ISSN 0975-6299 Vol 2/Issue 1/ Jan-Mar 2011

International Journal of Pharma and Bio Sciences

RESEARCH ARTICLE

BIOINFORMATICS

MOLECULAR EVOLUTION OF VIRULENCE GENES OF SWINE INFLUENZA VIRUS SUBTYPE-A H1N1 : AN ANALYSIS OF HOST RADIATION

Corresponding Author

RAJU PODDAR

Department of Biotechnology, Birla Institute of Technology, Mesra, India

Co Authors

BADVE ABHIJIT SADANAND AND KRISHNA KANHAIYA

Department of Biotechnology, Birla Institute of Technology, Mesra, India

ABSTRACT

In this research work, we studied the phenomenon of host radiation in Influenza virus subtype-A and subspecies H1N1 between different host species by building a codon based model. Host radiation is influenced by the rates of mutations in their virulence genes. We have analyzed molecular evolution of virulence genes HA, NA, NS1, NS1, PB1, PB2 of swine influenza virus H1N1 from different hosts swine, avian and humans so as to include all hosts. We used a site specific comparison of non-synonymous (amino acid altering) and synonymous (silent) nucleotide substitutions for the 6 selected genes Maximum likelihood genealogies were used to build two hypothesis models Null (H0) and Alternate (H1) hypothesis based on discrete gamma distribution rates. For null hypothesis, we assumed that the rate of change of a site vary with varying rate of gamma distribution and for alternative hypothesis H1, the rate varies with the fixed gamma distribution under codeml program of PAML. Likelihood ratio test was conducted between null and alternate hypotheses. It was found that null hypothesis had a higher rate of substitution and was found to be more suitable for four of the six selected genes. The study showed that HA gene to be having fastest rate of evolving followed by NS1, PB2 and NA gene. Codon usage studies of genes showing virulence according to our model, the most important codons responsible for molecular evolution are Asp and Glu for HA, Asp and Cys for NA, Glu and Thr for NS1, Lys and Arg for PB2 gene.

KEYWORDS

Avian influenza A virus (H5N1); adaptive molecular evolution; hemagglutinin; NS; PB2; host niche; nucleotide substitution rates; positive selection; Markov model; Likelihood ratio test

INTRODUCTION

Influenza A poses the greatest pandemic disease threat to humankind. The influenza viruses compose of 3 genera in single stranded RNA viral family *Orthomyxoviridae.* Influenza A has two important surface proteins called Hemagglutinin (HA) and Neuraminidase (NA). Influenza A can be classified into several serotypes ranging from H1-H16 and N1-N9. H1, H2 or H3 of influenza A are more common in humans.

Influenza A could potentially infect 30% of the world's population within a matter of months. The continual reinfection occurs because this parasite species are capable of adopting host range and causes the antigenic changes in influenza A. Antigenic drift occurs continually through series of silent mutations by amino acid replacements and major concern in pandemic surveillance.
The recent pandemics of H1N1 in April 2009 originated mainly in North American and Eurasian regions. This 2009 A (H1N1) virus contains a combination of gene segments that previously has not been reported in swine or human influenza viruses worldwide and due to antigenic pseudo shift. In 1918 Spanish flu epidemic originated due to process of resentment between two different viruses. 1918 strain of H1N1 called as "classical swine". There were three major influenza-A pandemics during the 20th century, namely (1) Spanish flu (H1N1) 1918-19, (2) Russian flu (H1N1) 1977, and (3) Flu pandemic (H1N1) 2009, which causes thousands of death.
Evolution of parasite species involves the phenomenon of host radiation and it's allows a parasite to expand its ecological niche by adapting to one or more novel hosts. Since, host radiation incorporates three major steps namely transmission of genetic material to other hosts, replication within the new hosts and finally the transmission between the individuals of new hosts. The new human hosts were immunologically deficient for this strain of the virus as its genetic content is differing from the seasonal virus as a result, lead to fatal deaths and heavy economy loss. So our work is based on the functional constraint and effort to build a codon model of nucleotide substitutions which illustrate the selective forces on the various virulent genes in different hosts including humans, birds and swine. Studies of the genetic basis revealed that out of ten genes HA, NA, PB1 (polymerase1), NS1 (non-structural1), NS2 (non-structural2) & finally PB2 (polymerase2) are responsible for virulence. Another study showed that through likelihood ratio test (LRT) null model and alternate model to detect possibility of protein under positive selection. To do so, we use a site-specific comparison of synonymous (silent) and non-synonymous (amino acid altering) mutations is performed in parasite populations from different hosts. Methods for performing the analyses on a site-specific level have focused on amino acid conservation as an indication of protein function. The purpose of the work is to gain a better understanding of the evolutionary processes in H1N1 avian influenza virus that has undergone host radiation from birds and swine to humans.
Forseberg et al.[1] has examined an efficient method can be developed to identify candidate sites for the molecular biological investigation of species-specific adaptation in parasites. Knudsen et al[2] has shown a maximum-likelihood method for detecting evolutionary rate shifts at specific protein positions and develop models for host radiation in H1N1 and study the adaptive evolution at amino acids level. Goldman et al[3] developed a codon based model for the evolution of protein-coding DNA

