Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Adaptive molecular evolution of virulence genes of avian influenza - A virus subtype H5N1: An analysis of host radiation

Rocky Kumar, Partho Halder and Raju Poddar*

Department of Biotechnology, Birla Institute of Technology, Mesra, India; Raju Poddar* - Email - rpoddar@bitmesra.ac.in; Phone: +91 651 2276223; Fax: +91 651 22755401;

* Corresponding author

received December 11, 2006; revised December 24, 2006; accepted December 25, 2006; published online December 26, 2006

Abstract:

The phenomenon of host radiation is strongly influenced by the rates of mutation of their virulence genes. We have studied the molecular evolution of virulence genes (HA, NS, PB2) of the Avian Influenza Virus H5N1 from avian to human hosts. We used a site-specific comparison of synonymous (silent) and non-synonymous (amino acid altering) nucleotide substitutions for the three chosen genes in parasite populations from different hosts. Analyses were made using Maximum Likelihood (ML) genealogies for the null and alternate hypothesis based on differential gamma distribution rates. The null hypothesis had a higher rate of substitution and was found to be more suitable for all the studied genes by Likelihood Ratio Test (LRT). The study showed the NS gene to be having the fastest rate of evolution.

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

Background:

A common phenomenon in parasite evolution is their ability to expand their niche to new host species. This process of acquiring of new host ranges is called host change event or host radiation. The selective force on a particular parasite gene may change after host radiation due to reasons like adaptive alteration of protein functionality or host's immune-mediated selection. Since a host change event comprises three stages namely transmission to new host species, replication within new host and transmission between individuals of the new host species, with the last two steps being rate limiting, host radiation events are rarely successful. According to the general theory of 'ecological specialization' the host specific adaptations act as ecological barriers. **[1]** These host change events though rare, do occur and may be detrimental to the new host, some eminent examples of which are the Spanish Flu pandemic and the latest H5N1 outbreaks.

The avian influenza A virus is commonly found in most wild birds especially migratory water fowl and wild ducks which act as their natural reservoirs and carriers. It is however fatally infectious to domestic birds like chicken, duck, turkeys etc. There are many subtypes of the virus differing in the combination of subtypes of HA (Hemagglutinin) and NA (Neuraminidase) surface proteins. Human infections were first reported in Hong Kong (1997) with the latest one being the H5N1 outbreak. **[2]** Though the outbreaks were local and man to man transmission was rare, the highly mutable nature of influenza virus raises concerns. This is due to the lack of 'proofreading' mechanisms and repair of errors that occur during replication. Little or no immune protection among human population due to lack of prior infection, may lead to pandemic outbreaks. **[3]**

Past influenza pandemics have led to high levels of illness, death, social disruption and economic loss. There were three major influenza-A pandemics during the 20th century, namely (1) Spanish flu (H1N1) 1918-19, (2) Asian flu (H2N2) 1957-58, and (3) Hong Kong flu (H3N2) 1968-69, which caused thousands of death in the United States. Of the various subtypes, H5N1 is of greatest pandemic concern currently because of the following reasons: rapid spread throughout poultry flocks in Asia; endemic outbreaks in eastern Asia; highly rapid rate of mutation; propensity to acquire genes from viruses infecting other animal species; causes severe disease in humans with a high fatality rate of approximately 70 %. **[4]**

Studies on the genetic basis revealed that three genes of avian influenza virus are responsible for their virulence. These are HA, NS and PB2 genes. HA (hemagglutinin) gene is responsible for high cleavability of the hemagglutinin glycoprotein. It is like glue that binds to the sialic acid on cell receptors and hampers its function, NS (non structural) gene antagonizes the induction of interferon protein levels and leads to lethal viral infection **[5]** and PB2 gene encodes an internal polymerase that influences the outcome of infection. **[6, 7]** Another study showed that a mutation at position 627 in the gene PB2**,** which identifying positions in the parasite genome underlying the phenotypic differences between host-specific strains may give insights about the molecular basis of species-specific adaptation.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

The fundamental objective behind our work is to study how selection acts on variants of genes responsible for virulence i.e. HA, NS and PB2 genes of different hosts comprising birds and humans. To do so, we use a site-specific comparison of synonymous (silent) and non-synonymous (amino acid altering) nucleotide substitutions in the 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 H5N1 avian influenza virus that has undergone host radiation from birds to humans.

