Infection, Genetics and Evolution 67 (2019) 7–16

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

DNA compositional dynamics and codon usage patterns of M1 and M2 matrix protein genes in influenza A virus

Himangshu Deka[a], Durbba Nath[a], Arif Uddin[b,*], Supriyo Chakraborty[a,**]

[a] Department of Biotechnology, Assam University, Silchar 788011, Assam, India
[b] Department of Zoology, Moinul Hoque Choudhury Memorial Science College, Hailakandi 788150, Assam, India

ARTICLE INFO

Keywords:
Influenza A virus
A/H3N2
Matrix protein
Codon usage
Neutrality
Mutation pressure

ABSTRACT

Influenza A virus subtype H3N2 has been a serious health issue across the globe with approximately 36 thousand annual casualties in the United States of America only. Co-circulation in avian, swine and human hosts has led to frequent mutations in the virus genome, due to which development of successful antivirals against the virus has become a formidable challenge. Recently, focussed research is being carried out targeting the matrix proteins of this strain as vaccine candidates. This study is carried out to unravel the key features of the genes encoding the matrix proteins that manoeuvre the codon usage profile in the H3N2 strains. The findings reveal differential codon choice for both matrix protein 1 and matrix protein 2. The overall codon usage bias is less pronounced in both the datasets which is evident from higher value of effective number of codons (> 55). Comparison of the codon usage for both the genes under study with that of humans revealed that the viral codon usage is not fully optimized for the human host conditions. Both the genes enrolled in the study showed variation which was reflected in almost all the indices used for codon usage studies. Neutrality analysis revealed a weak role of mutation pressure while selection was the major contributor towards codon usage.

1. Introduction

A member of *Orthomyxoviridae,* influenza A virus (IAV) has many subtypes circulating among a diverse range of hosts including human. Among the most important IAV subtypes infecting the humans, H3N2 has been one of the most severe strains. The seasonal IAV subtype H3N2 (A/H3N2) has been a major concern on global basis due to its high variability and reassortment. Last century had witnessed a H3N2 pandemic in the form of the Hong Kong flu during 1968–1969. This particular strain was, in fact, a reassortant of H2N2 subtype due to antigenic shift. Every season about 36,000 humans die in the United States of America as a result of H3N2 infection. The A/H3N2 mainly circulates among swine, avian and human hosts which have been reported to co-circulate with A/H1N1 and avian-origin A/H9N2 in pigs (Campitelli et al., 1997; Peiris et al., 2001).

Influenza A virus exerts tremendous virulence to warm blooded animals (birds and mammals) (Holmes et al., 2005; Rambaut et al., 2008). Based on the antigenic property of surface glycoproteins *viz.* haemagglutinin (HA) and neuraminidase(NA), the virus is classified into several subtypes (Lekcharoensuk et al., 2010). The genome consists of negative sense RNA fragmented into 8 single strands coding for minimum 11 proteins. The virus genome goes through consistent modifications by means of point mutations as well as genetic recombination and reassortment (Chan et al., 2006). The approximate molecular weight of the viral segments ranges from 1 × 10^6 to 2–4 × 10^5 (Palese and Schulman, 1976).

The dire need for an effective vaccine has prompted many workers to try different targets within the IAV genetic framework. Much effort has been laid on the surface proteins hemagglutinin (HA), neuraminidase (NA) and also the ion channel protein M2 (Mosier et al., 2016; Xiong et al., 2015). But due to high plasticity of IAV genome, development of an effective vaccine against these highly mutable surface proteins has been challenging. Alternative targets that are being tried out in recent times include nucleocapsid protein (NP) and integral matrix protein 1 (M1) (Antrobus et al., 2014; Berlanda Scorza et al., 2016). It makes sense in research to gain better understanding of the genetic features of these potential drug targets. Codon usage bias analysis is a useful technique which gives insights for understanding the underlying factors influencing the genetic architecture of an organism and also evolution at molecular level.

Codon usage bias is a well known phenomenon of unequal usage of synonymous codons in coding sequences of genes and has been

* Corresponding author.
** Corresponding author.
E-mail addresses: arif.uddin29@gmail.com (A. Uddin), supriyoch_2008@rediffmail.com (S. Chakraborty).

https://doi.org/10.1016/j.meegid.2018.10.015
Received 20 September 2018; Received in revised form 11 October 2018; Accepted 23 October 2018
Available online 24 October 2018

reported in almost all groups of organisms. A vast majority of viruses have been reported to display codon usage bias including different IAV subtypes (Ahn and Son, 2010; Butt et al., 2014; Goni et al., 2012; Li et al., 2010; Wu et al., 2014). Codon discrimination studies in A/H3N2 were earlier reported by Ahn and Son (2012) wherein they observed directional changes of codon usage between 1993 and 2010 (Ahn and Son, 2012). Similar findings with A/H3N2 in ten RNA segments of the virus were also reported (Lindstrom et al., 1998).