sequences is presented for use in phylogenetic estimation. Yang et al[4] has formulated recent statistical methods for detecting molecular adaptation, and discuss their limitations and possible improvements. Cristina et al[5] has shown Bayesian coalescent Markov chain Monte Carlo (MCMC) analysis of full-length neuraminidase (NA) gene sequences of 62 H1N1 IAV strains. Yang et al[6,7,8,9] has shown sequence length, sequence divergence and the strength of positive selection affects the power of the LRT. They have also proved that in small data sets the new BEB method does not generate false positives as did the old NEB approach, while in large data sets it retains the good power of the NEB approach for inferring positively selected sites. The simple models can be as efficient as complex models, and that use of complex models does not necessarily give more reliable trees compared with simple models proved by Piontkivska et al[10]. Shen et al[11] has shown that interplay between antigenic drift and receptor binding in HA evolution, and provide molecular signatures for monitoring future antigenically drifted 2009 pandemic and seasonal A (H1N1) influenza viruses. Zimmer and Valliant group al[12,13] has worked on summary of implications of past epidemics and the history and possible causes of the current epidemic of H1N1. Rocky K[14] has analyzed adaptive molecular evolution of virulence genes of avian influenza – A virus subtype H5N1. Sinha and Adeigo et al[15, 16, 17] has revealed the fact that strong purifying selection is responsible for the evolution of the novel influenza A (H1N1) virus among human.

Table-1
History of pandemics of Influenza due to Antigenic Shift

No.	Year and common name	Causative strain	Confirmed Deaths
1.	1889(Asiatic Flu)	H2N2	1 million
2.	1900(Russian Flu)	H3N8	1 million
3.	**1918**(Spanish Flu)	**H1N1**	50 million
4.	1957(Asian Flu)	H2N2	1.5 to 2 million
5.	1968(Hong Kong Flu)	H3N2	1 million
6.	**1977 (**Russian Flu Pandemic)	**H1N1**	Data not available
7.	2003(South East Asian Bird Flu)	H5N1	486
8.	**2009(**Flu Pandemic)	**H1N1**	17000

Theoretical model:

We are assuming, there are n codons (sites) in the sequence. Let the data at site s (s = 1, 2 . . . n) be represented by two vectors **b***s* and **h***s*, where **b***s* is a vector of codons from the sequences in the original host species, which in our case are birds, at site *s*, and **h***s* is a vector of codons from the sequences in the new host species, which in our case are humans, at site *s.* The codon based model given by Goldman and Yang (1994)[7,18] is a 61 x 61 matrix (stop codons not allowed) of the relative instantaneous substitution rates.

The transition probability matrix of codon substitution over a branch in the tree of length *t* is then calculated as $P(t) = e^{QT}$. The selective regime may differ at different positions in the gene, and potentially in different hosts. To reflect this difference arisen because of heterogeneous selection pressure, we allow a statistical distribution $p(\boldsymbol{\omega})$ of **ω** ratios among sites [8,19]. The probability of observing the data in a site is then obtained by integration,

$$P(data) = \int p(\omega)\, P(data \mid \omega)\, d\omega$$

Considering constant selection pressure after host radiation, we assume that the selective regime is constant in the viruses from the original host. The simplest event that may occur at the time of host radiation is the event *x*, that the selective regime in this codon position (and therefore) remains the same. Conditional on event *x* we have that the total probability of the data in site *s* is given by,

$$P(bs, hs \mid x, M) = 61\Sigma j=1\ \pi j \int P(cj, bs \mid M, \omega) P(cj, hs \mid M, \omega) p(\omega) d\omega$$

Here, *cj* is the codon state in the node of divergence. *M* represents the model parameters, and the tree, including topology and branch lengths. The probability of observing codon *cj* given the subtrees from the two different species are found by traversing the subtrees according to Felsenstein's pruning algorithm [20, 21, 22]. We can now construct the simplest hypothesis H_1 that no changes occur after host radiation. The probability of the data under this hypothesis is simply

$$Ps(H_1 \mid M) = P(bs, hs \mid x, M)$$

and under the assumption that sites evolve independently, the full likelihood of the data is

$$LH1 = \Pi s\ Ps(H_1 \mid M)$$

The probability of observing **r** (divergence in the selective regime affecting a site after host radiation), **b***s* and **h***s* is obtained by independently integrating over the '**ω**' distribution in the two parts of the tree and multiplying.

$$P(bs, hs \mid r, M) = 61\Sigma j=1\ \pi j \int P(rj, bs \mid M, \omega)) p(\omega) d\omega \times \int P(rj, hs \mid M, \omega)) p(\omega) d\omega .$$

However, it is not always necessary that the selective regime on all sites in a protein will change after host radiation. Therefore a hypothesis *H*0 is constructed that allows for host specific selection for the event 'r' with probability *pr*. We then have that the total probability of observing the data under *H*0 is

$$Ps(H0 \mid M) = pr\ P(bs, hs \mid r, M) + (1 - pd)\ P(bs, hs \mid x, M)$$

here '*pr'* is a parameter that describes the propensity of a site to experience a change in the selective regime. This equation is similar to the above equation

$$L_{H1} = \Pi s\ Ps(H_1 \mid M)$$

Next, from the chi square test consider that the simpler (null) model has p0 parameters and the more general (alternative) model has p1 parameters, and the (optimal) log likelihood values under the two models are *l*0 and *l*1. Then twice the log likelihood difference, $2\ l = 2(l1 - l0)$, has asymptotically a χ2 distribution with d.f. = p1 – p0 if the null model is true. So the test statistic $2\ l$ can be compared with that χ2 distribution to test whether the null model is rejected against the alternative model. Derek Gathereret al[22] has shown summary of implications of past epidemics and the history and possible causes of the current epidemic of H1N1.

