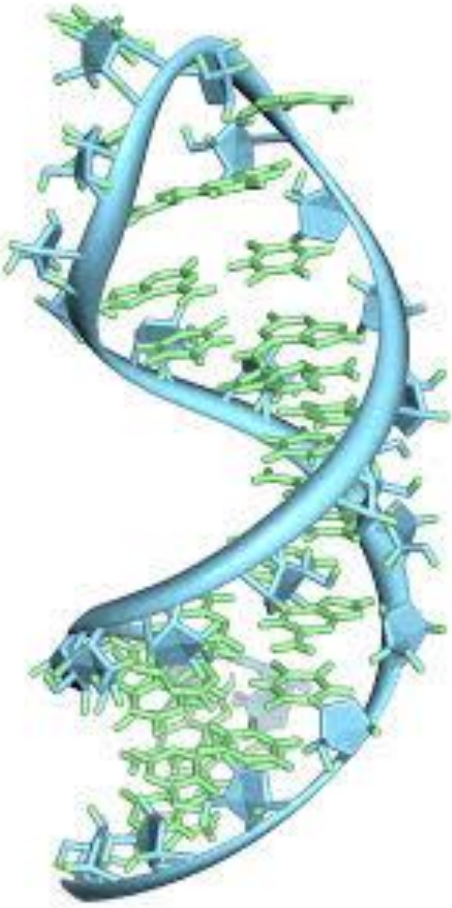


RESEARCH ARTICLES

J. Microbiol. Biotechnol. (2016), 26(11), 1972–1982
http://dx.doi.org/10.4014/jmb.1605.05016

jmb

Insights into the Usage of Nucleobase Triplets and Codon Context Pattern in Five Influenza A Virus Subtypes S

Himangshu Deka and Supriyo Chakraborty*

Department of Biotechnology, Assam University, Silchar-788011, Assam, India

Received: May 9, 2016
Revised: July 19, 2016
Accepted: July 20, 2016

First published online
July 27, 2016

**Corresponding author*
Phone: +91-9435700831;
Fax: +91-3842-270802;
E-mail: supriyoch2008@gmail.com

Supplementary data for this paper are available on-line only at http://jmb.or.kr.

pISSN 1017-7825, eISSN 1738-8872

Copyright© 2016 by
The Korean Society for Microbiology
and Biotechnology

Influenza A virus is a single-stranded RNA virus with a genome of negative polarity. Owing to the antigenic diversity and cross concrete shift, an immense number of novel strains have developed astronomically over the years. The present work deals with the codon utilization partialness among five different influenza A viruses isolated from human hosts. All the subtypes showed the homogeneous pattern of nucleotide utilization with a little variation in their utilization frequencies. A lower bias in codon utilization was observed in all the subtypes as reflected by higher magnitudes of an efficacious number of codons. Dinucleotide analysis showed very low CpG utilization and a high predilection of A/T-ending codons. The H5N1 subtype showed noticeable deviation from the rest. Codon pair context analysis showed remarkable depletion of NNC-GNN and NNT-ANN contexts. The findings alluded towards GC-compositional partialness playing a vital role, which is reflected in the consequential positive correlation between the GC contents at different codon positions. Untangling the codon utilization profile would significantly contribute to identifying novel drug targets that will pacify the search for antivirals against this virus.

Keywords: Codon usage bias, influenza A virus, preferred codon, dinucleotide, codon pair context

Introduction

Influenza A virus (IAV) belonging to the *Orthomyxoviridae* family, is an RNA virus with an 8-segmented genome, and it has been instrumental in causing serious mortality and morbidity across the globe with a good number of outbreaks over the years. Owing to its changing antigenic region, it has been difficult to develop a fruitful vaccine to tackle this highly infectious virus circulating in human, avian, and swine hosts. There are a large number of IAV strains reported worldwide; however, very few have been detected infecting humans. The IAV subtypes that have been found circulating in humans include H1N1, H1N2, H2N2, H3N2, and H5N1 [8, 11, 32, 34].

The phenomenon of codon usage bias (CUB) refers to the unequal usage of synonymous codons that have been reported in almost all groups of organisms, including humans [2, 6, 9, 12, 31]. Two major factors are being projected for inflicting the CUB; mutational pressure and natural selection. However, there are other factors responsible for CUB and that have been reported by various authors. These include nucleotide composition [26], gene length [21], hydrophobicity [29], environment effect [37], etc. The divergence from the standard genetic code may perhaps have a severe effect on the translational machinery. In opposition, unalterable changes to a species' translational machinery may compel it to adapt its CUB consequently. Genetically close species generally display a similar codon usage pattern. Thus, any dissimilarity among the organisms is reflected in the deviation of codon usage occurring among them [9]. Synonymous codon usage during translation is a non-arbitrary process, which makes it crucial to recognize the CUB patterns in order to determine the mode of translational selection of protein coding genes.

Typically, RNA viruses demonstrate a very low level of codon bias, which was been reported by the works of various investigators. There are several reports on IAV itself; most of them concentrate on the surface proteins hemagglutinin and neuraminidase [1, 35, 39]. However, the high variability and continual evolution of the different subtypes calls for a deeper insight into the CUB pattern among this highly variable infectious virus. The present investigation was undertaken to understand the codon usage patterns in a comparative manner among five IAV subtypes (*viz.*, H1N1, H1N2, H2N2, H3N2, and H5N1) isolated from human hosts.

Methods

Sequence Datasets

In this study, a total of 787 complete coding sequences (cds) of eight different genes (*viz.*, hemagglutinin (HA), neuraminidase (NA), nucleoprotein (NP), matrix protein (M1 and M2), polymerase acidic (PA), and polymerase basic (PB1 and PB2)), belonging to five IAV subtypes were used. All the sequences were retrieved from the GenBank database of NCBI (http://www.ncbi.nlm.nih.gov/). The accession numbers and other information about the genes are provided in Supplementary File S1.

Indices for Codon Usage Bias Study

The nucleotide composition has been traditionally regarded as the key player in shaping the codon usage pattern in the genes. The crucial role of the nucleotide composition is evident from the fact that most of the indices of CUB are based on the base composition of the genes. Among all the compositional parameters, the GC pattern has played a very highly influential role from the codon usage perspective in most of the genes. In the study, GC_3 is the frequency of the nucleotides G+C at the synonymous 3rd positions of the codons, excluding the *Met*, *Trp*, and the termination codons. Similarly, GC_1 and GC_2 represent the G+C frequency at the 1st and 2nd codon positions. GC_3 is a good indicator of the extent of base composition bias.

Relative synonymous codon usage (RSCU) [31] is a widely used index for investigating the synonymous codon usage pattern across genes and genomes. RSCU is defined as the ratio of the observed frequency to the expected frequency, assuming that all the synonymous codons for those amino acids are used equally. The synonymous codons are said to be randomly and equally used if the RSCU value is close to 1.0. The positively biased codons have a RSCU value of more than 1.0, whereas a value of less than 1.0 means a negative CUB.

The effective number of codons (Nc) is a parameter that reflects the extent of biasness towards the synonymous codons in a gene [36]. It is estimated to quantify the synonymous codon usage across the target sequence, which is calculated as given below:

$$\mathrm{Nc} = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

where F_k (k = 2, 3, 4, or 6) is the average of the F values for k-fold degenerate amino acids. The F value signifies the probability that the randomly chosen synonymous codons for an amino acid are identical. The boundary values of Nc are 20−61, the minimum being 20, when only one codon is used per amino acid, whereas a value of 61 means all the synonymous codons are equally used for each amino acid [7, 24, 36]. The codon bias is considered low if the Nc value is greater than 40.

The dinucleotide odds ratio was estimated as the ratio of observed count of a dinucleotide pair to the frequencies of the individual nucleotides constituting the dinucleotide pair. The following equation was used to calculate the odds ratio:

$$\rho xy = \frac{fxy}{fxfy}$$

where fx and fy represents the frequencies of mononucleotides x and y, respectively, and fxy denotes the frequency of the dinucleotide constituted by x and y [19]. The range 0.78−1.23 is considered as the boundary values of odds ratios. A value below 0.78 means a significantly low odds ratio, and any value greater than 1.23 is considered as over-representation [5].

The codon adaptation index (CAI) is a commonly used index to enumerate the adaptiveness of synonymous codons of a gene towards the codon usage of highly expressed genes. The CAI is also used as a predictor of gene expressivity. This index was first used by Sharp and Li [31] while studying the CUB in *E. coli.* CAI was originally proposed to provide a normalized estimate that can be used across genes and species. CAI values range from 0 to 1. A CAI value of 1 is assigned to the most frequent codons within a gene, whereas the least frequent codons are assigned a CAI value of 0 [10, 27]. CAI is estimated as

$$\mathrm{CAI} = \exp\frac{1}{L}\sum_{k=1}^{L} \ln w_{c(k)}$$

where L is the number of codons in the gene and $w_{c(k)}$ is the ω value for the k-th codon in the gene.

The CUB measures (*viz.*, GCs, RSCU, Nc, and CAI) for each coding sequence were estimated in our study by using an in-house Perl program developed by SC (author).

Neutrality Analysis

An analytical method for assessing codon usage is the neutrality plot. In this analysis, the mean GC contents at the first and second codon positions, represented by GC12, are plotted as the ordinates, and GC3 values are plotted as the abscissa in a scatterplot. In this plot, a statistically significant correlation between GC12 and GC3, and a regression line with close to 1 implies that mutation bias could be the central force influencing codon usage. On the contrary, selection in opposition to mutation bias may lead to a constricted distribution of GC content, which is reflected in a lack of correlation between GC12 and GC3 [33].

Multivariate Statistical Analysis

Correspondence analysis (CA) is a multivariate dimension reduction method for efficient comparison of large scale information in a two-dimensional plot. Using this method, variable types represented as rows and columns are displayed in a low-dimensional scatter diagram, which replaces the complexity of the original data [12, 14]. This analysis was implemented using the Past3 program [16].

Codon-Pair Context Analysis

Codon-pair context represents the codon pairs harbored by the ribosomal A- and P-sites. All codon context analyses were executed using the Anaconda program ver. 2.0 [22]. The alliance of codon-pairs is estimated using the chi-square test of independence. Based on the adjusted residual values for the contingency table, the preferential and the rejected codon-pairs are recognized and displayed in a 64 × 64 color-coded map. The map imparts an overall view of the codon-pair context data.

Results

Compositional Properties in the Influenza A Virus Genes

The 787 coding sequences (cds) were examined for their nucleobase composition, which reveals a lack of much deviance among the five selected subtypes (Table 1). The genes were found to possess a lower GC content (mean ± SD = 44.5 ± 1.8). The overall GC content in the M1 gene was found to be the highest in all subtypes, except for H5N1 where NP recorded the highest value for GC content, both overall as well as at the wobble position. The mononucleotides followed the decreasing order of A > G > T > C in almost all the subtypes and across all the genes, but with varying magnitudes. Whereas most of the genes across the subtypes showed inclination towards usage of A/T at the silent position, H5N1 showed sharp deviation from this observation by showing a preference for A/G at the third position.

Codon Usage Analysis in IAV Genes

To scrutinize whether IAV genes exhibit a similar codon usage pattern, the effective number of codon (Nc) values were estimated. The values were in the range of 44–56, with an average of 51.7 ± 2.3. The overall value of Nc >40 indicates weak bias prevailing in the genes of IAV. The Nc values showed significant positive correlations with GC (r = 0.308, $p < 0.001$), GC3 (r = 0.745, $p < 0.001$), and CAI (r = 0.171, $p < 0.05$).

The analysis of RSCU presented a complex picture of the codon usage in the IAV genes across the subtypes. The preference of codons in different genes was different, but in the majority of the cases, the preferred codon ended with A/T. When we compared the subtypes based on their RSCU values, a similarity in codon preferences was observed between H1N1and H1N2, whereas the subtypes H2N2, H3N2, and H5N1 presented different preference over codons (Fig. 1). Interestingly, we observed dissimilar codon preferences within subtypes as well with different genes opting for varied codon choices. For instance, leucine in H1N1 showed as many as four preferred codons in CTA (for HA and PB1), CTT (for M1 and PA), TTG (M2, NA and

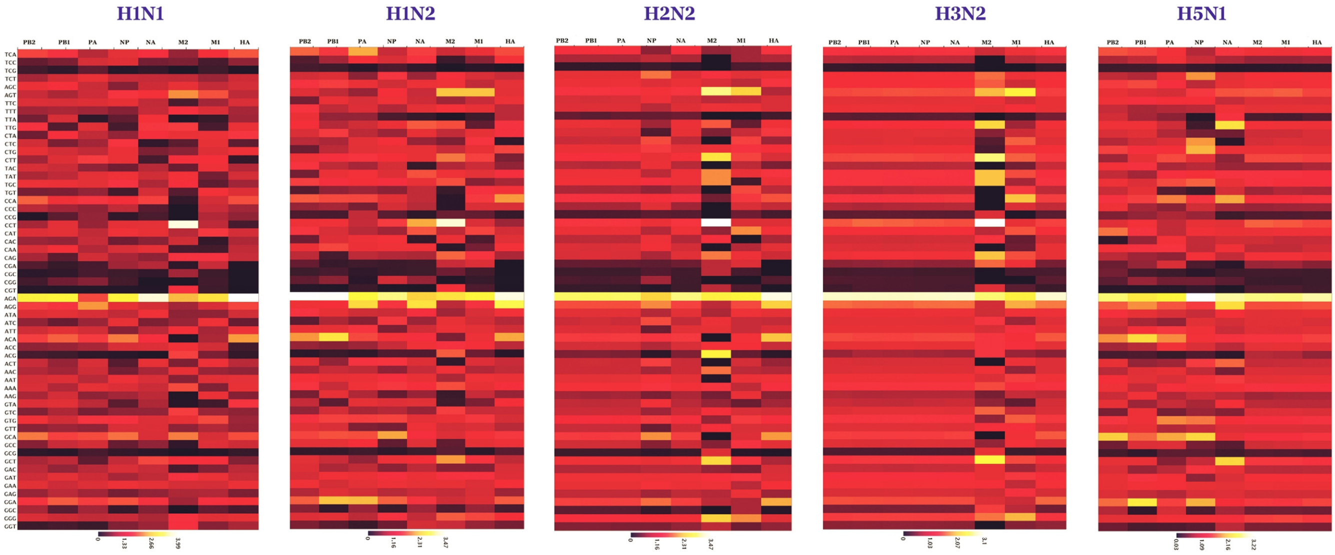


Fig. 1. Heat map of RSCU in five subtypes of IAV.

The darker colored blocks represent a lower magnitude of RSCU, whereas lighter ones represent higher RSCU values. Although some similarities existed between H1N1 and H1N2, the rest showed varied codon preferences.

Table 1. Compositional features of the genes in the IAV subtypes covered in this study.