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. **[8]** 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 H5N1 which infects two different host types, birds and humans. The sequences of HA**,** NS and PB2 genes of viruses isolated from aves and humans were downloaded from the Influenza Sequence Database at http://www.flu.lanl.gov/.

Multiple sequence alignment was done using ClustalX (ver 1.81) software. **[9]** The genealogy of the chosen isolates was inferred under the maximum likelihood (ML) criteria by the dnaML algorithm **[10]** provided in PHYLIP package (ver 3.6, 2004) which can be downloaded from http://evolution.genetics.washington.edu/phylip.html. The program was used to find the most significantly positive branches and their corresponding branch lengths. Plot of trees were made with the help of drawgram program of the PHYLIP package. A Bayesian estimate of the posterior genealogy distribution was performed using the MrBayes (ver 3.1.2) program. **[11]** This can be downloaded from http://mrbayes.net. 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' **[12]** which can be downloaded from the Web site http://abacus.gene.ucl.ac.uk/software/paml.html. A program 'codeml' of the package was used to find the base frequencies and different codon usage. All the statistical analyses were done by statistical software 'OriginPro 7.5 SRO' from Origin Lab Corp. USA.

Discussion:

Analyses were performed using the genealogies estimated by maximum likelihood. Results of the analysis of both hypotheses for the three genes are shown in table 1. 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.

Gene responsible for virulence	Likelihood value of hypothesis H_0	Likelihood value of hypothesis H_1	U (log likelihood ratio)	P value
HA	-7266.62128	-7755.29608	12.3834	0.8761
NS	-5745.40265	-6008.68716	11.1465	0.8853
PB2	-10923.83329	-12150.27347	14.2237	0.85776

Table 1: Analysis of the genes from the influenza A virus using ML method

The hypotheses were tested by using the LRT method. **[13]** The test statistics can be given by:

$$U = -2log\ L_1 / L_0$$

Where U is the log likelihood ratio for the models and L_1 and L_0 are the log likelihood values for hypothesis H_1 and H_0 respectively. Because H_1 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 'p' of not rejecting H_0, is given by the equation: $U \leq (1 - \alpha) \times 100\%$ (where α equals the probability of rejecting H_0).

The test showed that the hypothesis H_0 gives a better fit than hypothesis H_1. Besides the H_0 model having an additional class of invariant sites is responsible for faster rate of evolution. Further the probability of not rejecting hypothesis H_0 is highest in case of the NS gene followed by that of HA and PB2. Thus we can infer that NS has the fastest rate of evolution and seems to be most significant for molecular adaptation of the parasite. This was also confirmed by determining the trees generated for these genes by a maximum likelihood algorithm. The trees for the different genes are shown in fig-1, fig-2 and fig-3 for NS, HA and PB2 gene respectively.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