So, our work is based on the model proposed by Goldman and Yang (1994)[9, 23, 24,25,26] which is a codon based model for the evolution of protein-coding DNA sequences. In this model Markov process is used to describe substitutions between codons and transition/transversion rate bias and codon usage bias are allowed.

Methodology:

The assumption behind this approach is based on functional constraint i.e. functionally important residues and sequences are under stronger selective constraints that lower their evolutionary rates.

Investigation of changes in evolution was done by developing a likelihood ratio test based on Markov model of codon substitution for detecting significant rate shifts.

Our work is based on the model proposed by Goldman and Yang which has some benefits over Kimura model[27]. In this model Markov process is used to describe substitutions between codons and transition/ transversion rate bias and codon usage bias are allowed. Further selective restraints at the protein level are accommodated using physicochemical distances between the amino acids coded for by the codons. The utility of the model is illustrated on a data set of virulence gene sequences from the influenza A virus. The sample of coding sequences from homologous genes responsible for virulence was taken from influenza A virus of strain H1N1 which infects two different host types, birds and humans. Equal number of available avian, swine and human HA, NA, NS1, NS2, PB1 and PB2 sequences were downloaded from the influenza research database (http://www.fludb.org) and Chinese influenza virus database (http://influenza.psych.ac.cn/). The criteria for selecting the particular sequence is by sorting them according to host species, year and country which formed the representative set of sequences between the years 1919 till 2010.

Multiple sequence alignment was done using ClustalX (Version 2) software with the default parameters. The genealogy of the chosen isolates was inferred under the maximum likelihood (ML) criteria by the DNAML algorithm provided in the PHYLIP. The program was used to find the most significantly positive branches and their corresponding branch lengths. A very similar genealogy was inferred with the neighbor-joining algorithm provided in the PHYLIP package. Plot of trees were made with the help of drawgram program of the PHYLIP package. Editing of the trees was done with tree-view software.

A Bayesian estimate of the posterior genealogy distribution was performed using the MrBayes (ver. 3.1.2) program[23]. The estimation was performed with a general time reversible (GTR) model of substitution and a gamma distribution on rate heterogeneity.

Phylogenetic analysis by employing maximum likelihood method was done by a computer program 'PAML'. A program 'codeml' of the package was used. Branch lengths were estimated using the program under a model that let the ω ratio vary both among sites and among lineages (Yang and Nielsen, 2002). The models attempt to detect positive selection that affects only a few sites along a few lineages. The program was also used for the estimation of the most probable sites for the positive selection.

RESULT AND DISCUSSION

Analyses were performed using the genealogies estimated by maximum likelihood. Results of the analysis of both hypotheses for the six genes are shown in table. The hypothesis H_1 was approximated by a Gamma distribution of rates among the sites where as the hypothesis H_0 was approximated by a Gamma distribution plus a class of invariant sites.

The hypotheses were tested by using the LRT method. The test statistics can be given by:

$$U = -2\log L1 / L0$$

Where U is the log likelihood ratio for the models and L_1 and L_0 are the log likelihood values for hypothesis H1 and H0 respectively. Because H1 is a special case of H_0 (the hypotheses are nested), the likelihoods will always obey the relationship that $L_1 \leq L_0$. This means that *U* will never be negative. Minimum probability α of rejecting H0 is assumed to be at significance level 0.05. , is given by the equation: $U \leq (1 - \alpha) \times 100\%$.

Log Likelihood Ratio Test for H1N1 genes responsible for virulence

Virulen Gene Name	L_0	L_1	**U** (Log likelihood Ratio)	**P-value** (Probability of not H_0)
HA	- 16275.35060	-16402.01658	11.15	0.885
NA	-10451.52230	-10518.11400	8.39	0.916
NS1	-5253.752755	-5432.339626	10.37	0.896
NS2	-3018.719419	-2955.689620	-	0
PB1	-21019.46678	-20573.92549	-	0
PB2	-15809.343745	-15911.677467	9.2580	0.907

In codon usage analysis of **HA** gene average frequency of occurring of any base at any position of codon is highest for A at 0.33687 as well as for the first and third position. Yet for the position 2 the frequency of G is highest. The standard deviation of bases showed that in HA gene overall A is found to have highest standard deviation among the four bases. This result was expected according to its frequency observed. For the position 2 the standard deviation was observed at highest level for base C. Hence for position 1 and 3 A is the least conserved bases and are thus best candidate for molecular evolution. Thus the bases A and C are the most significant bases for molecular evolution process in HA gene.
Codon usage counts showed that Asparagine (Asn), Glutamate (Glu) and Lysine (Lys) are the potential amino acids for the molecular evolution under positive selection of HA gene.
In LRT test, the log-likelihood value of **NA** was observed to be at lowest. So it has less probability of undergoing positive selection. Codon frequency table shows that at position 1, frequency of G is highest while for positions 2 and 3 the frequency of A and T are more respectively. Among four bases, standard deviation value at position 1 is of base A while at other positions 2 and 3 least conservation is of G and C respectively. Though overall frequency of C in NA gene is less, it is rapidly causing substitutions at position 3. Thus position specific testing of bases reveal G at position 1, A at position 2, and C at position 3 are most evolving bases. The sum of codon usage counts elucidates the most important amino acids in NA genes as Asparagine (Asp) and Cysteine (Cys).