Subtype	Gene name	GC%	GC3%	Nc	Mononucleotides at synonymous position (%)			
					A3	T3	G3	C3
H1N1	HA	40.9	39.6	47	34.9	25.5	19.2	20.4
	M1	**48.4**	**48.3**	**53**	25.2	26.4	**28.6**	19.7
	M2	43.9	42.5	**54**	23.1	34.4	23.1	19.4
	NA	41.9	40	**52**	29.7	30.3	17.1	23
	NP	46	42.5	**51**	34.7	22.8	21.8	20.7
	PA	44	47.8	**55**	29.7	22.5	25.5	22.3
	PB1	41.9	43.2	**52**	33.6	23.2	23	20.1
	PB2	44.5	45.7	**51**	35	19.3	26.4	19.3
H1N2	HA	41.8	42.6	50	32.8	24.7	20.3	22.2
	M1	**48**	**46.7**	55	26.5	26.8	27.1	19.5
	M2	44.5	45.7	54	22.2	32.1	24.1	21.6
	NA	43.4	42.7	54	27	30.3	19.3	23.4
	NP	46.1	44.7	52	31.3	24	**24.8**	20
	PA	42.4	45.2	51	31.3	23.5	23.9	21.3
	PB1	42.3	43.7	49	33.5	22.8	23.4	20.4
	PB2	42.4	40.2	50	36.7	23.1	22.9	17.3
H2N2	HA	42.2	42.2	49	32.9	24.9	21.8	20.4
	M1	**49.3**	**49.1**	56	24	26.8	28.9	20.2
	M2	44.7	45	49	20.9	34.1	24.5	20.5
	NA	44.1	44.1	53	29.7	26.2	23.5	20.6
	NP	46.5	46.1	54	30.8	23.1	**24.4**	21.7
	PA	44.3	44.5	51	29.9	25.7	23.9	20.6
	PB1	44.4	44.5	50	29.6	25.8	23.9	20.6
	PB2	44.3	44.5	51	29.9	25.7	23.9	20.6
H3N2	HA	44.4	44.5	53	30.1	25.4	23.9	20.7
	M1	**48.2**	**47.5**	56	24.3	28.2	**29.1**	18.4
	M2	45.1	47.5	52	22.3	30.2	22.9	**24.6**
	NA	44.6	44.8	53	29	26.1	24.1	20.7
	NP	44.7	44.8	54	28.7	26.5	24	20.8
	PA	44.8	44.9	51	28.5	26.6	24.2	20.7
	PB1	44.6	44.8	49	28.6	26.5	24.1	20.7
	PB2	44.6	44.8	51	29	26.3	24.1	20.7
H5N1	HA	44.6	44.9	50	28.9	26.2	24.3	20.6
	M1	45.2	45.4	54	27.8	26.8	24.6	20.8
	M2	45	45.2	44	28	26.8	24.3	20.8
	NA	44	42.2	51	26.7	31.1	21.2	21
	NP	**47.7**	**47.3**	51	29.9	22.8	**26.3**	21
	PA	43.6	46.3	53	31.8	21.9	**23.8**	22.5
	PB1	43.5	46.4	52	31.6	22	**25**	21.4
	PB2	44.8	45.4	53	33.8	20.8	**26.5**	18.8

The values in boldface indicate deviations in magnitude as compared with the values from the rest of the members in the concerned group.

PB2), and CTC (for NP). A similar pattern was observed for other subtypes as well. Taken as a whole, AGA (arg), CCT (pro), ACA (thr), and AGT (ser) were some of the overwhelmingly favored codons. The 787 cds were examined for the rare codon analysis using Anaconda 2.0 software [22]. The codons of the make-up CGN and NCG were severely depleted. The codons CGC, TCG, and CGT were rarely used in all the subtypes, whereas some others like CCG, ACG, CGA, CGG and GCG were also suppressed to a great extent, albeit non-uniformly across the subtypes (Supplementary File S2).

It is envisaged that the preference for a specific codon to encode the amino acids has liaison with the expressivity of the gene [31]. To fish out such biasness and to execute a predictive estimation of gene expression, the CAI value was calculated for each gene. The CAI values had a mean of 0.83 and a standard deviation of 0.154. Interestingly, the M2 gene in each subtype showed a sharp decline in expressivity reflected by the mean CAI value of 0.48 ± 0.15. Ironically, H5N1 again proved to be anomalous with M2 expressivity (mean CAI of 0.77), catching up with the CAI value of the rest of the genes.

Dinucleotide Analysis and CpG Usage

We analyzed the enrolled cds for dinucleotide usage, which clearly suggested a severe diminution of dinucleotide CpG. The odds ratio values revealed that TpG with a mean odds ratio value of 1.45 ± 0.08 was the most over-represented dinucleotide, whereas CpG (0.53 ± 0.15) was the most under-represented one. The dinucleotides TpC, CpA, CpT, and GpA were also represented in elevated magnitude (mean odds ratio>1.23) as compared with the rest. GpT and TpA were among the under-represented (mean odds ratio < 0.78) ones following CpG. Nevertheless, this was an overall observation; hence, it did not show absolute uniformity *per se*, with slight variations among some of the representative genes.

Role of Compositional Constraint in Codon Usage in the IAV Genes

A plot of the average GC content of the first two codon positions (GC12) along the ordinates and GC3 along the abscissa, popularly known as the neutrality plot, has been widely utilized as an indicator of possible interplay of mutation and selection equilibrium in CUB. It has been postulated that a statistically significant correlation and a

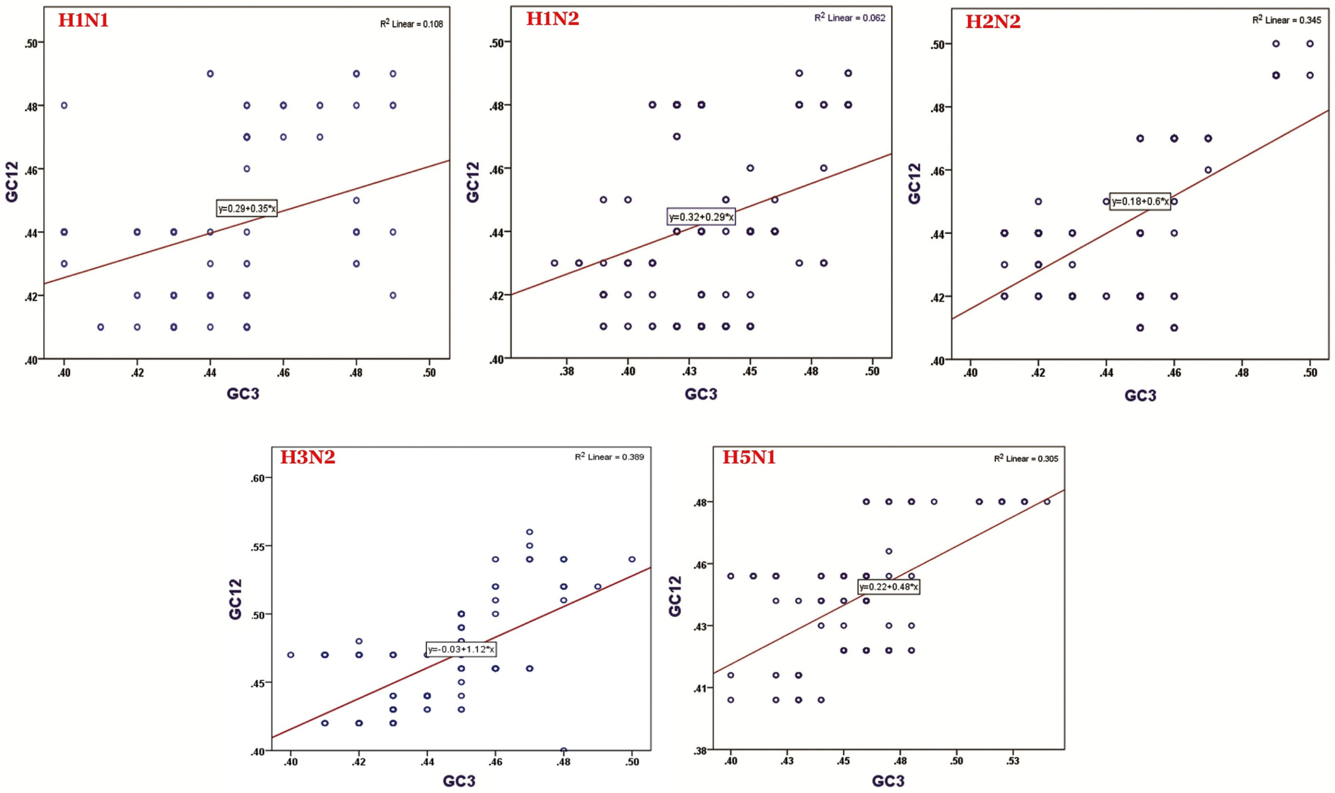


Fig. 2. Neutrality analysis for the genes in five IAV subtypes.

All the subtypes showed similar trends with varying magnitude, whereas H3N2 deviated slightly by exhibiting a different regression fit with a slope that increased at increasing rate.

Table 2. Regression curves of neutrality analysis (GC12 vs GC3).

IAV subtype	Regression line	R^2	*p*-Value	Slope
H1N1	y = 0.29+0.35x	0.108	<0.001	0.35
H1N2	y = 0.32+0.29x	0.082	0.001	0.29
H2N2	y = 0.18+0.6x	0.345	<0.001	0.59
H3N2	y = -0.03+1.12x	0.389	<0.001	1.12
H5N1	y = 0.22+0.48x	0.305	<0.001	0.48

regression line with a slope close to unity are indicative of mutation pressure being the prime evolutionary force; otherwise, selection against mutation is said to be operative in the case of weak correlation between the same [33]. A slope below 1 in the regression line would point to a tendency of non-neutral mutational pressure. To reveal any links amid the three codon positions, we constructed neutrality plots (GC12 vs. GC3) for each of the IAV subtypes (Fig. 2). We found statistically significant positive correlations between GC12 and GC3 in all the cases (Table 2). The slopes, however, showed differential magnitude, with visible deviation in the case of H3N2, where the slope was increasing at a growing rate unlike the rest. The results hint towards the possible role of mutational pressure inflicting CUB in the IAV genes.

PR2 Bias Plot Analysis

To inspect whether the unequal codon choices are limited to the genes with higher degree of bias, we employed a Parity Rule 2 (PR2) bias plot and examined the alliance between purines (A and G) and pyrimidines (C and T). For convenience of our analysis, we left out the three stop codons, codons for Met and Trp, and also the ATA codon of Ile. In PR2 analysis, at the mid-junction where both coordinates are 0.5, A becomes equal to T while G equals C (PR2), if there exists no substitution bias between the two complementary DNA strands [4]. All the subtypes showed a little bias. It appears from the allocation

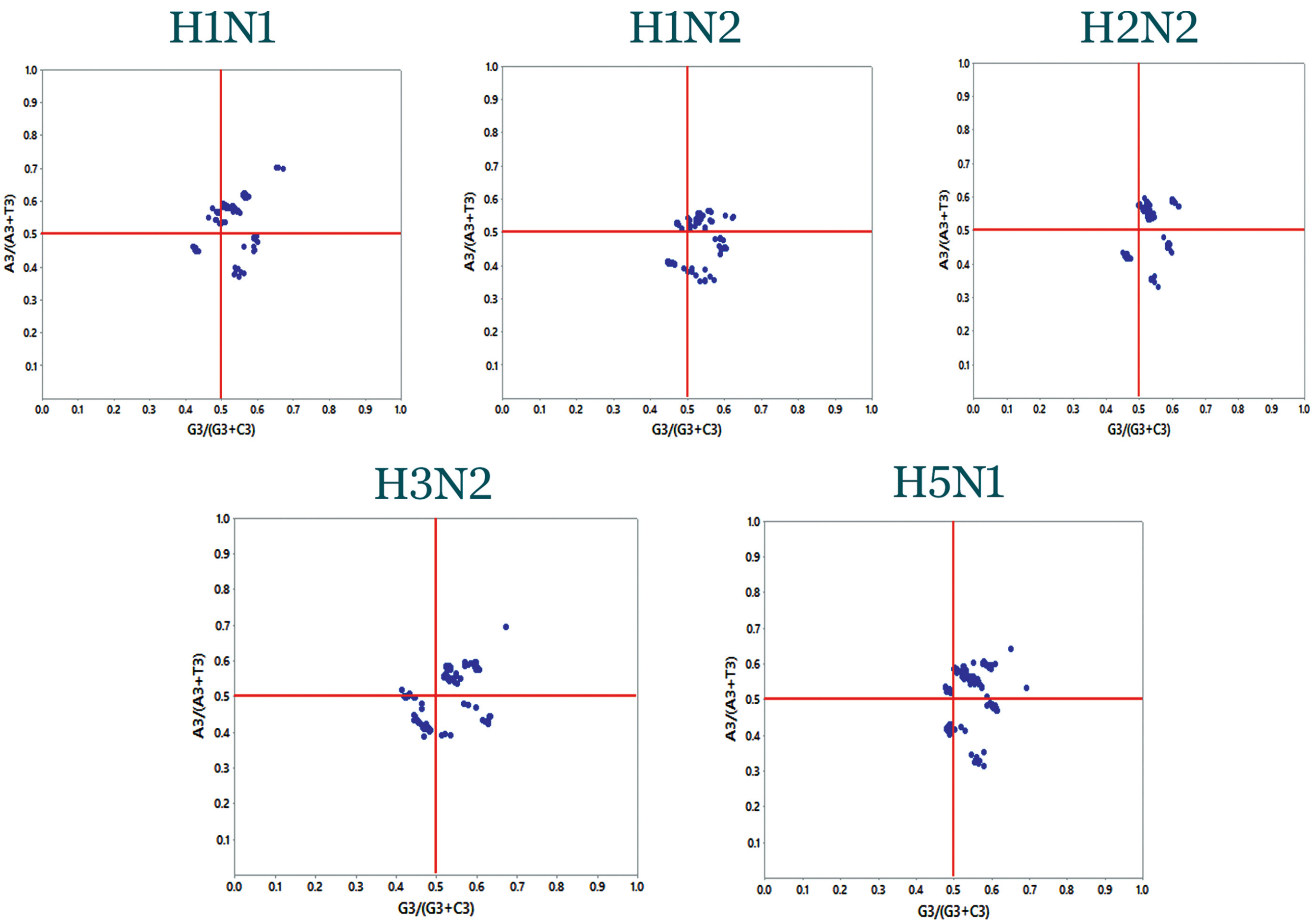


Fig. 3. PR2 analysis for the genes used in the study from five IAV subtypes.

of the points close to the midpoint in the plot that there exists only a meek PR2 bias in A3 and G3 (Fig. 3). However, the purines (A and G) seem to be used more frequently than the pyrimidines (C and T) at the synonymous sites, especially in the case of H2N2 and H5N1.