The PB1 gene in H3N2 virus strategizes to acclimatize the avian influenza virus to a new human host through evolution of its synonymous codons. Observations revealed enhanced replication of PB1 gene in IFN-treated human cells, thereby attenuating IFN activity of viral inhibition. It was possible in modern H3N2 virus due to selection of synonymous codons corresponding to the tRNA members in the IFN treated human cells, which drastically differ from untreated cells (Smith et al., 2018). Codon usage studies in Equine Influenza virus reported that polymerase genes of EQV had weaker codon choice. The factors influencing the synonymous codon bias in EQV polymerase genes involve nucleotide array, mutational force, aromaticity and hydropathicity whereas the dominating element is natural selection (Bera et al., 2017). Li et al. studied the synonymous codon usage pattern in H3N2 Canine Influenza virus (CIV), demonstrating similarity of CIV with other influenza viruses with respect to low codon bias. Interestingly, the existence of low codon bias contributes to the replication efficiency in CIV thus causing the CIV outbreak. As determined by neutrality plot and correlation studies, natural selection was found to act as the prime factor regulating the codon usage profile in CIV. The complete coding strands were found to be A/U biased and highly expressed (Bera et al., 2017).

This work is in an extension of our previous report (Deka and Chakraborty, 2016) where we presented the codon usage of eight major genes belonging to five IAV subtypes. In the earlier work we had found codon usage deviation in the genes encoding matrix proteins M1 and M2 from the rest in all the subtypes. The present study aims to investigate the possible factors that might have led to such deviations in codon usage.

2. Methodology

2.1. Sequence data retrieval

The complete coding sequences (cds) in this study for both M1 and M2 genes were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/). The dataset comprised of 1398 cds of M1 and 747 cds of M2 genes. The sequences are presented as text files in the S1a and S1b. We selected and analyzed only those sequences which are exact multiple of three nucleotides with correct start and stop codon.

3. Relative synonymous codon usage

Relative synonymous codon usage (RSCU) is a widely used parameter to unravel the degree of synonymous codon usage across genes and genomes (Sharp and Li, 1986). Assuming that all the synonymous codons for an amino acid are used in equal proportion, RSCU is estimated as the ratio of observed frequency of a codon to its expected frequency. It is calculated using the standard formula:

$$RSCU = \frac{g_{ij}}{\sum_{j}^{ni} gij} ni$$

where, g $_{ij}$ is the frequency of occurrence of the ith codon for the jth amino acid (any g_{ij} with a value of zero is arbitrarily assigned a value of 0.5) and n_i is the kind of synonymous codon.

4. Effective number of codons

The **effective number of codons (Nc)** is an index of the degree of synonymous codon usage bias in a gene. It was calculated after Wright (1990) as:

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

Here, F value denotes the likelihood that the arbitrarily chosen synonymous codons for an amino acid are identical. F_k (k = 2, 3, 4 or 6) stands for the mean of the F values for k-fold degenerate amino acids (Wright, 1990). The Nc values range from 20 to 61; more is the Nc value, lower is the codon usage bias of a gene and *vice versa* (Novembre, 2002; Wright, 1990).

5. Dinucleotide odds ratio

Dinucleotide odds ratio of a gene was calculated by dividing the observed frequency of a dinucleotide pair by the counts of the nucleotides making up that particular pair. Mathematically, it is represented as:

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

Here *fx* and *fy* represent the frequencies of nucleotides *x* and y respectively, whereas *fxy* is the frequency of the dinucleotide comprising of *x* and *y*.

6. Measure independent of length and composition (MILC)

The MILC is a sequence-based predictor of gene expression which is not influenced by the gene length and base composition (Supek and Vlahovicek, 2005). MILC is represented as:

$$MILC = \frac{1}{L} \sum_{a \in A} M_a - K$$

where, L is the number of codons in the cds, M_a is the statistical goodness of fit test for observed to expected codon usage, and *K* is the correction factor. The highest MILC value of the reference set is often used for estimating gene expression *i.e.* MELP (MILC-based expression measure) which is designated as:

$$MELP = \frac{MILC^{(gene)}}{MILC^{(ref)}}$$

In the above formula, $MILC^{(ref)}$ is the highest MILC value for the reference set, and $MILC^{(gene)}$ is the MILC value for the gene in question.

All the above indices of codon usage were estimated using an in-house Perl script developed by the corresponding author (SC).

7. tRNA adaptation index (tAI)

The tAI is used to measure the adaptation of the codons of a cds to tRNA molecules that are essential for translation of codons to amino acids. The tRNA availability is the dynamic force for translational selection. It estimates the level of adaptation of a gene to its genomic tRNA pool (dos Reis et al., 2004). The tAI is calculated as follows:

$$tAI_g = \left(\prod_{k=1}^{l_g} w_{ik} \right)^{1/l_g}$$

where, l_g is the length of the gene in codons and Wi_k is the relative adaptiveness value of the codon defined by the k^{th} triplet in the gene. It was calculated using the online tool (http://tau-tai.azurewebsites.net/#divTables).