In same LRT test the value of **NS1** for H0 and H1 is 10.37 and next to HA gene. This proves that the virulence gene NS1 is under strong selective regime in H1N1 virus. Codon frequency table shows that overall base frequency of A is highest. At position 1 base G has the highest frequency in codons while at position 2 and 3, T and A are with high frequency respectively. Among four bases, standard deviation value at position 1 is of base A while at other positions 2 and 3, G and T respectively are least conserved. Thus position specific substitutions of bases reveal A at position 1, G at position 2, and T at position 3 are most evolving bases.
The sum of codon usage counts elucidates the most important amino acids in NS1 genes as Glutamate (Glu) and Threonine (Thr).

In last when we do same test for **PB2** gene average frequency of occurring of any base at any position of codon is highest for A at 0.37786 as well as for the first and third position. Yet for the position 2 the frequency of T is highest. The standard deviation of bases showed that in PB2 gene overall A is found to have highest standard deviation among the four bases. For the position 2 the standard deviation was observed at highest level for base G. Hence for position 1 and 3 A is the least

conserved base and are thus best candidate for molecular evolution. Thus the bases A and G are the most significant bases for molecular evolution process in PB2 gene. Sum of codon usage counts is one parameter to identify the potential amino acids undergoing adaptive molecular evolution. In PB2 gene, Threonine (Thr) and Arginine (Arg) are important bases which can undergo non-synonymous substitution.

The trees for the different genes are shown in fig-1, fig-2, fig-3 and fig-4 for HA, NA, NS1 and PB2 gene respectively. The graph of Simulation Study of U of HA Gene is shown in fig-5.

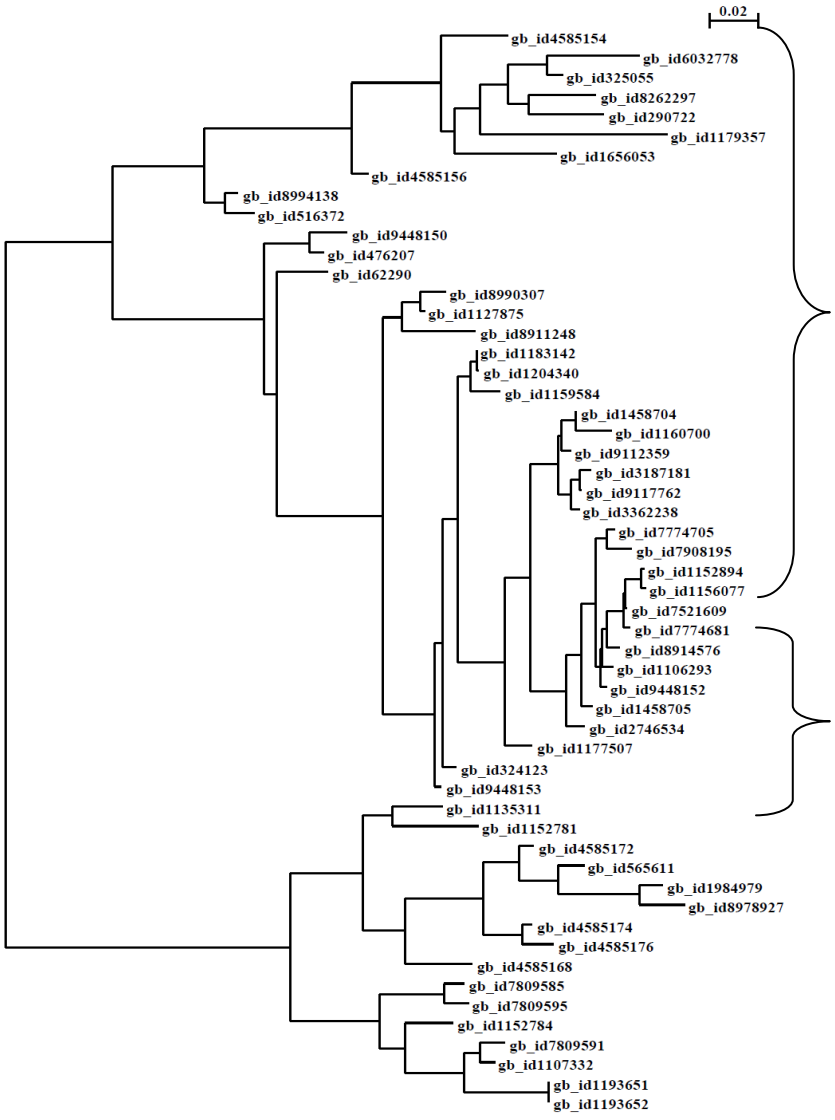

Figure 1

Genealogy of HA sequences used in the study using ML method. Sequences are listed in standard genebank accession numbers and the branch length in units of expected substitutions per codon is indicated by the scale bar.

