cycle
PRINTS	Family	HEMAGGLUTININ12	PR00329	Biological Process: Viral Infectious Cycle
PFAM	Family	Hemagglutinin	PF00509.7	Biological Process: Viral Fusion with host membrane Cellular Component: Viral Envelope Molecular Function: Host cell surface receptor binding
CATH (GENE3D)	Domain	Hemagglutinin Chain A, Domain 2: Beta-Ribbon Region	G3DSA: 2.10.77.10	
CATH (GENE3D)	Domain	PDB] Virus/Viral Protein: Hemagglutinin Stalk	G3DSA: 3.90.20.10	

Table 3: Blocks

Family	*Strand*	*Blocks*	*E-value*
IPB000149 Haemagglutinin HA1 chain signature	1	8 of 8	5.1e-118
IPB001364 Haemagglutinin HA1/HA2 chain Signature	1	6 of 6	6.1e-100

6. Secondary Structure Prediction



```
                     10        20        30        40        50        60        70       290       300       310       320       330       340       350
                      |         |         |         |         |         |         |         |         |         |         |         |         |         |
gb|ABY276531 MEKIVLLFAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKKHNGKLCDLDGVKPLILRDMKSELEYGNCNTKCQTPMGAINSSMPFHNIHPLTIGECPKYVKSNRLVLATGLRNSPQRETRGLFGAIAG

Sec.Cons.    ccceeeeeeeeeeccccceeeeeccccccchhhhhhhccceee?hhhhhhcccccceecccccccceeecceeccccccccccccccccccccccccccccccccccccccccccceeeecccccccccccccchhhhhhh

                     80        90       100       110       120       130       140       360       370       380       390       400       410       420
                      |         |         |         |         |         |         |         |         |         |         |         |         |         |
gb|ABY276531 CSVAGWLLGNPMCDEFINVPEWSYIVEKANPVNDLCYPGDFNDYEELKHLLSRINHFEKIQIIPKSSWSSFIEGGWQGMVDGWYGYHHSNEQGSGYAADKESTQKAIDGVTNKVNSIIDKMNTQFEAVGREFNNLERRIE

Sec.Cons.    cceeeeecccccccccccccccceeeeecccccccccccccccchhhhhhhhhhhccc?eeeeccccccccch?cccccceeccceeecccccccccc?hhhhhhhhhhhhhccchhhhhhhhhhhhhhhhhhhhhhhhhhh

                    150       160       170       180       190       200       210       430       440       450       460       470       480       490
                      |         |         |         |         |         |         |         |         |         |         |         |         |         |
gb|ABY276531 HEASLGVSSACPYQGKTSFFRNVVWLIKKNSTYPTIKRSYNNTNQEDLLVLWGIHHPNDAAEQTKLYQNPNLNKKMEDGFLDVWTYNAELLVLMENERTLDFHDSNVKNLYDKVRLQLRDNAKELGNGCFEFYHKCDNEC

Sec.Cons.    ccccccccccccccccc?hhhhheeeeccccccceeeeecccccccceeeeeeccccccchhhhhh?ccchhhhhcccc?eehhhhhhhhhhhhhhcccccccccchhhhhhhhhhhhhhhhhhccccceeeeccccchh


                    220       230       240       250       260       270       280       500       510       520       530       540       550       560
                      |         |         |         |         |         |         |         |         |         |         |         |         |         |
gb|ABY276531 TTYISVGTSTLNQRLVPRIATRSKVNGQSGRMEFFWTILKPNDAINFESNGNFIAPEYAYKIVKKGDSTIMESVRNGTYDYPQYSEEARLKREEISGVKLESIGIYQILSIYSTVASSLALAIMVAGLSLWMCSNGSLQCRIC

Sec.Cons.    cceeeeecccccccccccccccccccccccceeeeeeeecccceeeeecccccccchhhheeeccccceehhh?ccccccccchhhhhhhhhhhhceeeecccc?eeeeeeehhhhhhhhhhhhhhhheeeecccccceeeec
```

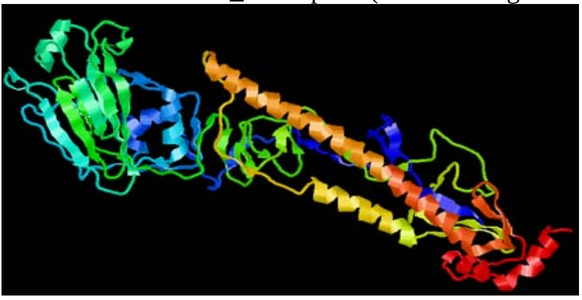
```
Sec.Cons. :
   Alpha helix     (Hh) :   157 is  27.89%
   3_10 helix      (Gg) :     0 is   0.00%
   Pi helix        (Ii) :     0 is   0.00%
   Beta bridge     (Bb) :     0 is   0.00%
   Extended strand (Ee) :   111 is  19.72%
   Beta turn       (Tt) :     0 is   0.00%
   Bend region     (Ss) :     0 is   0.00%
   Random coil     (Cc) :   286 is  50.80%
   Ambigous states (?)  :     9 is   1.60%
   Other states         :     0 is   0.00%
```

Figure 2: Prediction of consensus secondary structure

7. Homology Modeling

7.1 Result of Swiss Model Server

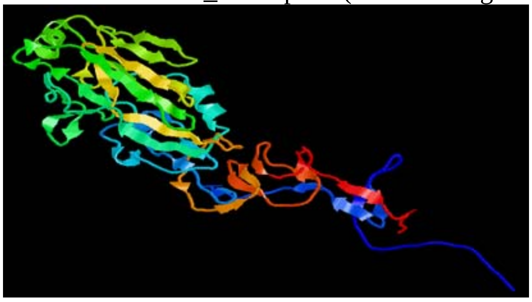
Figure 3: Model 1: 1ha0_a Template (viewed using Rasmol)