Trends of Codon Usage Variation in IAV

Correspondence analysis is a multivariate ordination technique used far and wide for its immensely effective way of reducing high-dimensional data in planar form [13]. The CA shows the allocation of genes based on their corresponding choice of codons, which helps uncover the latent influence on CUB. To resolve the trend in codon usage variation in the IAV genes, we executed CA on RSCU values, where all gene data were examined as a single dataset and the two major axes were put on view in a two-dimensional scatterplot. The first two major axes could account for 55.7% of the total variations with individual contributions of 44.5% and 11.3% by axis 1 and axis 2, respectively. The distribution of the genes in the CA plot showed the presence of at least three clusters marked in circles (Fig. 4). Genes M1 and M2 for all subtypes, except

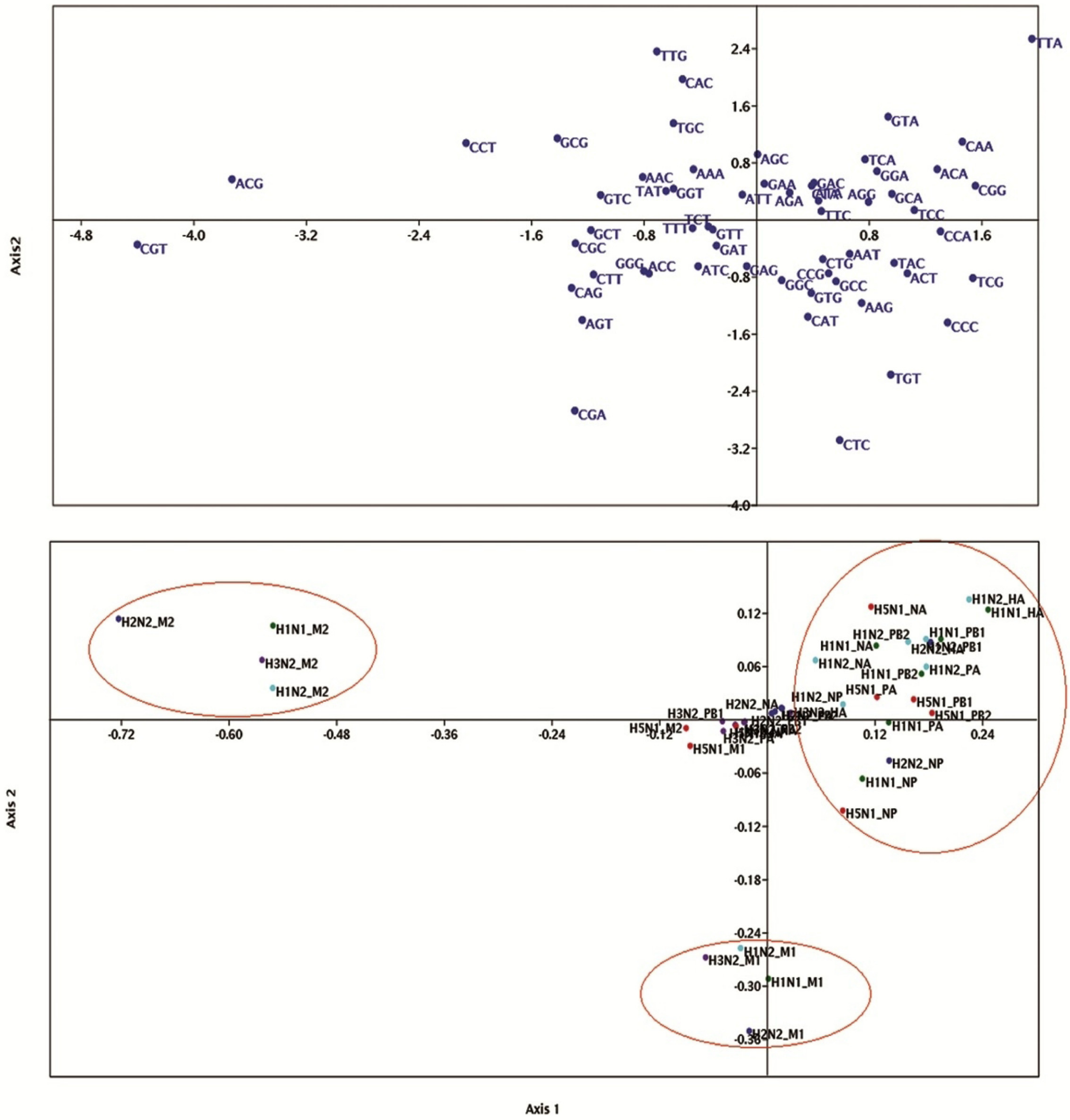


Fig. 4. Correspondence analysis on RSCU values in the IAV genes considered for the study.

The upper panel shows the distribution of the codons based on their preference by the genes, and the lower panel depicts the allocation of the genes in the five subtypes.

H5N1, clustered separately from the rest of the genes.

We also performed cluster analysis in Past3 [16] using UPGMA algorithm and taking the Euclidean similarity index. The results reiterate the findings with CA with three major clusters (Fig. 5). Here also, we noticed deviation of the H5N1 subtype from the rest. For instance, the M1 and M2 genes of all the subtypes except H5N1 formed a separate cluster whereas the same genes of the latter were seen clustering with PA, PB1, HA, NA, etc. of the rest. Three major groups were found among the IAV genes belonging to the five aforesaid subtypes. Taken as a whole, H2N2 and H3N2 showed a close similarity, whereas H5N1 turned out to be the most deviant one.

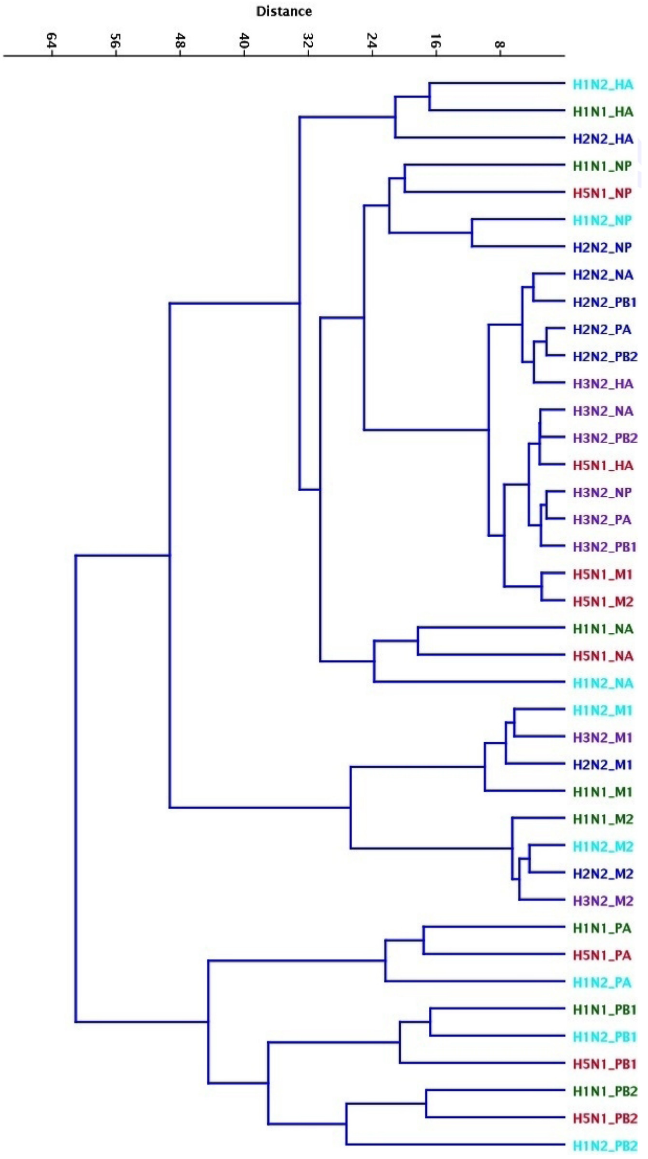


Fig. 5. Cluster analysis of the selected gene groups.
UPGMA algorithm and the Euclidean similarity index were used for constructing the dendrogram. H5N1 showed some deviation in a few occasions whereas H2N2 and H3N2 showed the closest resemblance.

Codon-Pair Context Analysis

An important but not very much extensively studied aspect in CUB studies is the codon-pair context in the genes. At the translation level, codon usage and codon-pair context are prone to selective forces, given that they have roles to play in the speed and accuracy of the mRNA decoding fidelity [3, 25]. Here, in a quest for the underlying codon-context, Anaconda 2.0 was used to compare codon pair associations with the help of a 64 × 64 codon-pair contingency table [23]. As per our findings, the individual contexts showed variations across the IAV subtypes.

The matrix plot of 5′context, considering all the genes as a whole, showed clusters of good as well as bad contexts, as can be seen marked in yellow and blue circles (Fig. 6). There were varying preferences over contexts of different make-up in the genes enrolled for the study. Contexts of the make-up NNC-GNN and NNT-ANN were severely depleted. Amino acid pairs like Arg-Gly, Glu-Lys, Ser-Gly, Ser-Ser are some of the most preferentially used contexts across the subtypes; however, these contexts did not occur at similar magnitudes. We also compared the subtypes against each other for codon context patterns, which presented more or less similar patterns in all the cases.

Discussion

This study amasses the codon usage profiles of five human influenza A viral subtypes covering the genes of the IAV genome encoding eight major proteins. Our findings point to a weak CUB in these genes as can be understood by higher Nc (>40) values. This observation is, however, not unique, as many authors have previously reported lower codon bias in IAV [1, 39]. In fact, a lower CUB has been found in many RNA viruses [17, 35]. Jenkins and Holmes [17] had reported an average Nc value of 50.9 in human RNA viruses-including IAV.

Dinucleotide biases may influence the codon usage patterns, and are reported in many viruses [18]. The RSCU and the dinucleotide analysis reveal a preference of A/T ending codons and remarkable suppression of codons having the dinucleotide CpG. The remarkable avoidance towards codons with dinucleotide CpG re-establishes the previous finding of low CpG usage in this single-stranded RNA virus. This CpG depletion is linked to its role for divergent evolutionary pressure and has been reported in

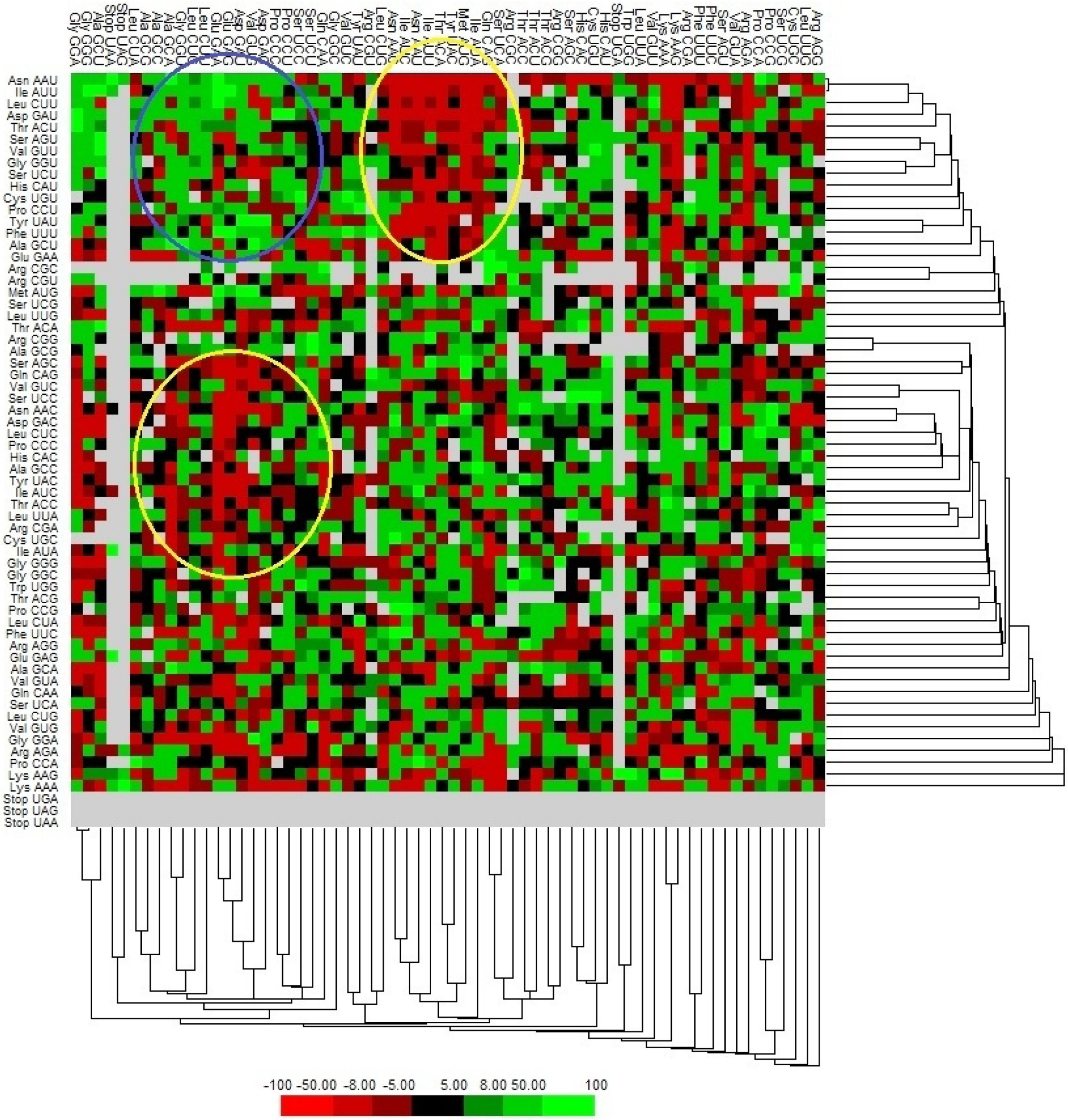


Fig. 6. Codon context analysis in IAV genes enrolled for the study.

The matrix plot represents 5′ context taking all the 787 genes as a whole. The blue and yellow circles depict good and bad contexts, respectively.

many RNA viruses in previous studies [28, 38]. The IAV strains evolving in avian hosts subsequent to the 1918 pandemic were believed to be selected under strong selection pressure to trim down their CpG content [9, 35]. The CpG shortage was also anticipated to have acquaintances with the immune response as unmethylated CpG is utilized as pathogen markers by the innate immune system of the host [15, 18, 30].

Considering the overall amino acid usage, there was not much variation among the IAV subtypes. Leucine was clearly the most favored amino acid (8.4%) followed by serine (7.8%) and glutamine (7.7%); whereas tryptophan (1.5%), histidine (1.9%), and cysteine (2.0%) were among the least abundant amino acids. Among the individual genes, we observed a little deviation in the form of the M1 gene, where alanine was the most preferred amino acid with an increased 10.5% of usage. Interestingly, H5N1 did not follow this trend with the alanine usage percentage of 6.6.

The deviation of H5N1 from the rest, however, is not limited to amino acid usage only. We observed preference of A/G at the third position in the case of H5N1, whereas the rest of the subtypes preferred A/T. There was difference in GC content as well for H5N1 as discussed in the Results section. This observation contradicts the reports of Zhou *et al*. [39] where they had not found any striking difference between the IAV subtypes. The reason behind this striking difference might be linked to the fact that, unlike the other subtypes, H5N1 is primarily caused by zoonosis and had crossed the avian-human species barrier only recently in 1997 [20]. Being primarily a poultry disease, its genetic setup is more adapted to the avian hosts, whereas the rest have been co-evolving with the human hosts for a longer span. Nevertheless, there have been reports of human-to-human transmission as well [34]. Correspondence analysis and cluster analysis on RSCU values represented the deviation of H5N1 from the rest. Here, the M1 and M2 genes of all subtypes were seen forming two separate clusters, leaving aside H5N1. Codon context analysis did not offer much variation among the different subtypes. However, rare codon analysis yet again showed a slightly different picture in H5N1. Codons CCG and CGG, which were rare in all other subtypes, were found well above the threshold line (Supplementary File S2).

The general relationship between base composition and codon usage mutational pressure is more pronounced than other selective forces. Neutrality analysis and PR2 bias analysis go in accordance with this observation. With the tremendous size of the RNA virus population, it appears unusual, but the effect of mutational pressure is too overwhelming for the effect of selection to make a mark [17]. However, there could be other factors responsible for the variations occurring in the IAV codon usage profile.

To summurize, this particular work highlights the codon usage profiles of five IAV strains infecting humans. Our findings suggest a low CUB in the IAV. The codon usage analysis using RSCU estimations of the codons presented a complicated picture with the varying preference of codons in different genes and different subtypes. However, these AT-rich genes showed an inclination towards A/T ending codons in most cases. Strikingly, the H5N1 subtype presented slight variation from the rest in the preference of A/G at the third codon position as well as amino acid usage. The variation of H5N1 from the rest is also supported by the correspondence analysis and cluster analysis. We found a significant positive correlation between GC12 and GC3, which implied GC composition as a crucial factor in shaping the codon usage in this virus. It gives the idea that a balance of mutation/selection exists in IAV, which permits it to re-adjust its codon usage to different conditions. More far-reaching examination concerning the CUB profile might help in a better comprehension of the various aspects of the virulence factors leading to the identification of suitable drug targets, which in turn would pave the way for the development of successful antivirals in combating the virus.

Acknowledgments

The authors are grateful to the University Grants Commission, New Delhi, India, for providing a UGC-BSR Fellowship to carry out this research work. We are also grateful to Assam University, Silchar-788011, Assam, India, for providing the research facility. We sincerely acknowledge the valuable inputs from Mr. Arup Kr. Malakar and Mr. Prosenjit Paul, Research Scholars, Department of Biotechnology, Assam University, while analyzing the data and in preparation of the manuscript.