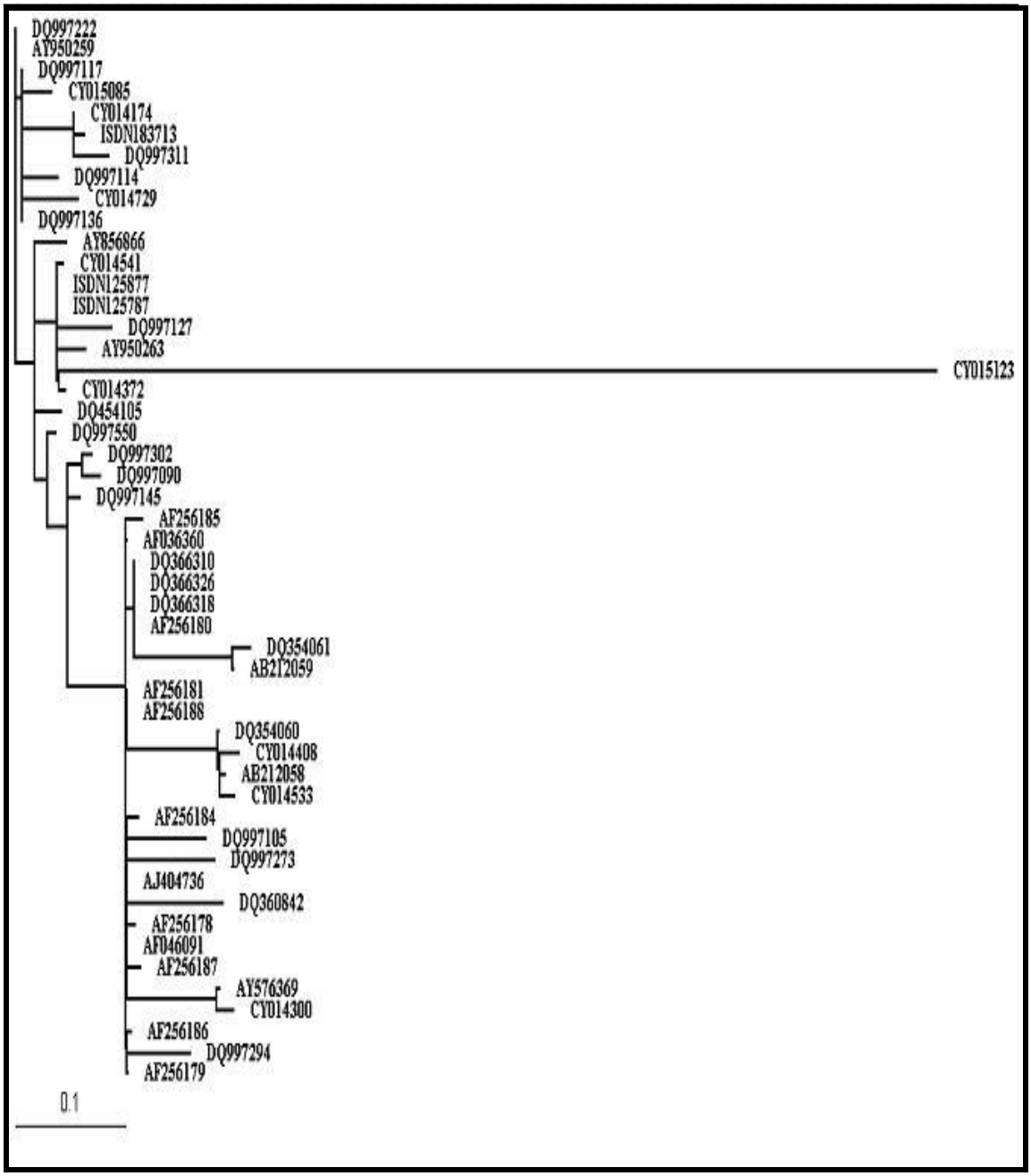

Figure 1: Genealogy of the NS sequences used in the study using ML method

As per table 2, it has also been shown that the hypothesis H_0 gives a wide range of values for the rates of change for all the considered genes, as compared to that of H_1 meaning that the chances of positive selection is greater in the model H_0.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

