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

Himangshu Deka and Supriyo Chakraborty*

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

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, GC1 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, GC2 and GC3 represent the G+C frequency at the 1st and 2nd codon positions. GC3 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:

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

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

\[ \text{CAI} = \exp \left( \frac{1}{L} \sum_{k=1}^{L} \ln w_{(k)} \right) \]

where \( L \) is the number of codons in the gene and \( w_{(k)} \) is the \( \omega \) 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 H1N1 and 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.](image-url)

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.
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

<table>
<thead>
<tr>
<th>Subtype</th>
<th>Gene name</th>
<th>GC%</th>
<th>GC3%</th>
<th>Nc</th>
<th>Mononucleotides at synonymous position (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A3</td>
</tr>
<tr>
<td>H1N1</td>
<td>HA</td>
<td>40.9</td>
<td>39.6</td>
<td>47</td>
<td>34.9</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>48.4</td>
<td>48.3</td>
<td>53</td>
<td>25.2</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>43.9</td>
<td>42.5</td>
<td>54</td>
<td>23.1</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>41.9</td>
<td>40</td>
<td>52</td>
<td>29.7</td>
</tr>
<tr>
<td></td>
<td>NP</td>
<td>46</td>
<td>42.5</td>
<td>51</td>
<td>34.7</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>44</td>
<td>47.8</td>
<td>55</td>
<td>29.7</td>
</tr>
<tr>
<td></td>
<td>PB1</td>
<td>41.9</td>
<td>43.2</td>
<td>52</td>
<td>33.6</td>
</tr>
<tr>
<td></td>
<td>PB2</td>
<td>44.5</td>
<td>45.7</td>
<td>51</td>
<td>35</td>
</tr>
<tr>
<td>H1N2</td>
<td>HA</td>
<td>41.8</td>
<td>42.6</td>
<td>50</td>
<td>32.8</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>48</td>
<td>46.7</td>
<td>55</td>
<td>26.5</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>44.5</td>
<td>45.7</td>
<td>54</td>
<td>22.2</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>43.4</td>
<td>42.7</td>
<td>54</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>NP</td>
<td>46.1</td>
<td>44.7</td>
<td>52</td>
<td>31.3</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>42.4</td>
<td>45.2</td>
<td>51</td>
<td>31.3</td>
</tr>
<tr>
<td></td>
<td>PB1</td>
<td>42.3</td>
<td>43.7</td>
<td>49</td>
<td>33.5</td>
</tr>
<tr>
<td></td>
<td>PB2</td>
<td>42.4</td>
<td>40.2</td>
<td>50</td>
<td>36.7</td>
</tr>
<tr>
<td>H2N2</td>
<td>HA</td>
<td>42.2</td>
<td>42.2</td>
<td>49</td>
<td>32.9</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>49.3</td>
<td>49.1</td>
<td>56</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>44.7</td>
<td>45</td>
<td>49</td>
<td>20.9</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>44.1</td>
<td>44.1</td>
<td>53</td>
<td>29.7</td>
</tr>
<tr>
<td></td>
<td>NP</td>
<td>46.5</td>
<td>46.1</td>
<td>54</td>
<td>30.8</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>44.3</td>
<td>44.5</td>
<td>51</td>
<td>29.9</td>
</tr>
<tr>
<td></td>
<td>PB1</td>
<td>44.4</td>
<td>44.5</td>
<td>50</td>
<td>29.6</td>
</tr>
<tr>
<td></td>
<td>PB2</td>
<td>44.3</td>
<td>44.5</td>
<td>51</td>
<td>29.9</td>
</tr>
<tr>
<td>H3N2</td>
<td>HA</td>
<td>44.4</td>
<td>44.5</td>
<td>53</td>
<td>30.1</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>48.2</td>
<td>47.5</td>
<td>56</td>
<td>24.3</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>45.1</td>
<td>47.5</td>
<td>52</td>
<td>22.3</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>44.6</td>
<td>44.8</td>
<td>53</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>NP</td>
<td>44.7</td>
<td>44.8</td>
<td>54</td>
<td>28.7</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>44.8</td>
<td>44.9</td>
<td>51</td>
<td>28.5</td>
</tr>
<tr>
<td></td>
<td>PB1</td>
<td>44.6</td>
<td>44.8</td>
<td>49</td>
<td>28.6</td>
</tr>
<tr>
<td></td>
<td>PB2</td>
<td>44.6</td>
<td>44.8</td>
<td>51</td>
<td>29</td>
</tr>
<tr>
<td>H5N1</td>
<td>HA</td>
<td>44.6</td>
<td>44.9</td>
<td>50</td>
<td>28.9</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>45.2</td>
<td>45.4</td>
<td>54</td>
<td>27.8</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>45</td>
<td>45.2</td>
<td>44</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>44</td>
<td>42.2</td>
<td>51</td>
<td>26.7</td>
</tr>
<tr>
<td></td>
<td>NP</td>
<td>47.7</td>
<td>47.3</td>
<td>51</td>
<td>29.9</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>43.6</td>
<td>46.3</td>
<td>53</td>
<td>31.8</td>
</tr>
<tr>
<td></td>
<td>PB1</td>
<td>43.5</td>
<td>46.4</td>
<td>52</td>
<td>31.6</td>
</tr>
<tr>
<td></td>
<td>PB2</td>
<td>44.8</td>
<td>45.4</td>
<td>53</td>
<td>33.8</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>IAV subtype</th>
<th>Regression line</th>
<th>R²</th>
<th>p-Value</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1N1</td>
<td>y = 0.29+0.35x</td>
<td>0.108</td>
<td>&lt;0.001</td>
<td>0.35</td>
</tr>
<tr>
<td>H1N2</td>
<td>y = 0.32+0.29x</td>
<td>0.082</td>
<td>0.001</td>
<td>0.29</td>
</tr>
<tr>
<td>H2N2</td>
<td>y = 0.18+0.6x</td>
<td>0.345</td>
<td>&lt;0.001</td>
<td>0.59</td>
</tr>
<tr>
<td>H3N2</td>
<td>y = -0.03+1.12x</td>
<td>0.389</td>
<td>&lt;0.001</td>
<td>1.12</td>
</tr>
<tr>
<td>H5N1</td>
<td>y = 0.22+0.48x</td>
<td>0.305</td>
<td>&lt;0.001</td>
<td>0.48</td>
</tr>
</tbody>
</table>

**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
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

![Correspondence analysis on RSCU values in the IAV genes considered for the study.](image)

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.

**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. 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.](image-url)
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].

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.
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 summarize, 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