References

1. Ahn I, Son HS. 2010. Comparative study of the nucleotide bias between the novel H1N1 and H5N1 subtypes of influenza A viruses using bioinformatics techniques. *J. Microbiol. Biotechnol.* **20:** 63-70.
2. Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136:** 927-935.
3. Boycheva S, Chkodrov G, Ivanov I. 2003. Codon pairs in the genome of *Escherichia coli*. *Bioinformatics* **19:** 987-998.
4. Chen Y. 2013. A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection. *Biomed. Res. Int.* **2013:** 406342.
5. Cheng X, Virk N, Chen W, Ji S, Sun Y, Wu X. 2013. CpG usage in RNA viruses: data and hypotheses. *PLoS One* **8:** e74109.
6. Chiapello H, Ollivier E, Landes-Devauchelle C, Nitschke P, Risler JL. 1999. Codon usage as a tool to predict the cellular

location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases. *Nucleic Acids Res.* **27:** 2848-2851.

7. Comeron JM, Aguade M. 1998. An evaluation of measures of synonymous codon usage bias. *J. Mol. Evol.* **47:** 268-274.
8. Gill PW. 1971. Hong Kong 68 variant of influenza A2. *Br. Med. J.* **3:** 308.
9. Goni N, Iriarte A, Comas V, Sonora M, Moreno P, Moratorio G, *et al.* 2012. Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development. *Virol. J.* **9:** 263.
10. Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* **10:** 7055-7074.
11. Gramer MR, Lee JH, Choi YK, Goyal SM, Joo HS. 2007. Serologic and genetic characterization of North American H3N2 swine influenza A viruses. *Can. J. Vet. Res.* **71:** 201-206.
12. Grantham R, Gautier C, Gouy M, Mercier R, Pave A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8:** r49-r62.
13. Greenacre M, Hastie T. 1987. The geometric interpretation of correspondence analysis. *J. Am. Stat. Assoc.* **82:** 437-447.
14. Greenacre M, Vrba E. 1984. Graphical display and interpretation of antelope census data in African wildlife areas, using correspondence analysis. *Ecology* **65:** 984-997.
15. Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog.* **4:** e1000079.
16. Hammer Ø, Harper D, Ryan P. 2001. PAST – PAlaeontological STatistics, ver. 1.89. *Palaeontol. Electron.* **4:** 1-9.
17. Jenkins GM, Holmes EC. 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.* **92:** 1-7.
18. Karlin S, Doerfler W, Cardon LR. 1994. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *J. Virol.* **68:** 2889-2897.
19. Karlin S, Mrazek J. 1997. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* **94:** 10227-10232.
20. Korteweg C, Gu J. 2008. Pathology, molecular biology, and pathogenesis of avian influenza A (H5N1) infection in humans. *Am. J. Pathol.* **172:** 1155-1170.
21. Moriyama EN, Powell JR. 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res.* **26:** 3188-3193.
22. Moura G, Pinheiro M, Arrais J, Gomes AC, Carreto L, Freitas A, *et al.* 2007. Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLoS One* **2:** e847.
23. Moura G, Pinheiro M, Silva R, Miranda I, Afreixo V, Dias G, *et al.* 2005. Comparative context analysis of codon pairs on an ORFeome scale. *Genome Biol.* **6:** R28.
24. Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. *Mol. Biol. Evol.* **19:** 1390-1394.
25. Ogle JM, Ramakrishnan V. 2005. Structural insights into translational fidelity. *Annu. Rev. Biochem.* **74:** 129-177.
26. Osawa S, Ohama T, Yamao F, Muto A, Jukes TH, Ozeki H, Umesono K. 1988. Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. *Proc. Natl. Acad. Sci. USA* **85:** 1124-1128.
27. Post LE, Nomura M. 1980. DNA sequences from the *str* operon of *Escherichia coli*. *J. Biol. Chem.* **255:** 4660-4666.
28. Rabadan R, Levine AJ, Robins H. 2006. Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. *J. Virol.* **80:** 11887-11891.
29. Romero H, Zavala A, Musto H. 2000. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res.* **28:** 2084-2090.
30. Shackelton LA, Parrish CR, Holmes EC. 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *J. Mol. Evol.* **62:** 551-563.
31. Sharp PM, Li WH. 1987. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15:** 1281-1295.
32. Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, *et al.* 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459:** 1122-1125.
33. Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **85:** 2653-2657.
34. Ungchusak K, Auewarakul P, Dowell SF, Kitphati R, Auwanit W, Puthavathana P, *et al.* 2005. Probable person-to-person transmission of avian influenza A (H5N1). *N. Engl. J. Med.* **352:** 333-340.
35. Wong EH, Smith DK, Rabadan R, Peiris M, Poon LL. 2010. Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. *BMC Evol. Biol.* **10:** 253.
36. Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* **87:** 23-29.
37. Xiang H, Zhang R, Butler RR 3rd, Liu T, Zhang L, Pombert JF, Zhou Z. 2015. Comparative analysis of codon usage bias patterns in microsporidian genomes. *PLoS One* **10:** e0129223.
38. Zhong J, Li Y, Zhao S, Liu S, Zhang Z. 2007. Mutation pressure shapes codon usage in the GC-rich genome of foot-and-mouth disease virus. *Virus Genes* **35:** 767-776.
39. Zhou T, Gu W, Ma J, Sun X, Lu Z. 2005. Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. *Biosystems* **81:** 77-86.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2014, Article ID 349139, 7 pages
http://dx.doi.org/10.1155/2014/349139

Research Article

Compositional Constraint Is the Key Force in Shaping Codon Usage Bias in Hemagglutinin Gene in H1N1 Subtype of Influenza *A* Virus

Himangshu Deka and Supriyo Chakraborty

Department of Biotechnology, Assam University, Silchar, Assam 788011, India

Correspondence should be addressed to Supriyo Chakraborty; supriyoch_2008@rediffmail.com

Received 20 May 2014; Accepted 8 July 2014; Published 17 July 2014

Academic Editor: Ian Dunham

Copyright © 2014 H. Deka and S. Chakraborty. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

It is vital to unravel the codon usage bias in order to gain insights into the evolutionary forces dictating the viral evolution process. Influenza *A* virus has attracted attention of many investigators over the years due to high mutation rate and being cross-specific shift operational in the viral genome. Several authors have reported that the codon usage bias is low in influenza *A* viruses, citing mutational pressure as the decisive force shaping up the codon usage in these viruses. In this study, complete coding sequences of hemagglutinin genes for H1N1 subtype of influenza *A* virus have been explored for the possible codon usage bias acting upon these genes. The results indicate overall low bias with peaking ENC values. The GC content is found to be substantially low as against AT content in the silent codon sites. Significant correlations were observed in between the compositional parameters versus AT_3, implying the possible role of the latter in shaping codon usage profile in the viral hemagglutinin. The data showed conspicuously that the sequences were *A* redundant with most codons preferring nucleotide *A* over others in the third synonymous codon site. The results indicated the pivotal role of compositional pressure affecting codon usage in this virus.

1. Background

Influenza *A* virus (IAV), a member of *Orthomyxoviridae,* remains a serious health concern on a global basis with a number of epidemics since early 19th century till date. With several variants of varying pathogenic profile, IAV is causing significant mortality every year throughout the globe. In the year 2009, the world has seen its only second global pandemic, an H1N1 pandemic which was declared as phase 6 alert level by the World Health Organization (WHO). It was the first of its kind since 1968 when Hong Kong flu was declared a global pandemic by the WHO. Reports say that about 214 countries have been affected by the pandemic influenza H1N1 of 2009 taking 18,138 lives, as updated in May 2010 (http://www.who.int/csr/don/2010_06_04/en/index.html).

What makes influenza *A* such a deadly virus? Generally, upon exposure to a pathogen, the host develops specific immunity against it, thus, preventing the same pathogen infecting for a second time. The IAV escapes the specific immunity of the host by a process termed as antigenic drift. This is achieved by frequent mutation in the hemagglutinin (HA) and neuraminidase (NA) genes which encode the main antigenic determinant proteins in the virus, due to which immunogenically distinct strains develop which cause the seasonal outbreaks [1]. Another process, differently termed by different authors as cross-specific shift [2] or reassortment [3], is responsible for the frequent changes in the antigenic region of the virus, as happened in case of 2009 H1N1 pandemic. The viral HA or NA or other gene segments of different subtype of IAV are exchanged resulting into a novel subtype of IAV. These two genes, HA in particular, provide virulence to the virus making it as a potential drug target for the prevention of the spread of influenza infection [1].

The degeneracy of the genetic code has rendered the privilege of using more than one codon to code for the same amino acid. The phenomenon is called synonymous use of codons. The use of synonymous codons, however, is not uniform in different species ranging from prokaryotes to complex organisms as well as in viruses; certain synonymous codons are used preferentially. This tilted use of codons is termed as codon usage bias (CUB). With the rapidly growing stockpile of sequences in public databases after whole genome sequencing of large number of species, investigators have engaged in research in the context of codon usage bias in specific genes as well as whole genome of a vast range of organisms [4–7].

The preferential use of synonymous codons is governed by different evolutionary forces [8]. Over the years many authors have reported a number of measures to assess codon usage bias across genes and genomes. Among these measures, GC content, relative synonymous codon usages (RSCU), and effective number of codons (ENC), are some of the most widely used parameters for codon bias study. Much has been debated regarding the inclination towards the selection of optimal codons in genes; many advocated increased efficiency of translation process as the main reason behind selection of optimal codons [9]. However, the exact mechanisms behind synonymous codon variation are yet to be understood clearly.

Several workers have reported that the overall codon usage bias in RNA viruses is low, which is attributed to GC compositional properties and dinucleotide content in these viruses [5, 10–12]. Mutational bias has been projected as the main factor that drives the codon usage variation among the influenza *A* viruses which are phylogenetically conserved [10, 12, 13].

2. Materials and Methods

2.1. Datasets. In this study, a total of 32 complete coding sequences of the hemagglutinin (HA) gene of human-host derived influenza *A* virus subtype H1N1 reported from India were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/). The serial numbers (SN), accession numbers, and other information are presented in supplementary Table 1 available online at http://dx.doi.org/10.1155/2014/349139.

2.2. Parameters for Codon Usage Bias Study. Relative synonymous codon usage (RSCU) [14] is one of the most widely used parameters for querying the pattern of synonymous codon usage across genes and genomes without confounding influence of the amino acid composition. To examine the synonymous codon usage in the genes, RSCU values were calculated. RSCU is defined as the ratio of the observed frequency to the expected frequency if all the synonymous codons for those amino acids are used equally. If the RSCU value of a codon is more than 1.0, it is said to have a positive codon usage bias, while a value of less than 1.0 means a negative codon usage bias. When the RSCU value is close to 1.0, it means that this codon is chosen randomly and equally with other synonymous codons.

The effective number of codons (ENC) estimates the enormity of codon usage bias in a gene [15]. ENC is estimated to quantify the synonymous codon usage across the target sequence which is calculated as given below:

$$\mathrm{ENC} = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}, \tag{1}$$

where, F_k ($k = 2, 3, 4$ or6) is the average of the F_k values for k-fold degenerate amino acids. The F value denotes the probability that two randomly chosen codons for an amino acid with two codons are identical. The values of ENC range from 20 (when only one codon is used per amino acid) to 61 (when all synonymous codons are equally used for each amino acid) [15–17]. The codon bias is considered low if the ENC value is greater than 40.

Nucleotide composition plays a crucial role in the codon usage pattern in the genes because most of the indices of codon usage bias are based on the base composition of the genes. GC_3 is the frequency of the nucleotides G+C at the synonymous 3rd positions of the codons excluding the *Met*, *Trp*, and the termination codons. Similarly, GC_{1s} and GC_{2s} represent G+C frequency at 1st and 2nd codon positions. GC_{3s} is a good indicator of the extent of base composition bias.

Gene expressivity was measured by codon adaptation index (CAI) as given by Sharp and Li [14]. CAI has been used as a simple and effective parameter to measure the adaptiveness of synonymous codon usage bias of a gene towards the codon usage of highly expressed genes. CAI, with the boundary values 0-1, was originally proposed to provide a normalized estimate that can be used across genes and species. A value of 1 is assigned to the most frequent codons within a gene (CAI = 1) while the least frequent codons are assigned a CAI value of 0 [18, 19]. CAI is estimated as

$$\mathrm{CAI} = \exp \frac{1}{L} \sum_{k=1}^{L} \ln w_{c(k)}, \tag{2}$$

where L is the number of codons in the gene and $w_{c(k)}$ is the ω value for the kth codon in the gene.

Frequency of optimal codon (Fop), originally proposed by Ikemura in the year 1981, is one of the first estimators used in the study of codon usage bias. As an index, Fop shows the optimization level of synonymous codon choice in each gene to translation process [8]. Fop is defined as the ratio of total number of optimal codons in a gene to the total number of synonymous as well as nonsynonymous codons in that gene.

The codon usage bias measures, namely, RSCU, ENC, GCs, Fop, and CAI for each coding sequence, were estimated in our study by using an in-house Perl program developed by SC.

3. Results and Discussion

3.1. Nucleotide Compositional Properties. The coding sequences were analyzed thoroughly for their nucleotide composition. Individual nucleotides as well as GC and AT content in three synonymous codon positions were estimated. The nucleotide composition in the analyzed genes is summarised in Table 1. The results reveal that the viral hemagglutinin is *A* redundant with overall *A* content of 35.3% with a range of 34.9% to 35.6% and standard deviation (SD) of 0.167. On the other hand, the *C* content in all the accessions is consistently low ranging from 18.2% to 18.8% with average and SD of 18.5 and 0.145, respectively.

The frequency of codons containing dinucleotide TpA is much higher in comparison to those containing dinucleotide

Table 1: Nucleotide composition of the genes used in the study.

Sl No.	A%	T%	G%	C%	A_3%	T_3%	G_3%	C_3%	GC%	GC_3%	AT%	AT_3%	ENC
1	35.4	24.2	22	18.3	36.4	27	16.6	20	40.3	38.4	59.7	61.6	57
2	35.2	23.9	22.3	18.6	35.5	26	17.5	20.9	40.9	40.2	59.1	59.8	58
3	35.4	24.2	22.2	18.2	36.6	27	16.4	20	40.4	38.3	59.6	61.7	57
4	35.2	23.9	22.3	18.6	35.3	26	17.7	20.9	40.9	40.4	59.1	59.6	58
5	35	24	22.4	18.6	35.3	26	17.7	20.9	41	40.4	59	59.6	58
6	35.3	24	22.2	18.4	36.5	26.6	16.4	20.4	40.6	38.8	59.4	61.2	58
7	35.3	23.9	22.3	18.5	35.7	26	17.3	20.9	40.8	40	59.2	60	58
8	35.4	23.9	22.2	18.5	35.5	26.2	17.5	20.8	40.7	40	59.3	60	58
9	35.4	24.2	22	18.3	36.7	27	16.2	20.1	40.4	38.3	59.6	61.7	57
10	35.4	24.2	22.2	18.2	36.7	27	16.2	20.1	40.4	38.3	59.6	61.7	57
11	35.6	24.2	21.9	18.4	37.2	26.6	15.8	20.4	40.3	38.6	59.7	61.9	57
12	35.3	23.9	22	18.8	36.1	25.9	16.6	21.4	40.8	39.9	59.2	60.1	58
13	35.2	23.8	22.2	18.8	35.9	25.5	17.1	21.5	41	40.4	59	59.6	58
14	35.3	24	22.2	18.4	36.5	26.6	16.2	20.6	40.6	39	59.4	61.2	58
15	35.4	24	22.2	18.5	36.4	26.1	16.8	20.7	40.6	38.8	59.4	61	58
16	34.9	24.1	22.4	18.6	36	26.1	17.1	20.8	41	39.2	59	60	58
17	35.3	24	22.1	18.5	36.7	26.2	16.2	20.9	40.7	38.7	59.3	60.9	58
18	35.3	24.2	22.1	18.4	36.4	26.4	16.6	20.7	40.6	38.7	59.4	60.8	58
19	35.1	24.1	22.4	18.4	36.4	26.5	16.5	20.5	40.8	38.8	59.2	61	58
20	35	24.2	22.2	18.6	36.4	25.9	16.7	21	40.8	38.6	59.2	60.2	58
21	35.3	24.1	22.2	18.5	36.4	26.5	16.4	20.7	40.6	38.7	59.4	61	58
22	35.1	24.1	22.4	18.3	36.2	26.7	16.7	20.4	40.7	38.8	59.3	61	58
23	35.1	24.2	22	18.6	36.3	26.3	16.3	21.3	40.7	38.7	59.3	60.4	58
24	35.3	24.1	22.2	18.5	36.4	26.4	16.5	20.7	40.6	38.5	59.4	60.8	58
25	35.2	24	22.2	18.5	36.3	26.3	16.3	21.1	40.8	38.6	59.2	60.7	58
26	35.2	24	22.3	18.5	36	26.4	16.7	20.9	40.8	38.7	59.2	60.5	58
27	34.9	24	22.3	18.7	36.3	25.9	16.8	20.9	41	39	59	60.2	58
28	35.3	24	22.3	18.5	36.4	26.4	16.5	20.7	40.7	38.6	59.3	60.8	58
29	35	24.2	22.2	18.6	36.6	26.1	16.4	20.8	40.8	39	59.2	60.4	58
30	35.5	24	22.2	18.3	37	26.3	16.3	20.3	40.4	38.6	59.6	61.4	58
31	35.3	24	22.2	18.5	36.6	26.2	16.6	20.6	40.7	39.2	59.3	61.8	58
32	35.4	23.8	22.1	18.7	36.4	26	16.9	20.7	40.8	38.9	59.2	60.7	58

CpG. Four codons, that is, CGA, CGC, CGG, and CGT, out of possible nine codons containing CpG, are absent in the analyzed gene; the frequencies of the remaining codons are also very low with the highest value of 9 for GCC. In contrast, most of the codons (5 out of 6) containing TpA showed higher frequency with the highest value of 17 for GTA and the lowest 6 for TTA. While three codons containing TpA are preferred, there are no preferential codons containing CpG.