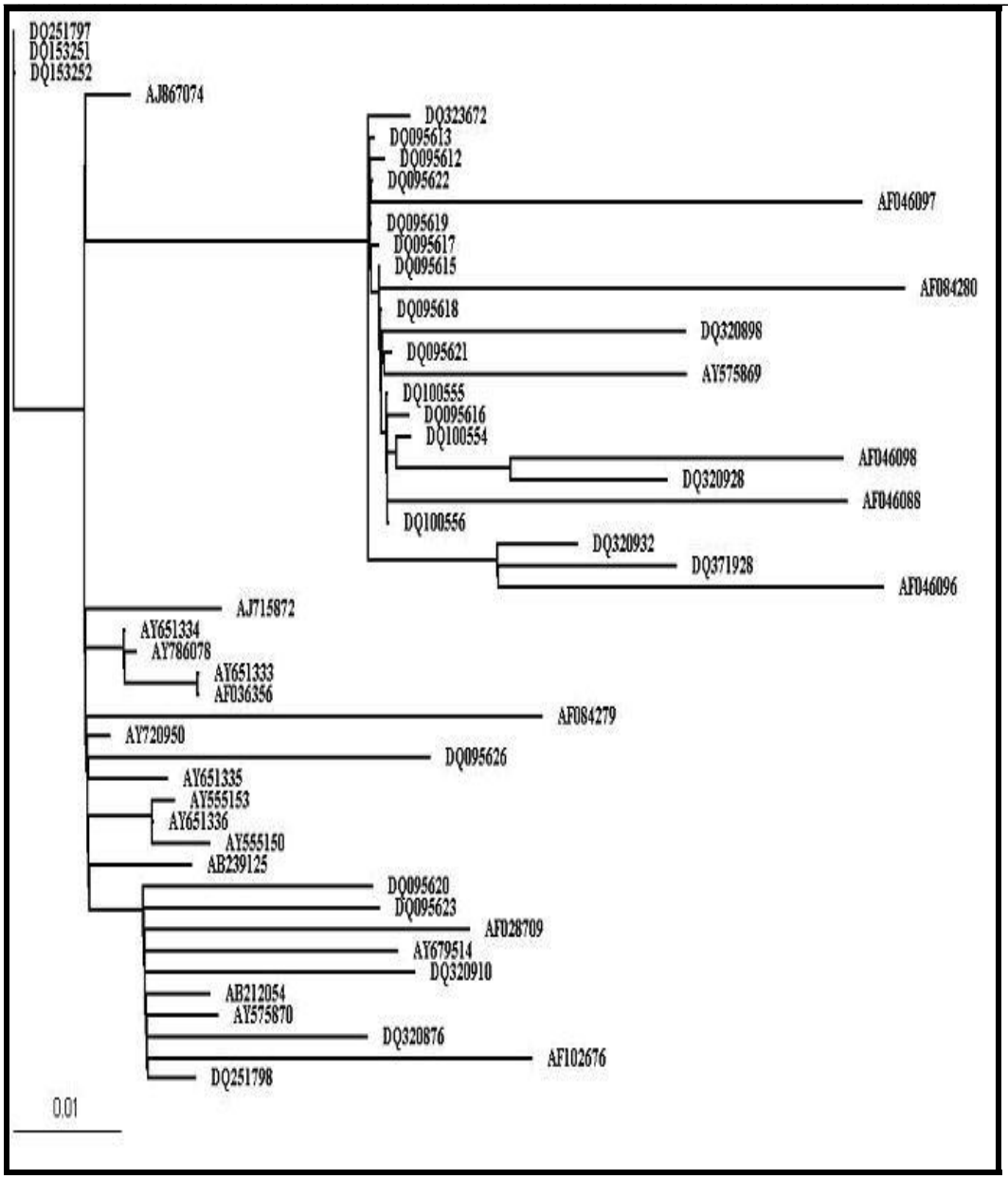

Figure 2: Genealogy of the HA sequences used in the study using ML method. Sequences are listed with GeneBank accession (A/C) numbers and the branch length in units of expected substitutions per codon is indicated by the scale bar

State in HMM	Rates of change		Probability	
	H_1	H_0	H_1	H_0
1	0.528	0.359	0.474	0.895
2	9.472	6.110	0.026	0.103
3	0.000	20.531	0.500	0.0025