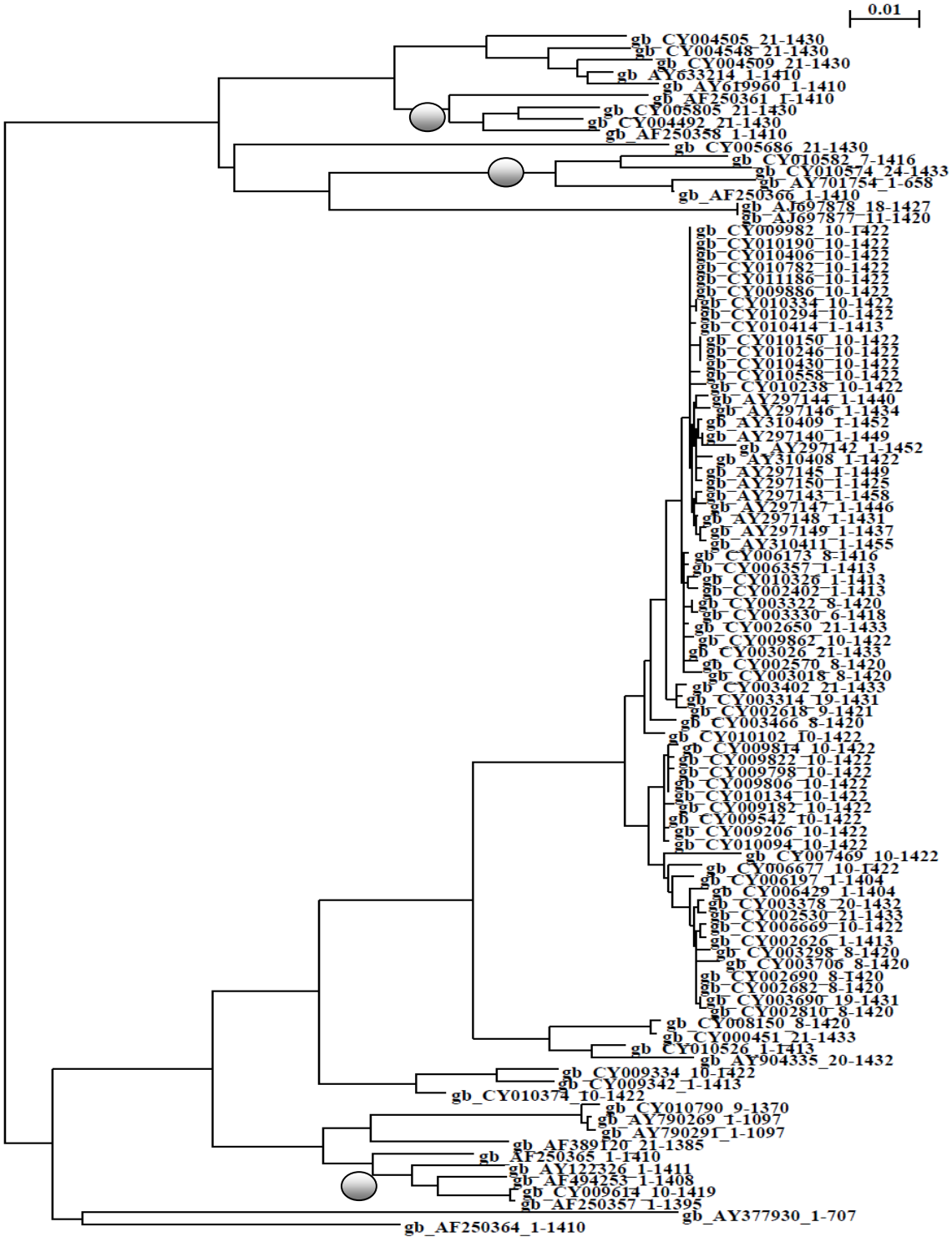

Figure 2

Genealogy of NA sequences used in the study using ML method.