The overall GC content in the dataset was found to be much lower in comparison to overall AT content (40.7% and 59.3%, resp.). The suppression of GC content as compared to AT content is also evident from GC/AT content at the silent position. The overall GC_3 was found to be low (39.0%) as against AT_3 (60.7%) (Figure 1). To detect any possible relation of base composition at different synonymous codon positions, the estimated values of the four nucleotides A, T, G, and C and the AT and GC content were compared with the values of the nucleotides in third synonymous positions (i.e., A_3, T_3, G_3, and C_3). The results indicate a strongly significant and complicated correlation which is presented in Table 2. The correlation coefficients were highly significant in majority of the parameters taking both positive and negative values except a few showing insignificant correlation. Negative correlation was also observed between GC_{1+2} and GC_3 ($r = -0.478$, $P < 0.001$). The correlation results indicate the possible role of mutational pressure acting on these genes. The base composition was most likely influenced by AT_3 as revealed by the highly significant correlation coefficients.

Previous studies have revealed that the CpG underrepresentation is attributable to immunologic escape, in order to avoid host immune system using the unmethylated CpGs as a pathogen marker [20, 21]. CpG deficiency has also been reported in some other RNA viruses as well [10, 20, 22]. Thus, combating the host immune response may constitute a selection pressure in these viruses.

The general trend of the ENC values suggests the absence of strong codon bias in the hemagglutinin gene. The ENC values were consistently found in higher range with an average

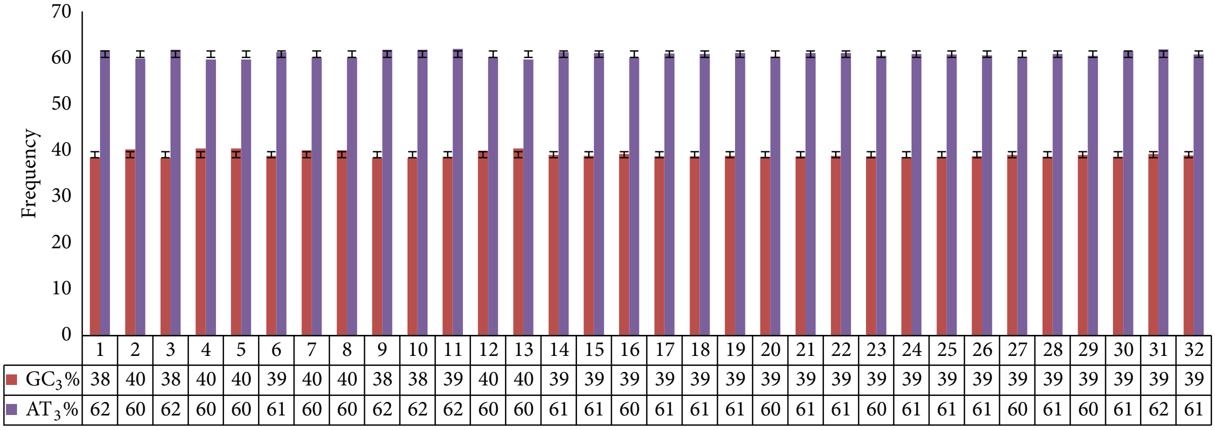
Table 2: Correlation between different nucleotide compositional parameters.

	A_3%	T_3%	G_3%	C_3%	GC_3%	AT_3%
A%	$r = 0.425^*$	$r = 0.444^*$	$r = -0.366^*$	$r = -0.471^{**}$	$r = -0.264^\wedge$	$r = 0.613^{**}$
T%	$r = 0.539^{**}$	$r = 0.653^{**}$	$r = -0.577^{**}$	$r = 0.512^{**}$	$r = -0.695^{**}$	$r = 0.537^{**}$
G%	$r = -0.515^{**}$	$r = -0.241^\wedge$	$r = 0.542^{**}$	$r = 0.089^\wedge$	$r = 0.329^\wedge$	$r = -0.426^{**}$
C%	$r = -0.478^{**}$	$r = -0.899^{**}$	$r = 0.468^{**}$	$r = 0.883^{**}$	$r = 0.618^{**}$	$r = -0.776^{**}$
GC%	$r = -0.693^{**}$	$r = -0.810^{**}$	$r = 0.664^{**}$	$r = 0.767^{**}$	$r = 0.669^{**}$	$r = -0.886^{**}$
AT%	$r = 0.693^{**}$	$r = 0.810^{**}$	$r = -0.664^{**}$	$r = -0.767^{**}$	$r = -0.669^{**}$	$r = 0.886^{**}$

*Means correlation is significant at the level of 0.05.

**Means correlation is significant at the level of 0.001.

$^\wedge$Means no correlation.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
🟥 GC_3%	38	40	38	40	40	39	40	40	38	38	39	40	40	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39
🟪 AT_3%	62	60	62	60	60	61	60	60	62	62	62	60	60	61	61	60	61	61	61	60	61	61	60	61	61	61	60	61	60	61	62	61

Figure 1: Comparison of AT and GC content at synonymous third codon positions in the genes under study. Clearly, AT_3 is much higher than GC_3 in all the accessions.

value of 58 ± 0.363. Based upon these observations, it appears that the extent of codon usage bias in these genes is generally constant. The ENC values were analyzed for possible correlations with the nucleotide compositional parameters, particularly GC_3 content which has been shown previously to correlate with the former [12]. The results of our analyses are in accordance with the significant positive correlations between ENC and GC_3 ($r = 0.431$, $P = 0.014$) as well as ENC and overall GC content ($r = 0.724$, $P = 0.0001$).

3.2. Characteristics of Synonymous Codon Usage. In an attempt to find out the nature of codon usage bias in the genes under study, the RSCU values of the 59 codons were analyzed (Table 3). Interestingly, most of the preferred codons ended with nucleotide *A*. Among the preferred codons, dinucleotide CpG is markedly suppressed while dinucleotides TpA and CpA were found to be abundant in most of them.

In quest for possible under- and over-representation of codons, RSCU values were sorted from lower to higher values. We observed that majority of the codons, both preferred as well as non-preferred, fall under unbiased or randomly used category (0.6 < RSCU < 1.6). Seven codons (GCA, AGA, CTA, TCA, ACA and GTA) showed very high RSCU values (RSCU > 1.6) and hence, were considered to be "over-represented". Similarly there were ten under-represented codons (RSCU < 0.6) (Figure 2).

All the amino acids showed preference over a particular codon except *Asp* where both the codons were used equally (Figure 4). Surprisingly, in all the accessions, out of six possible codons for *Arg,* only two codons, *namely,* AGA and AGG, were used omitting the rest four. Among these two codons, there was a high bias towards AGA with RSCU values 4.61 as compared to that of 1.32 for AGG. *Ser* and *Leu* were the most frequently used amino acids, while *Cys, Gln* and *His* were used least frequently. Frequency of the amino acids *Lys, Gly, Asn, Thr, Val* etc. were also towards higher side (Figure 3).

Highly expressed genes show a tendency of high biasness towards some codons and tend to use those codons frequently. To find out such biasness and predict the expression of the genes, CAI values were estimated, values of which range from 0 to 1. The CAI values for the hemagglutinin genes were found in the range of 0.3143–0.3447 with an average of 0.3829 and standard deviation of 0.0391, indicating that the codons are not translationally optimized for expression of these genes.

The frequency of optimal codons (Fop) in a gene can be used as an indicative measure to check if the codons are optimized for efficient translation [23]. The optimized codons refer to the codons with highest transfer RNA (tRNA) copy number. The results showed a similar trend of Fop to that of RSCU values; the codons with higher RSCU values also tend to have higher Fop values (Figure 4). These two parameters

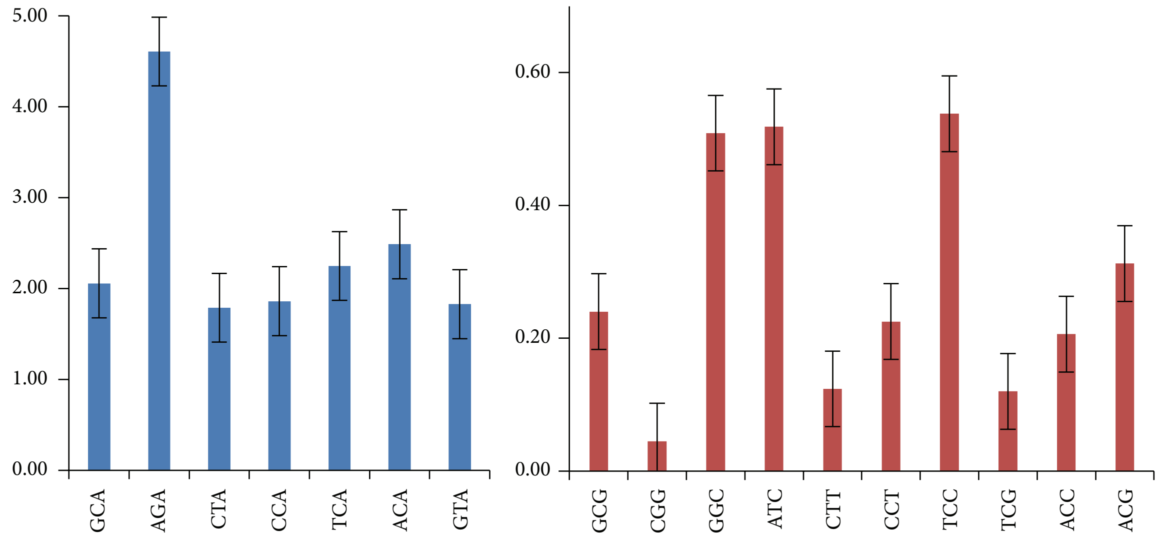


FIGURE 2: Over- and underrepresented codons in the genes used in the study. The overrepresented codons (RSCU > 1.6) are shown in blue, while the underrepresented (RSCU < 0.6) ones are shown in red.

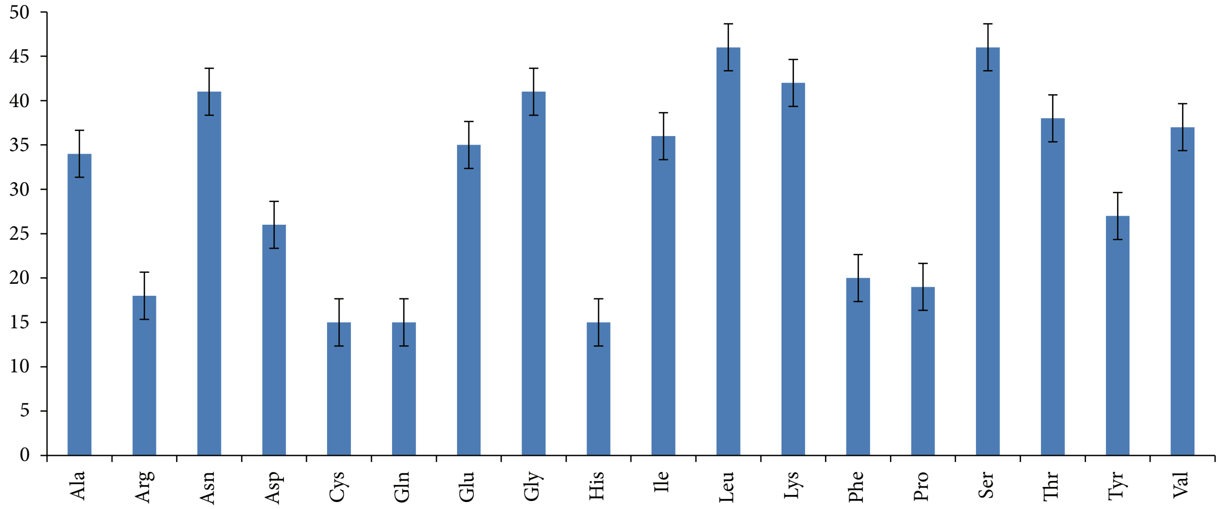


FIGURE 3: Frequency of the amino acid usage in the genes under study. Leucine and serine are clearly the most frequent amino acids.

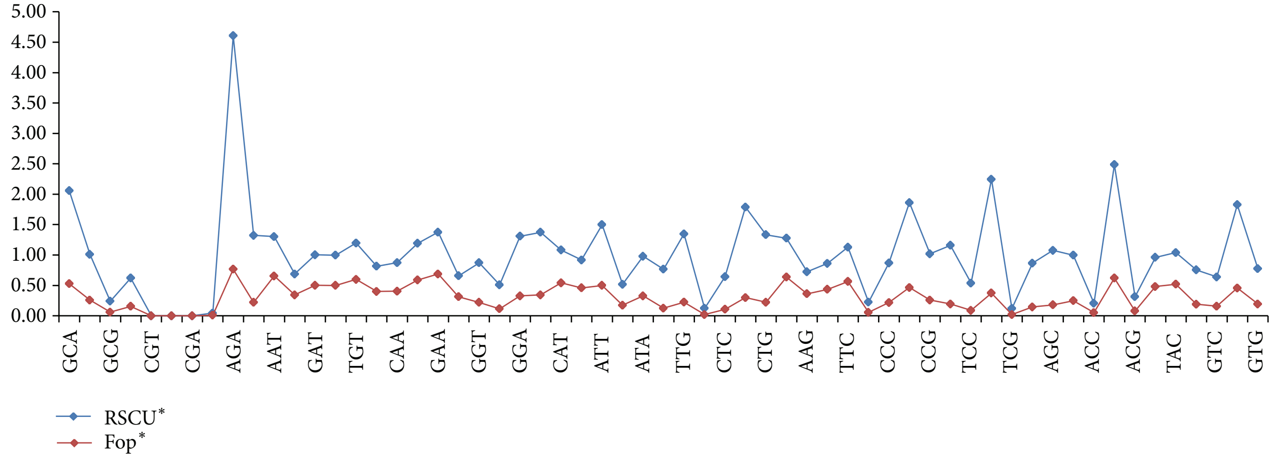


FIGURE 4: Trend of RSCU and Fop values in the coding sequences of the genes.