Table 2: Estimation of rates of change and their probability for both hypotheses

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

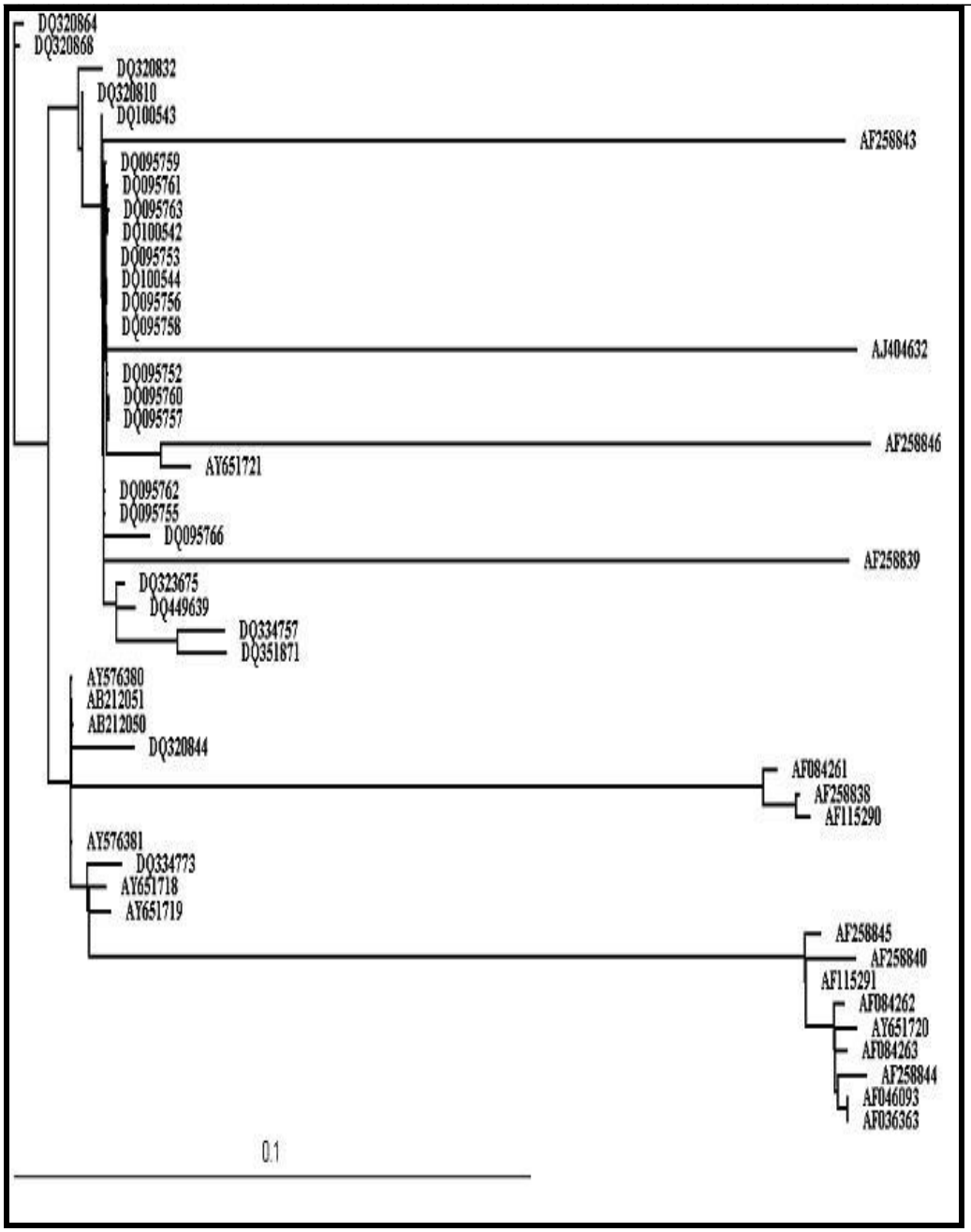

Figure 3: Genealogy of the PB2 sequences used in the study using ML method

Further in our study on codon usage, we found that the most important base responsible for non-synonymous substitution in case of HA genes is T at first position and C for second position as they have the highest standard deviation among the four bases for a given position which means that they are the least conserved bases and have high chances of getting substituted and thus are important candidate for molecular evolution. Similarly for PB2 the most important base at first and second position are A and C respectively and for NS gene the most important base at first and second position are T and C respectively. The results are shown in the table 3.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Base	HA	HA	PB2	PB2	NS	NS
	*SD for 1st position	SD for 2nd position	SD for 1st position	SD for 2nd position	SD for 1st position	SD for 2nd position
T	0.00609	0.00337	0.00129	0.00196	0.00632	0.00254
C	0.00422	0.00446	0.0019	0.00256	0.00327	0.00442
A	0.00345	0.00313	0.00452	0.0014	0.00475	0.0016
G	0.00338	0.00333	0.0035	0.00139	0.00312	0.00297

Table 3: Values of standard deviation of base frequency for different codon position and different bases
*SD = Standard deviation

The result also showed that the standard deviation for substitution of bases of NS gene was highest among all the three considered genes which shows and confirms that NS gene is the most evolving gene followed by HA and PB2. These results were also confirmed with the simulation studies for Markov Chain Monte Carlo samples based on a total of 19502 samples from 2 runs. Each run produced 10001 samples of which 9751 samples were included (data not shown).

Conclusion:

We found out that during the process of molecular evolution of avian influenza virus from birds to humans, the most important gene responsible for causing virulence in humans is NS followed by HA and PB2. Our motivation here was to study the evolution of virus undergoing host radiation, but the study addresses the more general problem of describing the substitution process in two groups of related organisms. The codon-based approach which we used uses a comparison between the two different types of nucleotide substitution, thus enabling use of the full information in the nucleotide sequence and also includes the known biological phenomenon of transition-transversion bias. 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 nonsynonymous substitution scale identically with changes in the mutation rate after a viral host change because of alterations. This scaling can be hold for sites where all amino acid substitutions are either neutral or strongly deleterious. However, for sites undergoing positive selection if the process of fixation is limited by factors other than the availability of mutations, this may not be the case. 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.

Acknowledgement:

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:

[01] P. E. Turner & S. F. Elena, *Genetics,* 156:1465 (2000) [PMID: 11102349]
[02] http://www.who.int/csr/disease/avian_influenza/en/
[03] R. J. Webby & R. G. Webster, *Science*, 302:1519 (2003) [PMID: 14645836]
[04] www.cidrap.umn.edu
[05] Z. Li, *et al*., *J Virol.,* 80:11115 (2006) [PMID: 16971424]
[06] M. T. Hughes, *et al*, *J Virol*., 74:5206 (2000) [PMID: 10799596]
[07] M. Hatta, *et al*., *Science,* 293:1840 (2001) [PMID: 11546875]
[08] N. Goldman & Z. Yang, *Mol. Bio. Evol.,* 11:725 (1994) [PMID: 7968486]
[09] J. D. Thompson *et al*., *Nucleic Acids Res.*, 24:4876 (1997) [PMID: 9396791]
[10] J. Felsenstein & G. A. Churchill, *Mol Biol Evol.,* 13:93 (1996) [PMID: 8583911]
[11] J. P. Huelsenbeck & F. Ronquist, *Bioinformatics*, 17:754 (2001) [PMID: 11524383]
[12] Z. Yang, *Computer Applications in BioSciences*, 13:555 (1997) [PMID: 9367129]
[13] B. Knudsen & M. M. Miyamoto, *PNAS,* 98:14512 (2001) [PMID: 11734650]

Edited by P. Kangueane

Citation: Kumar *et al.*, Bioinformation 1(8): 321-326 (2006)