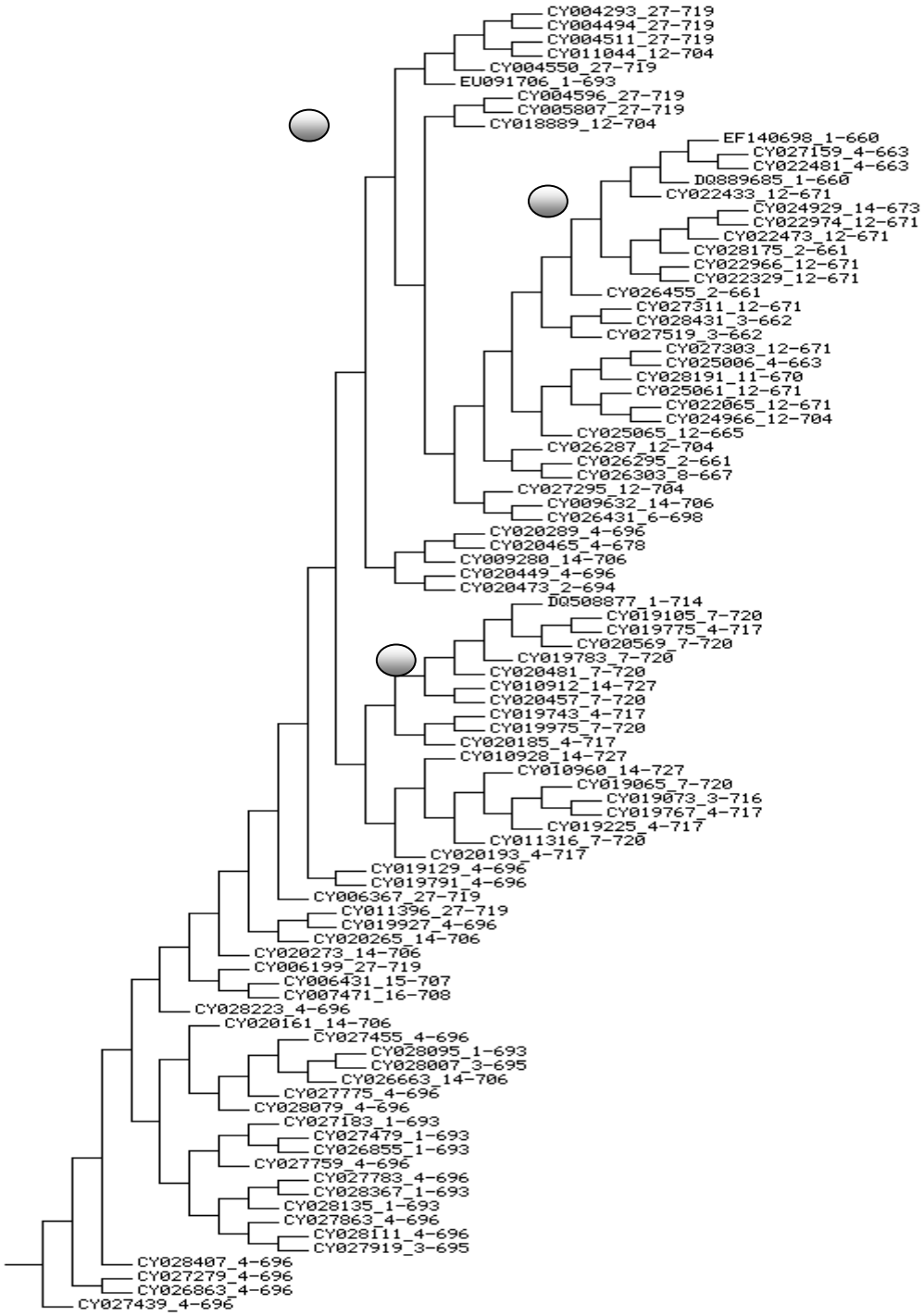

Figure- 3
The genealogy study of NS1 gene by ML method

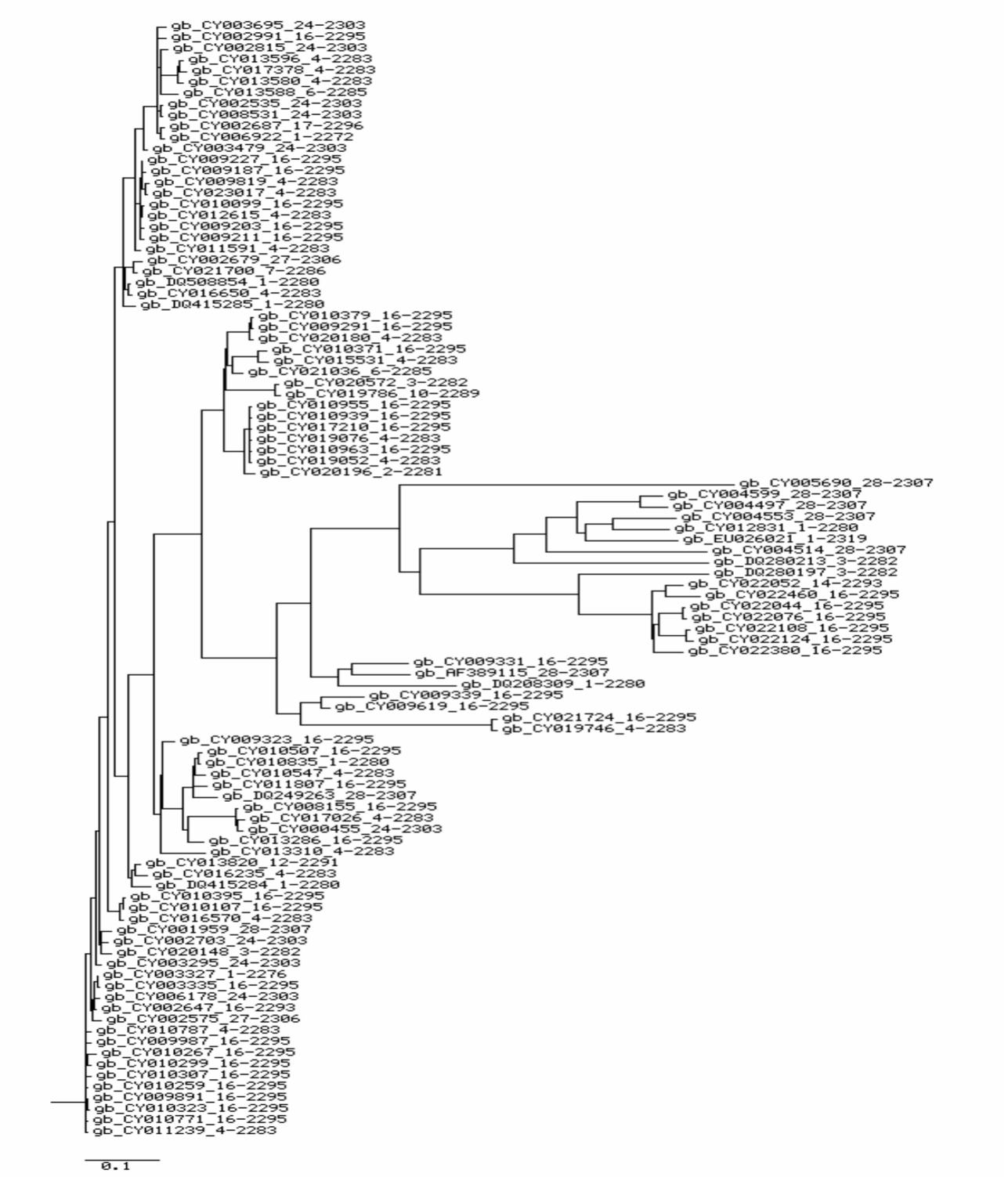

Figure- 4
The genealogy study of PB2 Gene by ML method

Codon position vs. Base table of HA, NA, NS1, PB2 genes

Base ⇩	**HA**(1st position)	**HA**(2nd Position)	**HA**(3rd position)	**NA**(1st position)	**NA**(2nd position)	**NA**(3rd position)	**NS1**(1st position)	**NS1**(2nd position)	**NS1**(3rd position)	**PB2**(1st position)	**PB2**(2nd position)	**PB2**(3rd position)
A	0.33687	0.18601	0.32693	0.28835	0.36186	0.23619	0.30776	0.28517	0.27620	0.37786	0.29616	0.35730
T	0.21865	0.24853	0.25593	0.23423	0.18258	0.31557	0.14375	0.30305	0.27358	0.14398	0.30741	0.23184
G	0.29926	0.35569	0.19878	0.34144	0.26247	0.20948	0.32584	0.19883	0.23623	0.31073	0.18976	0.23827
C	0.14522	0.20977	0.21835	0.13598	0.19309	0.23876	0.22266	0.21295	0.21399	0.16743	0.20667	0.17259