Table 3: Synonymous codon usage pattern in 32 coding sequences.

AA	Codon	RSCU*	Fop*	*N**
Ala	**GCA**	**2.06**	**0.53**	**18**
	GCC	1.01	0.26	9
	GCG	0.24	0.06	2
	GCT	0.62	0.16	5
Arg	CGT	0.00	0.00	0
	CGC	0.00	0.00	0
	CGA	0.00	0.00	0
	CGG	0.05	0.01	0
	AGA	**4.61**	**0.77**	**14**
	AGG	1.32	0.22	4
Asn	**AAT**	**1.30**	**0.66**	**27**
	AAC	0.69	0.34	14
Asp	GAT	1.00	0.50	13
	GAC	1.00	0.50	13
Cys	**TGT**	**1.20**	**0.60**	**9**
	TGC	0.81	0.40	6
Gln	CAA	0.87	0.40	6
	CAG	**1.19**	**0.59**	**9**
Glu	**GAA**	**1.38**	**0.69**	**24**
	GAG	0.66	0.31	11
Gly	GGT	0.87	0.22	9
	GGC	0.51	0.12	5
	GGA	1.31	0.33	13
	GGG	**1.37**	**0.34**	**14**
His	**CAT**	**1.08**	**0.54**	**8**
	CAC	0.92	0.46	7
Ile	**ATT**	**1.50**	**0.50**	**18**
	ATC	0.52	0.17	6
	ATA	0.98	0.33	12
Leu	TTA	0.77	0.12	6
	TTG	1.35	0.22	10
	CTT	0.12	0.02	1
	CTC	0.64	0.11	5
	CTA	**1.79**	**0.30**	**14**
	CTG	1.33	0.22	10
Lys	**AAA**	**1.28**	**0.64**	**27**
	AAG	0.72	0.36	15
Phe	TTT	0.86	0.44	9
	TTC	**1.13**	**0.56**	**11**
Pro	CCT	0.23	0.06	1
	CCC	0.87	0.22	4
	CCA	**1.86**	**0.47**	**9**
	CCG	1.02	0.26	5
Ser	TCT	1.16	0.19	9
	TCC	0.54	0.09	4
	TCA	**2.25**	**0.37**	**17**
	TCG	0.12	0.02	1
	AGT	0.87	0.14	7
	AGC	1.08	0.18	8
Thr	ACT	1.00	0.25	9
	ACC	0.21	0.05	2
	ACA	**2.49**	**0.62**	**24**
	ACG	0.31	0.08	3
Tyr	TAT	0.96	0.48	13
	TAC	**1.04**	**0.52**	**14**
Val	GTT	0.76	0.19	7
	GTC	0.64	0.16	6
	GTA	**1.83**	**0.46**	**17**
	GTG	0.78	0.19	7

Note: *All values are mean values; *N* represents the number of codons; the preferentially used codons for each amino acid are described in bold.

showed a significant positive correlation with correlation coefficient of $r = 0.710$ ($P = 0.0001$).

4. Conclusion

Amidst much debate, mutational pressure and natural selection have been cited as the major stimulants in framing the codon usage profiles of different viruses [5, 20, 24]. As in most of the RNA viruses, mutation rate of IAV is very high and the effects of codon usage bias are too small for natural selection to act effectively [25]. One possible explanation for lower codon preferences might be due to the fact that it helps the virus to replicate readily in alternate hosts with different codon choices [5].

Hemagglutinin constitutes one of the most important sites for human immune system to act on, thus, making it a potential drug target against this virus. Untangling the underlying mechanisms operating behind the synonymous codon usage profile of the virus will possibly bring up new avenues in the research involving development of antiviral drugs against this hazardous virus.

Disclosure

The authors did not avail any financial assistance from any source in undertaking the present study.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors are grateful to Assam University, Silchar, Assam, India, for providing the research facility.

References

[1] J. B. Plotkin and J. Dushoff, "Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 12, pp. 7152–7157, 2003.

[2] G. J. D. Smith, D. Vijaykrishna, J. Bahl et al., "Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza a epidemic," *Nature*, vol. 459, no. 7250, pp. 1122–1125, 2009.

[3] M. R. Hilleman, "Realities and enigmas of human viral influenza: pathogenesis, epidemiology and control," *Vaccine*, vol. 20, no. 25-26, pp. 3068–3087, 2002.

[4] R. Grantham, C. Gautier, and M. Gouy, "Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type," *Nucleic Acids Research*, vol. 8, no. 9, pp. 1893–1912, 1980.

[5] G. M. Jenkins and E. C. Holmes, "The extent of codon usage bias in human RNA viruses and its evolutionary origin," *Virus Research*, vol. 92, no. 1, pp. 1–7, 2003.

[6] W. Gu, T. Zhou, J. Ma, X. Sun, and Z. Lu, "The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*," *BioSystems*, vol. 73, no. 2, pp. 89–97, 2004.

[7] D. B. Levin and B. Whittome, "Codon usage in nucleopolyhedroviruses," *Journal of General Virology*, vol. 81, part 9, pp. 2313–2325, 2000.

[8] T. Ikemura, "Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs," *Journal of Molecular Biology*, vol. 158, no. 4, pp. 573–597, 1982.

[9] P. M. Sharp, L. R. Emery, and K. Zeng, "Forces that influence the evolution of codon bias," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1544, pp. 1203–1212, 2010.

[10] B. D. Greenbaum, A. J. Levine, G. Bhanot, and R. Rabadan, "Patterns of evolution and host gene mimicry in influenza and other RNA viruses," *PLoS Pathogens*, vol. 4, no. 6, Article ID e1000079, 2008.

[11] Z. P. Li, D. Ying, P. Li, F. Li, X. Bo, and S. Wang, "Analysis of synonymous codon usage bias in 09H1N1," *Virologica Sinica*, vol. 25, no. 5, pp. 329–340, 2010.

[12] T. Zhou, W. Gu, J. Ma, X. Sun, and Z. Lu, "Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses," *BioSystems*, vol. 81, no. 1, pp. 77–86, 2005.

[13] E. H. M. Wong, D. K. Smith, R. Rabadan, M. Peiris, and L. L. M. Poon, "Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus," *BMC Evolutionary Biology*, vol. 10, no. 1, article 253, 2010.

[14] P. M. Sharp and W. Li, "The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications," *Nucleic Acids Research*, vol. 15, no. 3, pp. 1281–1295, 1987.

[15] F. Wright, "The "effective number of codons" used in a gene," *Gene*, vol. 87, no. 1, pp. 23–29, 1990.

[16] J. M. Comeron and M. Aguadé, "An evaluation of measures of synonymous codon usage bias," *Journal of Molecular Evolution*, vol. 47, no. 3, pp. 268–274, 1998.

[17] J. A. Novembre, "Accounting for background nucleotide composition when measuring codon usage bias," *Molecular Biology and Evolution*, vol. 19, no. 8, pp. 1390–1394, 2002.

[18] L. E. Post and M. Nomura, "DNA sequences from the str operon of *Escherichia coli*," *Journal of Biological Chemistry*, vol. 255, no. 10, pp. 4660–4666, 1980.

[19] M. Gouy and C. Gautier, "Codon usage in bacteria: correlation with gene expressivity," *Nucleic Acids Research*, vol. 10, no. 22, pp. 7055–7074, 1982.

[20] J. Zhong, Y. Li, S. Zhao, S. Liu, and Z. Zhang, "Mutation pressure shapes codon usage in the GC-rich genome of foot-and-mouth disease virus," *Virus Genes*, vol. 35, no. 3, pp. 767–776, 2007.

[21] N. Goñi, A. Iriarte, V. Comas et al., "Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development," *Virology Journal*, vol. 9, article 263, 2012.

[22] S. Karlin, W. Doerfler, and L. R. Cardon, "Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses?" *Journal of Virology*, vol. 68, no. 5, pp. 2889–2897, 1994.

[23] Y. Lavner and D. Kotlar, "Codon bias as a factor in regulating expression via translation rate in the human genome," *Gene*, vol. 345, no. 1, pp. 127–138, 2005.

[24] P. M. Sharp, M. Stenico, J. F. Peden, and A. T. Lloyd, "Codon usage: mtational bias, translational selection, or both?" *Biochemical Society Transactions*, vol. 21, no. 4, pp. 835–841, 1993.

[25] A. J. Leigh Brown, "Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 5, pp. 1862–1865, 1997.

Available Online at http://www.recentscientific.com

ISSN: 0976-3031

International Journal of Recent Scientific Research
Vol. 5, Issue, 5, pp.965-969, May, 2014

International Journal of Recent Scientific Research

RESEARCH ARTICLE

CODON USAGE BIAS IN H1N1 NEURAMINIDASE: SELECTION OR MUTATIONAL BIAS?

Himangshu Deka, Supriyo Chakraborty*

Department of Biotechnology, Assam University, Silchar-788011, Assam, India

ARTICLE INFO

Article History:

Received 15th, April, 2014
Received in revised form 27th, April, 2014
Accepted 17th, May, 2014
Published online 28th, May, 2014

Key words:

Codon usage bias, H1N1, Influenza A, Neuraminidase, Synonymous codon

ABSTRACT

Influenza A virus (IAV) has been a major concern worldwide as a cause of high mortality and morbidity. In the present study, the complete coding regions of viral neuraminidase (NA) gene of IAV subtype H1N1 reported from India were analyzed for the possible codon usage bias using statistical and bioinformatics tools. A total of 34 NA coding sequences were used in the study. The results show a low bias in the coding region of the NA gene sequences. The RSCU values suggest a very low preference of the codons having dinucleotide CpG whereas most of the codons showed a preferred use of the dinucleotides CpA and TpA. The results suggest that there exists a balance between mutational pressure and natural selection to shape the codon usage bias in the IAV subtype which helps the virus adapt to different host conditions.

© Copy Right, IJRSR, 2014, Academic Journals. All rights reserved.

INTRODUCTION

The exponential increase in the volume of sequence information during the early '90s facilitated for the first time the detailed statistical analyses of codon usage (Grantham, Gautier *et al*. 1980). It has been established that there exists a bias in the usage of synonymous codons in the biological system ranging from prokaryotes to complex organisms including the viruses. With the rapid availability of vast number of sequences after whole genome sequencing of large number of species, scientists are now trying to look the codon bias phenomena in holistic manner. Accordingly, on a global basis, investigators have focussed research interest in the context of codon bias phenomenon in specific genes as well as whole genome (Grantham, Gautier *et al*. 1980; Plotkin and Kudla 2011).

The major concern regarding the negative-stranded RNA virus, Influenza A virus, can be understood by the fact that roughly one-fifth of the human populations are infected by the virus every year, causing significant mortality and negative economic impacts on society worldwide. Among the major influenza pandemics two was caused by the H1N1 strain, one in the year 1918 and the latest in 2009 (Cox, Black *et al*. 1989; Dawood, Jain *et al*. 2009). The first outbreak of the H1N1 of this century originated in Mexico in 2009 which later spread to about 207 countries worldwide with a death toll of more than 7,800. Apart from these two several other outbreaks of H1N1 have been reported in 1950s and in 1970s (Goni, Iriarte *et al*. 2012).

While human immune system develops resistance against most pathogens upon exposure to them, IAV poses serious threat to the host immunity by presenting a moving antigenic target. This process, termed as antigenic drift, helps it to escape the specific immunity caused by earlier infections. Drift is the result of the selective fixation of mutations in the hemagglutinin (HA) and neuraminidase (NA) genes (Goni, Iriarte *et al*. 2012). The viral neuraminidase (NA) is frequently used as an antigenic determinant found on the surface of the Influenza virus. While, in some other variants of the influenza neuraminidase confers more virulence to the virus than others making it as a potential drug target for the prevention of the spread of influenza infection (Liu, Eichelberger *et al*. 1995).

Apart from genetic drift, which tends to be a slow evolutionary process, shift is another process by which IAV evolves. The genetic information is shared between the IAV strains triggering rapid evolutionary change in the virus. As happened in case of the 2009 pandemic, such rapid change may result in cross-species shift (Dawood, Jain *et al*. 2009).

Several workers have reported that the overall codon usage bias in RNA viruses is low and there is little variation in bias between genes (Jenkins and Holmes 2003; Gu, Zhou *et al*. 2004; Goni, Iriarte *et al*. 2012). The low codon usage bias in the RNA viruses is attributed to GC compositional properties and dinucleotide content in these viruses. Mutational bias has been projected as the main factor that drives the codon usage variation among the influenza A viruses which are phylogenetically conserved (Gu, Zhou *et al*. 2004).

The analysis of synonymous codon usage is used to investigate the interplay between the mutational pressure exerted by the pathogen on host and the selection pressure on the former by the latter (Jenkins and Holmes 2003). Over the years many authors have reported a number of tools which can be used to measure codon usage bias across genes and genomes. Among these measures, GC content, relative synonymous codon usages (RSCU), effective number of codons (ENC) are some most widely used parameters for codon bias study. RSCU measures the frequency of a particular codon compared to the expected frequency if all synonymous codons are used equally (Sharp and Li 1987; Novembre 2002). While ENC measures the deviation of the codon usage from equal usage of the synonymous codons in a gene or genome, it does not give the direction of bias (Plotkin and Dushoff 2003).

MATERIALS AND METHODS

Datasets

In this study, a total of 34 complete coding sequences of the neuraminidase (NA) gene of human-host derived influenza A virus subtype H1N1 reported from India were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/). The serial numbers (SN), accession numbers and other information are presented in **table 1**.

Parameters for codon usage bias study

To examine the synonymous codon usage in the genes RSCU values were calculated. RSCU is defined as the ratio of the observed frequency to the expected frequency if all the synonymous codons for

* Corresponding author: **Supriyo Chakraborty**
Department of Biotechnology

those amino acids are used equally (Sharp and Li 1987). If the RSCU value of a codon is more than 1.0 it is said to have a positive codon usage bias, while a value of less than 1.0 means a negative codon usage bias. When the RSCU value is close to 1.0, it means that this codon is chosen randomly and equally with other synonymous codons.

Table 1 Information of the complete coding sequences of the 34 NA genes

SN	Accession No	Gene Length
1	KF280657	1410
2	KF280665	1410
3	KF280673	1410
4	KF280681	1410
5	KF280689	1410
6	KF280697	1410
7	KF280705	1410
8	KF280713	1410
9	KF280721	1410
10	KF280729	1410
11	KF280737	1410
12	KF280745	1410
13	KF280753	1410
14	JX262202	1410
15	JX262201	1410
16	HM460506	1413
17	JF265672	1410
18	JF265671	1410
19	HM241726	1411
20	HM241719	1429
21	HM241712	1428
22	HM241705	1428
23	CY088710	1428
24	CY088703	1428
25	CY088696	1413
26	CY088689	1428
27	CY088682	1410
28	CY088675	1438
29	CY088668	1421
30	CY088661	1428
31	CY088654	1421
32	CY088647	1429
33	CY088640	1428
34	CY088633	1438

The effective number of codons (ENC) is estimated to quantify the synonymous codon usage across the target sequence which is given below:

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

where, F_k (k = 2, 3, 4 or 6) is the average of the F_k values for k-fold degenerate amino acids. The F value denotes the probability that two randomly chosen codons for an amino acid with two codons are identical. The values of ENC range from 20 (when only one codon is used per amino acid) to 61 (when all synonymous codons are equally used for each amino acid) (Wright 1990; Novembre 2002).

GC_{3s} is the frequency of the nucleotides G+C at the synonymous 3rd positions of the codons excluding Met, Trp and the termination codons. Similarly $GC_{1s\ \mathrm{and}}$ GC_{2s} represent G+C frequency at 1st and 2nd codon positions. GC_{3s} is a good indicator of the extent of base composition bias.