Final Total Energy: - 15422.940 KJ/mol

Sequence Identity [%]: 44
Residue Range: 19–515

Figure 4: Model 2: 2ibx_c Template (viewed using Rasmol)



Final Total Energy: - 16669.939 KJ/mol

Sequence Identity [%]: 99
Residue Range: 17-337

Table 4: The Model Assessment and Validation

	ANOLEA Total energy (E/kT units)	Procheck Residues in most favored regions of Ramachandran -plot	Verify3D 3D-1D average score
1ha0_a-model	-1101.254	74.2%	0.68
2ibx_c-model	-3205.581	79.2%	0.72

Note

1. In Anolea, lower the value of energy better is the model.
2. In Procheck, More percentage (>90%) of residues in the most favored region better the model.
3. Low values (<0.3) indicate a problem, whereas high values (>0.5) indicate that the structure is good.

8. Discussion

Our analysis through gene sequences shows that same genes like HA, PA, PB1, PB1-F2; NS, PB2, M1, M2, NA, NP were present in all strains. It reflected that H5N1, H2N2, H9N2, H3N2, H1N1 were evolved from the same common ancestor at the same rate. In case of gene like HA remain more conserved in H9N2, H2N2, H3N2, and H1N1 than in H5N1. [Table 1 (A)] In case of genes like NS2, NS1, PA, PB2 they remain more conserved in H5N1, H3N2, H2N2, and H1N1 than in H9N2. [Table 1 (B)] In contrast, for the genes like NP, M1, and M2 in H3N2 strain appears to diverge more from the common ancestor than H1N1, H2N2, H5N1, H9N2. [Table 1 (C)] In case of genes like NA, PB1 and PB1-F2 are highly conserved in H3N2, H2N2, H9N2, and H5N1 than in H1N1. [Table 1(D)] Therefore, from this observation it had been finished that within the course of evolution, the genes underwent appropriate modifications in strains H1N1, H3N2, H5N1 and H9N2, as compared to H2N2. This proves that H2N2 is a smaller amount pandemic as compared to others that are main causative of pandemic bird flu now a day.

The HA protein of selected strain is predicted to have following functions- Biological Process: Viral Fusion with host membrane, Cellular Component: Viral Envelope and Molecular Function: Host cell surface receptor binding. From the different validation methods for the predicted models, we concluded that both of the models are good enough.

9. Conclusion

After analyzing different subtypes of influenza a virus sequences we come to the conclusion that though they all are closely related, they have distinctly different pathogenic behavior which plays an important role in survival in different species. It is interesting to have closer look at the matter by studying at the gene level. A phylogenetic analysis can be very helpful in understanding the evolutionary pattern. So based on current analysis, it can be said that different subtypes get diverged at different level. So from our current analysis it can be said that from the common ancestor these strains are diverged additional within the course of evolution. So as to adopt a better survival strategy this drift is more distinguished. With the finishing of the ongoing gene sequencing project on Avian Influenza, we hope it will be possible to draw conclusive decision about the true picture of evolution in near future and gene responsible for pathogenesis can also be identified.

We concluded that Hemagglutinin protein that is coded by HA gene is one of the reasons of pathogenicity of Influenza A virus. Till now the structures submitted is using X-ray crystallography or NMR techniques. We forward step to present a theoretical model using available online modelling tools.

10. Acknowledgement

The author is grateful to Vandna Chawla, SRF, Studio of Structural and Computational Biology, IHBT, Palampur, Himachal Pradesh for her help and support to carry out this work.

References

[1] Nitar Nwe, Qigai He, Sudarat, Ivanus Manopo, Damrongwatanapokin, Qingyun Du, Yukol Limlamthong, Beau James Fenner, Lynn Spencer and Jimmy Kwang et al., *Expression of hemagglutinin protein from the avian influenza virus H5N1 in a baculovirus/insect cell system significantly enhanced by suspension culture*, BMC Microbiology, **6**:16 doi: 10.1186/1471-2180-6-16, 2006.
[2] E. Allan, and B. W. Wren, Genes to genetic immunization: identification of bacterial vaccine candidates. *Methods*, 31, 193-198, 2003.
[3] M. Mora, D. Veggi, L. Santini, M. Pizza, and R. Rappuoli et. al, Reverse vaccinology. Drug Discov. Today, 8, 459-464, 2003.
[4] K. Paine, and D. R. Flower, Bacterial bioinformatics: pathogenesis and the genome. J. Mol. Microbiol. Biotechnology, 4, 357-365, 2002.
[5] H. Chen, GJ Smith, SY Zhang, Qink, Wangj, Liks, et al. Avian flu: H5N1 virus outbreak in migratory water fowl. Nature; 436: 191-2, 2005.
[6] A. Tamanna, SK Lal and AU Khan. In silico analysis of genes nucleoprotein, neuraminidase and hemagglutinin: A comparative study on different strains of influenza A

(Bird flu) virus subtype H5N1. In Silico Biology. 6, 0015, 2006.

Author Profile

Manish Kumar is pursuing PhD in Bioinformatics, from Shri Venkateshwara University, Uttar Pradesh. He has also completed M. Sc (Bioinformatics) and B. Sc (Biosciences) from Jamia Millia Islamia University, New Delhi. He has three years of teaching and research experience. He has been earlier associated with Guru Nanak Dev University, Amritsar, in area of Computer Aided Drug Design and Sequence Analysis. He has published number of research papers in national and international journals. He has also attended number of conferences, workshops and refresher course within India. His areas of interest are Computer Aided Drug Design, Sequence Analysis and Computational & Structural Biology.