Standard Deviation for a given position among bases for HA, NA, NS1 and PB2 genes

Base ⇩	**HA**(1st position)	**HA**(2nd position)	**HA**(3rd position)	**NA**(1st position)	**NA**(2nd position)	**NA**(3rd position)	**NS1**(1st position)	**NS1**(2nd position)	**NS1**(3rd position)	**PB2**(1st position)	**PB2**(2nd position)	**PB2**(3rd position)
A	0.008607	0.006657506	0.01964	0.016216	0.006129	0.012949	0.01857	0.005886	0.009671	0.002674	0.001976	0.0111
T	0.005	0.004347152	0.016397	0.011071	0.00554	0.016244	0.012559	0.005074	0.011265	0.002511	0.00167	0.012565
G	0.00607	0.004876	0.01286	0.010286	0.007532	0.011891	0.007739	0.007243	0.009656	0.00246	0.003215	0.012855
C	0.005867	0.00880915	0.014822	0.007712	0.005254	0.016943	0.012618	0.006261	0.009027	0.002177	0.001096	0.010399

Simulation Study of U of HA Gene:

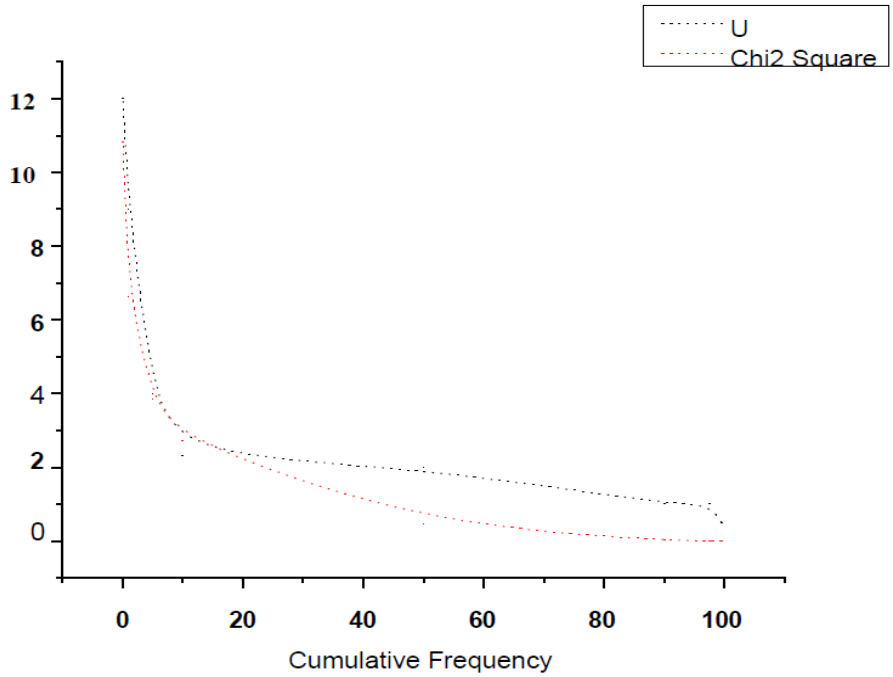

Figure- 5

The Chi-2 Distribution with one degree of freedom compared with a simulation study of U for HA gene