Gene expressivity was measured by codon adaptation index (CAI) as given by Sharp and Li (Sharp and Li 1987). CAI has been used as a simple and effective measure of the overall synonymous codon usage bias of a gene. CAI was originally proposed to provide a normalized estimate that can be used across genes and species, ranging from 0 to 1. The boundary values refer to the cases in which only the most frequent codons (CAI = 1) or only the least frequent codons (CAI = 0) are used within a gene. CAI is given by the following formula:

$$CAI = exp\frac{1}{L}\sum_{k=1}^{L} \ln w_{c(k)}$$

where, L is the number of codons in the gene and $w_{c(k)}$ is the ω value for the k-th codon in the gene.

Frequency of optimal codon (Fop) in a codon is used as an index to show the optimization level of synonymous codon choice in each gene to translation process (Ikemura 1982). Fop is defined as the ratio of total number of optimal codons in a gene to the total number of synonymous as well as non synonymous codons in that gene.

The codon usage bias measures namely RSCU, ENC, GCs and CAI for each coding sequence were estimated by using a Perl program developed by SC.

RESULT AND DISCUSSION

The nucleotide content and the overall GC content at the three codon positions reveal that most of the preferential codons use A at the synonymous third codon position. The overall percentage of A, T, G, C and overall GC content in the three codon positions are shown in table 2. Throughout the accessions A% is higher than the rest of the nucleotides with an average value of 30. As evident from the results, GC3% is higher than GC1% and GC2% in all the accessions with a mean of 46.0%. (Fig 1).

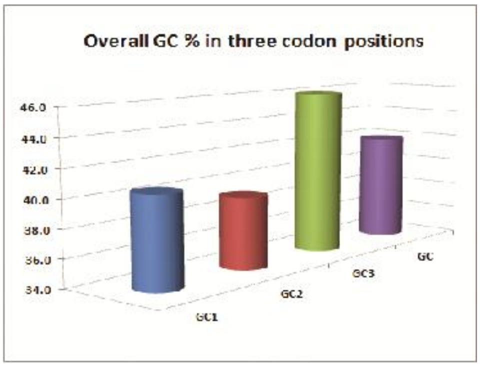


Fig 1 GC content in three codon positions: Percentage of GC content in the three codon positions (GC1, GC2 and GC3) and overall GC content in the coding sequences of the NA gene used in this study.

Previous studies have revealed that influenza A virus strains infecting human hosts since 1918 have been selected under strong pressure to reduce the frequency of CpG in its genome. The possible explanation for low CpG may be immunologic escape as unmethylated CpGs are recognized by the host’s innate immune system as a pathogen signature (Greenbaum, Levine *et al*. 2008). Marked CpG deficiency has also been reported in several RNA viruses including H1N1 (Greenbaum, Levine *et al*. 2008; Wong, Smith *et al*. 2010). Thus, escape from the host antiviral response could act as a selective pressure contributing to codon usage in H1N1.

To peruse the possible effects of CpG under-represented on codon usage bias, the RSCU values were examined for the eight codons having dinucleotide combination of CpG (CCG, GCG, TCG, ACG, CGC, CGG, CGT, and CGA). Our analysis indicated that CGA and CGG were more preferred but the rest of the codons containing CpG are markedly suppressed. Similarly, out of six codons containing TpA (TTA, CTA, GTA, TAT, TAC, ATA) only two codons are preferred (i.e., TAT and ATA). However, codons containing CpA (TCA, CCA, ACA, GCA, CAA, CAG, CAT and CAC) show a remarkable preferentiality over others with five preferred codons out of eight. Codons containing the dinucleotide TpG (TTG, GTG, TGT, TGC, and CTG) also show a low preferentiality. However, it was interesting to note that most of the preferred codons contain dinucleotide CpA (Fig 2).

Table 2 Nucleotide composition of the 34 coding sequences of NA gene

SN	A%	T%	G%	C%	GC%	GC1%	GC2%	GC3%	ENC
1	31.8	26.1	23.6	18.5	42.1	40.6	39.4	46.4	59
2	32.1	26.0	23.2	18.7	41.9	41.1	38.7	46.0	59
3	32.1	26.0	23.3	18.6	41.9	40.4	39.1	46.3	58
4	32.1	26.0	23.4	18.6	42.0	40.9	39.1	46.0	59
5	32.1	26.0	23.3	18.7	42.0	40.9	39.1	46.0	59
6	32.1	26.1	23.3	18.5	41.8	40.4	38.7	46.4	58
7	32.1	26.0	23.3	18.6	41.9	41.3	38.9	45.5	59
8	31.9	26.0	23.5	18.7	42.1	41.3	39.1	46.0	59
9	31.8	26.1	23.5	18.7	42.1	40.9	38.9	46.6	59
10	32.0	26.0	23.5	18.6	42.1	40.6	39.1	46.4	59
11	32.1	26.0	23.3	18.6	41.9	40.4	39.1	46.1	58
12	31.8	26.2	23.6	18.4	42.0	40.0	39.8	46.2	58
13	32.5	26.0	22.8	18.7	41.5	40.4	38.3	45.7	58
14	32.0	26.1	23.3	18.6	41.9	40.6	39.1	46.1	59
15	32.0	26.1	23.3	18.6	41.9	40.6	39.1	46.1	59
16	32.7	26.3	23.2	17.8	41.0	38.0	38.6	46.3	57
17	32.1	26.1	23.3	18.6	41.8	40.6	38.9	46.0	59
18	32.0	26.0	23.3	18.7	42.0	40.8	39.1	46.0	59
19	31.9	26.1	23.4	18.6	42.0	40.8	39.3	46.0	59
20	32.1	26.2	23.1	18.7	41.8	40.3	38.9	46.2	58
21	31.9	26.3	23.2	18.6	41.7	39.9	39.5	45.7	58
22	31.9	26.3	23.2	18.6	41.8	40.3	39.5	45.7	58
23	31.9	26.1	23.2	18.8	41.7	40.0	39.9	45.6	59
24	31.9	26.3	23.2	18.6	41.8	40.3	39.5	45.5	58
25	31.8	26.2	23.4	18.6	42.0	40.6	39.5	46.0	59
26	31.9	26.2	23.2	18.7	41.9	40.5	39.5	45.7	59
27	31.8	25.9	23.5	18.8	42.3	41.3	39.6	46.1	59
28	31.8	26.4	23.1	18.7	41.8	40.4	39.2	45.7	59
29	31.8	26.2	23.4	18.6	42.0	40.5	39.3	46.1	59
30	31.9	26.2	23.2	18.6	41.8	40.3	39.5	45.8	58
31	32.0	26.0	23.3	18.6	41.9	40.7	39.5	45.7	59
32	32.1	26.1	23.1	18.7	41.7	40.3	39.0	46.0	58
33	31.9	26.2	23.2	18.7	41.8	40.3	39.5	45.7	58
34	31.8	26.4	23.1	18.7	41.8	40.4	39.2	45.7	58

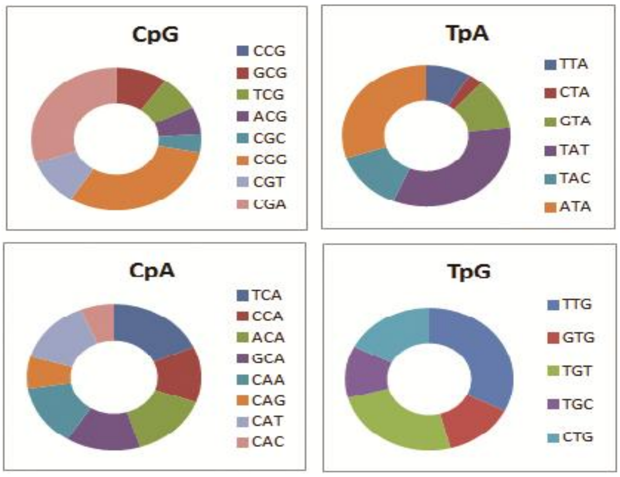


Fig 2 Preference of the dinucleotide in the codons. Most of the preferred codons use CpA while frequency of CpG is very less.

The general trend of ENC values is consistent throughout, (range 57.0-59.0) with an average of 58.6 and standard deviation of 0.0261. The high ENC values signify that the majority of the NA genes of H1N1 do not show a strong codon bias. This is in accordance with the previously published literature (Comeron and Aguade 1998; Jenkins and Holmes 2003; Zhou, Gu *et al*. 2005). The published data suggest that the reason for weak bias in different RNA viruses may be a strategy of these viruses to replicate efficiency in the vertebrate host cells with distinct codon choices (Sharp and Li 1987; Zhang, Wang *et al*. 2011).

Highly expressed genes tend to use limited number of codons and show a tendency of high biasness towards those codons. The CAI value directly corresponds to the expression of the genes. It has been used to measure the extent of codon bias in a gene to examine the adaptation of its codons towards the codon usage of highly expressed genes (Sharp and Li 1987).

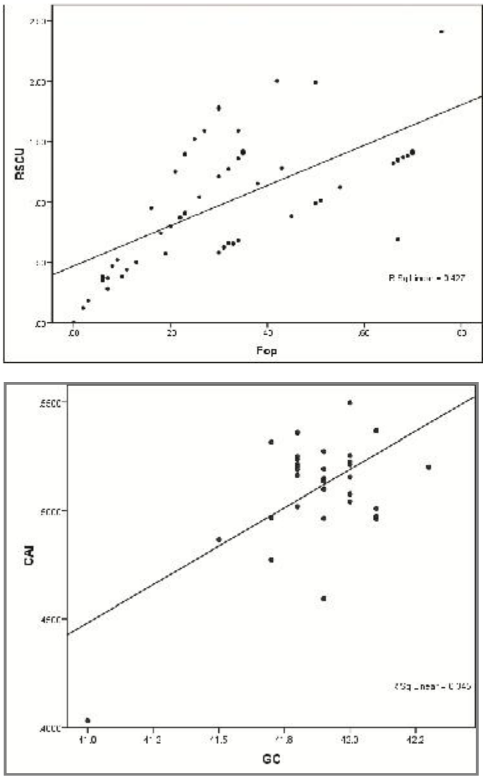


Fig 3 Correlation between a) RSCU and Fop and b) between GC and CAI values in the coding sequences of the NA gene

A value close to 1.0 indicates very high expression while lower values suggest a low expression and hence low codon bias (Grantham, Gautier *et al*. 1980; Sharp and Li 1987).

Table 3 Synonymous codon usage pattern in the 34 coding sequences

AA	Codon	RSCU*	N*	Fop*	AA	Codon	RSCU*	N*	Fop*
Ala	GCA	**1.36**	**3**	**0.34**	Leu	TTA	0.37	3	0.06
	GCC	0.91	2	0.23		TTG	**1.77**	**14**	**0.30**
	GCG	0.44	1	0.11		CTT	1.52	12	0.25
	GCT	1.27	**3**	0.32		CTC	1.25	10	0.21
Arg	CGT	0.47	3	0.08		CTA	0.12	1	0.02
	CGC	0.18	1	0.03		CTG	0.95	7	0.16
	CGA	0.38	2	0.06	Lys	AAA	**1.34**	**19**	**0.67**
	CGG	1.39	8	0.23		AAG	0.66	9	0.33
	AGA	**2.00**	**12**	**0.42**	Phe	TTT	0.63	4	0.31
	AGG	1.59	9	0.34		TTC	**1.40**	**10**	**0.70**
Asn	AAT	**1.37**	**22**	**0.68**	Pro	CCT	0.38	1	0.10
	AAC	0.66	10	0.32		CCC	**2.41**	**6**	**0.76**
Asp	GAT	**1.12**	**10**	**0.55**		CCA	1.21	3	0.30
	GAC	0.88	8	0.45		CCG	0.00	0	0.00
Cys	TGT	**1.35**	**12**	**0.67**	Ser	TCT	0.37	2	0.07
	TGC	0.65	6	0.33		TCC	1.59	9	0.27
Gln	CAA	**1.32**	**24**	**0.66**		TCA	**1.78**	**10**	**0.30**
	CAG	0.68	12	0.34		TCG	0.35	2	0.06
Glu	GAA	0.99	**8**	0.50		AGT	1.40	8	0.23
	GAG	**1.01**	**8**	**0.51**		AGC	0.52	3	0.09
Gly	GGT	0.87	5	0.22	Thr	ACT	0.90	3	0.23
	GGC	1.04	6	0.26		ACC	1.41	**5**	0.35
	GGA	**1.40**	**8**	**0.35**		ACA	**1.42**	**5**	**0.35**
	GGG	0.69	4	0.67		ACG	0.28	1	0.07
His	CAT	**1.38**	**13**	**0.69**	Tyr	TAT	**1.42**	**8**	**0.70**
	CAC	0.62	6	0.31		TAC	0.58	3	0.30
Ile	ATT	1.15	8	**0.38**	Val	GTT	0.80	3	0.20
	ATC	0.57	4	0.19		GTC	**1.99**	**8**	**0.50**
	ATA	**1.28**	**9**	0.43		GTA	0.50	2	0.13
						GTG	0.74	3	0.18

*****RSCU, N and Fop values are mean values; AA means Amino acid and N stands for No of codons used.
The preferentially used codons for each amino acid are described in bold

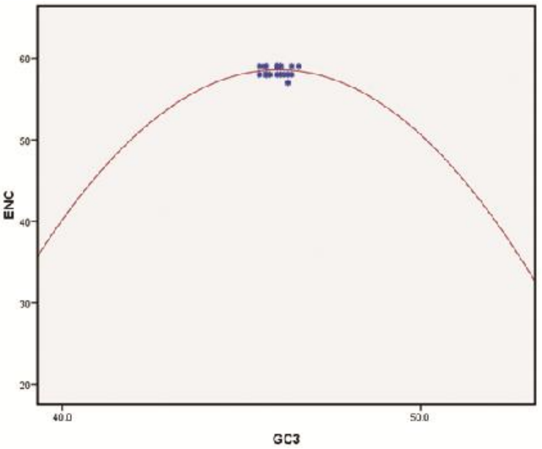


Fig 4: Relationship of ENC and GC3%. It shows the codon usage if GC compositional constraints accounts for codon usage bias alone

The CAI values in the present study are in the range of 0.4031-0.5496 with an average of 0.5100 and standard deviation of 0.0261. This suggests the absence of a strong bias in the gene under study.

Correlation analysis was performed between GC content, ENC, CAI, RSCU and Fop values. There was a strong positive correlation between GC content and ENC (r=0.700, p<0.01) and also between GC and CAI (r=0.587, p<0.01) (Fig 3b). A strong positive correlation was observed between RSCU and Fop values (r= 0.653, p<0.01) (Fig 3a) and between RSCU and ENC (r= 0.361, p<0.05). No significant correlation was found between RSCU and CAI.

The ENC plot (Fig 4) was constructed to investigate the general pattern of synonymous codon usage. The plot was constructed by taking the GC3% values along the X-axis and ENC values along the Y-axis. If the GC3% is the only major factor playing role in the codon choice, the curve of the predicted values will lie above the ENC plots (Wright 1990).

The plot shows most of the points lying on inner side while a few points lying on outer side of the curve indicating that mutational pressure is not the sole force acting in the codon usage bias in the NA gene. There possibly exists a balance between mutational bias and natural selection to shape the codon usage which allows the virus to re-adapt its codon usage to different host environments over time (Zhou, Gu *et al*. 2005; Goni, Iriarte *et al*. 2012).

CONCLUSION

Natural selection and mutational pressure are two major factors which have been reported to affect codon usage bias in various organisms (Sharp and Li 1987). Earlier studies have revealed that mutational pressure, rather than natural selection, is the main factor playing crucial role in shaping the codon usage in most RNA viruses. Apart from mutation pressure in determining patterns of codon usage bias in RNA viruses, the analysis has revealed that the virus is under host immune selection pressure.