CONCLUSION

We infer from our study that the most evolving gene responsible for virulence in swine influenza virus A of serotype H1N1 is HA followed by NS1, PB2 and NA. For two genes that we studied NS2 and PB1 the LRT statistics fail to fit the null model. Thus for NS2 and PB1 the null model is rejected. Our motivation for doing this analysis was to examine the H1N1 virus for host radiation but the study reveals the problem of substitution rates among different groups of organisms and determining whether the selective pressures differ in them. The codon based approach which we used uses comparisons of two different nucleotide substitutions are done namely synonymous and non-synonymous substitutions. The inference of this was done using maximum likelihood approach in codeml of PAML package, Thus enabling use of full information in nucleotide sequence and also include the known biological phenomenon called transition vs transversion ratio (Ts/ Tv). However the gamma parameter used in the study is a very crude indicator and the indicators such as, e.g., shifts between biochemically different groups of amino acids etc. can improve the study. We have assumed that the rates of synonymous and non synonymous substitutions are relative to changes in mutational rates obtained after the viral parasites alter their hosts. However for the genes under strong positive selection this assumption does not hold as the process of fixation is governed by factors other than availability of mutations. Further codon usage studies of genes showing virulence according to our model, The most important codons

responsible for molecular evolution are Asp and Glu for HA, Asp and Cys for NA, Glu and Thr for NS1, Lys and Arg for PB2 gene. . We believe that our study may prove to be useful to identify candidate genes and codons for the molecular biological investigation of species-specific adaptation in viruses.

ACKNOWLEDGMENT

We are thankful to our Department of Biotechnology, Government of India for the funds to set up a Sub- Distributed Information Center (BTISnet SubDIC) at our Department of Biotechnology, Birla Institute of Technology, Mesra where this work was done.

REFERENCES

1. Forsberg R, Bugge F, Christiansen, A Codon-Based Model of Host-Specific Selection in Parasites, with an Application to the Influenza A Virus*. Mol. Bio. Evol, Vol. 20, No. 8, (2003).*
2. Knudsen, B, Miyamoto M M, A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc. Natl. Acad. Sci, USA* 98:14512–14517, (2001).
3. Goldman N, Yang, A Codon-based Model of Nucleotide Substitution for Protein coding DNA Sequences. *Mol. Bio. Evol*, 11(5):725-736, (1994).
4. Yang Z, Nielsen N, Goldman N, Petersen, Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics,* 155:431–449, (2000).
5. Yang Z, PAML: A program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences*, 13:555-556, (1997).
6. Yang Z, Biewelski J, Statistical methods for detecting molecular adaptation. *TREE* vol. 15(12), (2000).
7. Yang Z, Nielsen R, Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Bio. Evol*, 19:908-917, (2002).
8. Yang Z, Bielawski JP, Anisimova M, Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution. *Mol. Biol. Evol,* 18(8):1585–1592, (2001).
9. Yang Z, Wong WSW, Nielsen R, (2005), Bayes Empirical Bayes Inference of Amino Acid Sites under Positive Selection. *Mol Biol Evol*, 22:1107–1118, (2005).
10. Piontkivska H, Efficiencies of maximum likelihood methods of phylogenetic inferences when different substitution models are used. *Mol. Phyl. And Evol,* 31 865–873, (2002).
11. Shen J, Ma J, Wang Q, Evolutionary Trends of A (H1N1) Influenza Virus Hemagglutinin since 1918. *PLoS ONE 4(11): e7789*, (2009).
12. Zimmer SM, Burke DS, Historical perspective—emergence of influenza A (H1N1) viruses. N Engl J Med, 361(3):279-85, (2009).
13. Vaillant L, La Ruche G, Tarantola A, et al. Epidemiology of fatal cases associated with pandemic H1N1 influenza 2009.Eurosurveill, 14(33), (2009).
14. Rocky K, Partho H and Raju P, Adaptive molecular evolution of virulence genes of avian influenza – A virus subtype H5N1: An analysis of host radiation, *Bioinformation,* 1(8): 321-326 (2006).
15. Sinha NK, Roy A, Das B, Das S, Basak S, (2009), Evolutionary complexities of swine flu H1N1 gene sequences of 2009. *Biochem and Biophysic Res Comm.*, (2009).

16. Goni N, Fajardo A, Moratorio G, Colina R, Cristina J, Modeling gene sequences over time in 2009 H1N1 Influenza A Virus population. *Virology Journal*, 6:215, (2009).
17. Adiego SB, Omenaca TM, Martinez CS, et al. Human cases of swine influenza A (H1N1), Aragon, Spain. Eurosurveill, 14(7):19120, (2009).
18. Fraser C, Donnelly CA, Cauchemez S, et al. Pandemic potential of a strain of influenza A (H1N1): early findings. Science, 324:1557-1561, (2009).
19. Gani R, Hughes H, Fleming D, et al. Potential impact of antiviral drug use during influenza pandemic. Emerg Infect Dis, 11(9):1355-62, (2009).
20. Crill WD, Wichman HA, Bull JJ, Evolutionary reversals during viral adaptation to alternating hosts. *Genetics*, 154:27-37, (2000).
21. Felsenstein, J, Churchill GA, A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Bio.Evol,* 13: 93-104, (1996).
22. Gatherer D, The 2009 H1N1 influenza outbreak in its historical context. *Journal of Clinical Virology* 45, 174–178, (2009).
23. Huelsenbeck JP, Ronquist F, MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics,* 17:754–755, (2001).
24. Thompson J.D, Gibson T.J, Plewniak F, Jeanmougin F, and Higgins D.G, The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acid Res,* 24:4876-4882, (1997).
25. Turner P. E, Elena SF, Cost of host radiation in an RNA virus. *Genetics,* 156:1465-1470, (2000).
26. Ren F, Tanaka H, Yang Z, An Empirical Examination of the Utility of Codon-Substitution Models in Phylogeny Reconstruction Systematic Biology. 54(5):808-818, (2005)
27. Kimura M, The Neutral Theory of Molecular Evolution, 1st edition. Cambridge University Press, Cambridge, (1983).