Vector-borne RNA viruses are said to have a lower codon usage bias than other RNA viruses (Jenkins and Holmes 2003). One possible explanation that can be attributed here is that a low bias is advantageous to viruses replicating in two different cell types with potentially distinct codon preferences. The replication cycle of IAV is dependent on host machinery and hence the viral replication is affected by the codon usage in the host as well as in the viral genomes. As in case of other RNA viruses, mutation rate of IAV is very high and the effects of codon usage bias too small for natural selection to operate efficiently (Brown 1997). RNA secondary structure may also influence the codon choice in synonymous sites (Simmonds and Smith 1999). The viral neuraminidase presents one of the most important target sites for human immune system (Plotkin and Dushoff 2003). Hence, detailed information about the synonymous codon usage profile may aid in the development of vaccines against the virus.

Acknowledgement

The authors are grateful to Assam University, Silchar-788011, Assam and India for providing the research facility.

DISCLOSURE

The authors do not have any competing interest. The authors did not avail any financial assistance from any source in undertaking the present study.

References

Brown, A. J. (1997). "Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population." Proc Natl Acad Sci U S A 94(5): 1862-1865.

Comeron, J. M. and M. Aguade (1998). "An evaluation of measures of synonymous codon usage bias." J Mol Evol 47(3): 268-274.

Cox, N. J., R. A. Black, *et al*. (1989). "Pathways of evolution of influenza A (H1N1) viruses from 1977 to 1986 as determined by oligonucleotide mapping and sequencing studies." J Gen Virol 70 (Pt 2): 299-313.

Dawood, F. S., S. Jain, *et al*. (2009). "Emergence of a novel swine-origin influenza A (H1N1) virus in humans." N Engl J Med 360(25): 2605-2615.

Goni, N., A. Iriarte, *et al*. (2012). "Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development." Virol J 9: 263.

Grantham, R., C. Gautier, *et al*. (1980). "Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type." Nucleic Acids Res 8(9): 1893-1912.

Greenbaum, B. D., A. J. Levine, *et al*. (2008). "Patterns of evolution and host gene mimicry in influenza and other RNA viruses." PLoS Pathog 4(6): e1000079.

Gu, W., T. Zhou, *et al*. (2004). "The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens." Biosystems **73**(2): 89-97.

Ikemura, T. (1982). "Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs." J Mol Biol 158(4): 573-597.

Jenkins, G. M. and E. C. Holmes (2003). "The extent of codon usage bias in human RNA viruses and its evolutionary origin." Virus Res 92(1): 1-7.

Liu, C., M. C. Eichelberger, *et al*. (1995). "Influenza type A virus neuraminidase does not play a role in viral entry, replication, assembly, or budding." J Virol 69(2): 1099-1106.

Novembre, J. A. (2002). "Accounting for background nucleotide composition when measuring codon usage bias." Mol Biol Evol **19**(8): 1390-1394.

Plotkin, J. B. and G. Kudla (2011). "Synonymous but not the same: the causes and consequences of codon bias." Nat Rev Genet 12(1): 32-42.

Plotkin, J. B. and J. Dushoff (2003). "Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus." Proc Natl Acad Sci U S A 100(12): 7152-7157.

Sharp, P. M. and W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res 15(3): 1281-1295.

Simmonds, P. and D. B. Smith (1999). "Structural constraints on RNA virus evolution." J Virol 73(7): 5787-5794.

Wong, E. H., D. K. Smith, *et al*. (2010). "Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus." BMC Evol Biol 10: 253.

Wright, F. (1990). "The 'effective number of codons' used in a gene." Gene 87(1): 23-29.

Zhang, J., M. Wang, *et al*. (2011). "Analysis of codon usage and nucleotide composition bias in polioviruses." Virol J **8**: 146.

Zhou, T., W. Gu, *et al*. (2005). "Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses." Biosystems 81(1): 77-86.

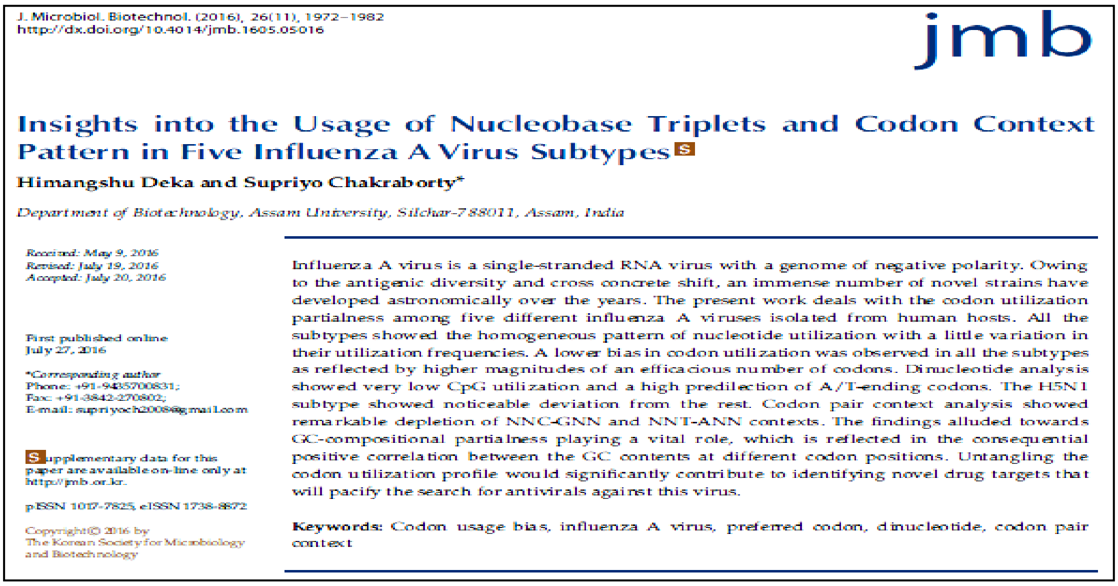
****

APPENDIX B

Lists of Publications

Published Research Articles

1. **Deka, H.,** and Chakraborty, S., **2016**. Insights into the usage of nucleobase triplets and codon context pattern in five influenza A virus subtypes. *J. Microbiol. Biotechnol.* 26(11), pp. 1972-1982.



J. Microbiol. Biotechnol. (2016), 26(11), 1972–1982
http://dx.doi.org/10.4014/jmb.1605.05016

jmb

Insights into the Usage of Nucleobase Triplets and Codon Context Pattern in Five Influenza A Virus Subtypes S

Himangshu Deka and Supriyo Chakraborty*

Department of Biotechnology, Assam University, Silchar-788011, Assam, India

Received: May 9, 2016
Revised: July 19, 2016
Accepted: July 20, 2016

First published online
July 27, 2016

**Corresponding author*
Phone: +91-9435700831;
Fax: +91-3842-270802;
E-mail: supriyoch2008@gmail.com

S upplementary data for this paper are available on-line only at http://jmb.or.kr.

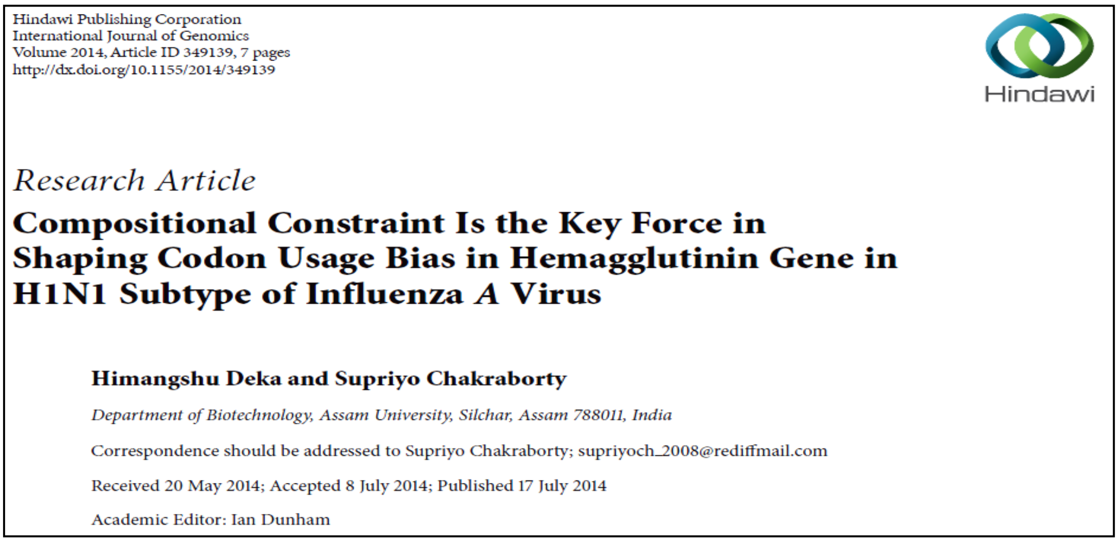
pISSN 1017-7825, eISSN 1738-8872

Copyright© 2016 by
The Korean Society for Microbiology and Biotechnology

Influenza A virus is a single-stranded RNA virus with a genome of negative polarity. Owing to the antigenic diversity and cross concrete shift, an immense number of novel strains have developed astronomically over the years. The present work deals with the codon utilization partialness among five different influenza A viruses isolated from human hosts. All the subtypes showed the homogeneous pattern of nucleotide utilization with a little variation in their utilization frequencies. A lower bias in codon utilization was observed in all the subtypes as reflected by higher magnitudes of an efficacious number of codons. Dinucleotide analysis showed very low CpG utilization and a high predilection of A/T-ending codons. The H5N1 subtype showed noticeable deviation from the rest. Codon pair context analysis showed remarkable depletion of NNC-GNN and NNT-ANN contexts. The findings alluded towards GC-compositional partialness playing a vital role, which is reflected in the consequential positive correlation between the GC contents at different codon positions. Untangling the codon utilization profile would significantly contribute to identifying novel drug targets that will pacify the search for antivirals against this virus.

Keywords: Codon usage bias, influenza A virus, preferred codon, dinucleotide, codon pair context

2. **Deka, H.,** and Chakraborty, S., **2014**. Compositional constraint is the key force in shaping codon usage bias in hemagglutinin gene in H1N1 subtype of influenza A virus. *Int. J. Genomics.* Article ID 349139.



Hindawi Publishing Corporation
International Journal of Genomics
Volume 2014, Article ID 349139, 7 pages
http://dx.doi.org/10.1155/2014/349139

Hindawi

Research Article

Compositional Constraint Is the Key Force in Shaping Codon Usage Bias in Hemagglutinin Gene in H1N1 Subtype of Influenza *A* Virus

Himangshu Deka and Supriyo Chakraborty

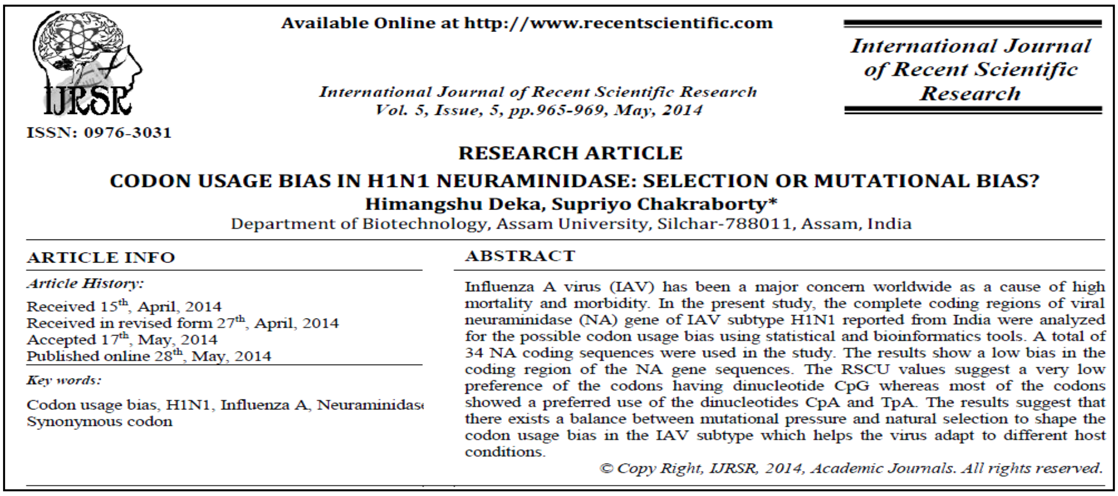
Department of Biotechnology, Assam University, Silchar, Assam 788011, India

Correspondence should be addressed to Supriyo Chakraborty; supriyoch_2008@rediffmail.com

Received 20 May 2014; Accepted 8 July 2014; Published 17 July 2014

Academic Editor: Ian Dunham

3. **Deka, H.,** and Chakraborty, S., **2014**. Codon usage bias in H1N1 neuraminidase: Selection or mutational bias? IJRSR. 3(5), pp 965-969.



Available Online at http://www.recentscientific.com

ISSN: 0976-3031

International Journal of Recent Scientific Research
Vol. 5, Issue, 5, pp.965-969, May, 2014

International Journal of Recent Scientific Research

RESEARCH ARTICLE

CODON USAGE BIAS IN H1N1 NEURAMINIDASE: SELECTION OR MUTATIONAL BIAS?

Himangshu Deka, Supriyo Chakraborty*
Department of Biotechnology, Assam University, Silchar-788011, Assam, India

ARTICLE INFO

Article History:

Received 15th, April, 2014
Received in revised form 27th, April, 2014
Accepted 17th, May, 2014
Published online 28th, May, 2014

Key words:

Codon usage bias, H1N1, Influenza A, Neuraminidase
Synonymous codon

ABSTRACT

Influenza A virus (IAV) has been a major concern worldwide as a cause of high mortality and morbidity. In the present study, the complete coding regions of viral neuraminidase (NA) gene of IAV subtype H1N1 reported from India were analyzed for the possible codon usage bias using statistical and bioinformatics tools. A total of 34 NA coding sequences were used in the study. The results show a low bias in the coding region of the NA gene sequences. The RSCU values suggest a very low preference of the codons having dinucleotide CpG whereas most of the codons showed a preferred use of the dinucleotides CpA and TpA. The results suggest that there exists a balance between mutational pressure and natural selection to shape the codon usage bias in the IAV subtype which helps the virus adapt to different host conditions.

© Copy Right, IJRSR, 2014, Academic Journals. All rights reserved.

4. **Deka, H.,** and Chakraborty, S., **2016**. Molecular overview of codon usage in M1 and M2 genes of influenza A virus. (Communicated)

APPENDIX C

Lists of Workshops and Conferences attended

1. Participated and presented a poster in “Innopharm1”, the international conference organized by Innovare Academic Sciences, at M.P. Council of Science and Technology, Bhopal, MP, during 10-11 October, 2014
2. Participated in “Immunocon-2014”, the 41st Annual conference of Indian Immunology Society held in Madurai Kamaraj University, Madurai, India, during 12-14 December, 2014.
3. Participated and presented a paper in UGC sponsored National Seminar “MBBRNEI”, at Pub Kamrup College during 19-21 August, 2015.
4. Participated in the workshop on “Data analysis for Biological Sciences” conducted by the Applied Statistics Unit of Indian Statistical Institute, Kolkata, held at Department of Statistics, Assam University, Silchar during 28-30 March, 2016.
5. Participated in the DBT sponsored workshop on “R Programming for Genomics Application” organized by DBT-BIF Centre, St. Anthony’s College, Shillong, Meghalaya during 21-23 September, 2016.
6. Participated in the DBT sponsored workshop on “Microbial Transcriptome Data Anlaysis” organized by DBT-BIF Centre, St. Anthony’s College, Shillong, Meghalaya during 21-23 September, 2016.
7. Participated in the DBT sponsored workshop on “Computer Aided Drug Designing: Basics to Molecular Dynamics Simulation” organized by DBT-BIF Centre, Assam University, Silchar during 12-18 September, 2016.