8. Correspondence analysis

Correspondence analysis, a widely used multivariate statistical method in codon usage analysis, was employed to investigate the major trends in codon usage variation among genes. The method used for our analysis was implemented in Past software (Hammer et al., 2001).

9. Neutrality plot

A neutrality plot is an analytical method widely used to determine the role of evolutionary forces like mutation and selection on codon usage (Li et al., 2016). In this method, average GC contents at first and second codon positions (GC12s) are plotted in the vertical axis and GC3s values in the horizontal axis in a 2-dimensional scatter plot. If the correlation between GC12s and GC3s is statistically significant, and the slope of regression is close to 1, it is assumed that mutational bias could be the major force operating on codon usage bias. In contrast, if selection effect was dominant over mutational pressure, a narrow distribution of GC content would be observed (Li et al., 2016).

10. Mutational responsive index (MRI)

The MRI is used measure the mutational drift in codons. A positive MRI value indicates directional mutational pressure while a negative value indicates translational selection operating on the gene (Gatherer and McEwan, 1997).

11. Translational selection (P2)

The P2 is used to measures the codon-anticodon interaction efficiency, which indicates the translational efficiency of a gene. The formula used for estimating P2 of each cds is as follows:

$$P2 = (WWC + SSU)/(WWY + SSY)$$

Where W = A or T, S = C or G and Y = C or T.

P2 value > 0.5 reveals bias in favour of translational selection (Gouy and Gautier, 1982).

11.1. Statistical analysis

Correlation analysis was used to identify the relationship between overall nucleotide composition and each nucleotide at 3rd codon position, between gene expression level and codon usage bias. All the statistical analyses were done using the SPSS software.

12. Results

12.1. Base composition of M1 and M2 genes in H3N2

The base compositional analyses in both the datasets were carried out excluding three termination codons and two non-synonymous codons *i.e.* ATG and TGG encoding the amino acids Met and Trp, respectively. The nucleobase usage in M1 dataset of H3N2 followed the decreasing order of A > G > C > T (29.0%, 26.0%, 23.1%, 21.9%, respectively). The base composition at third codon position, however showed a different pattern, preferring T to A at the synonymous sites following the nucleobase preference in the order T3 > A3 > G3 > C3 (29.4, 26.0, 24.3, 20.2%, respectively). The M2 dataset showed a slightly different pattern by favoring A (28.0%) followed by T (26.6%). At silent third codon position, however, the nucleobase T (32.1%) was the most favoured base while G (20.2%) was the least preferred one. The GC content for M1 was 49.1% overall and 44.5% at synonymous third position. Similarly, the GC content for M2 stood at 45.4% overall and 45.2% at silent site. The AT-rich genome and preference of A/T at the silent codon position had been shown in IAV and other RNA viruses in several previous works (Goni et al., 2012; Jenkins and Holmes, 2003; Zhou et al., 2005). The ANOVA results showed that the coding sequences differed among themselves in terms of base composition. The difference in base composition between cds was statistically significant for both M1 (F-test value of cds difference = 19.872, $p < .001$) and M2 datasets (F-test value of cds difference = 48.554, $p < .001$) (S2).

Table 1
Preference of codons in M1 and M2 genes of IAV subtype H3N2.

Amino Acid	Codon	RSCU (M1)	RSCU (M2)	Amino Acid	Codon	RSCU (M1)	RSCU (M2)
Ala	GCA	1.11	0.14		TTA	0.27	0.05
	GCC	1.19	0.59		TTG	0.63	1.88
	GCG	0.26	0.90		CTA	0.77	0.61
	GCT	**1.44**	**2.36**	**Leu**	CTC	1.60	0.46
Arg	CGA	1.02	1.58		CTG	1.15	0.50
	CGC	0.35	0.10		**CTT**	**1.62**	**2.44**
	CGG	0.07	0.03	**Lys**	**AAA**	**1.17**	**1.52**
	CGT	0.36	0.86		AAG	0.83	0.48
	AGA	**2.40**	**2.55**	**Phe**	TTC	0.86	0.93
	AGG	1.81	0.88		**TTT**	**1.14**	**1.07**
Asn	AAC	0.64	1.22		**CCA**	**1.94**	0.06
	AAT	**1.36**	**0.78**	**Pro**	CCC	0.97	0.01
Asp	GAC	0.64	0.91		CCG	0.10	0.80
	GAT	**1.36**	**1.09**		CCT	1.00	**3.13**
Cys	TGC	0.64	**1.97**		TCA	1.09	0.80
	TGT	**1.36**	0.03		TCC	0.73	0.07
Gln	CAA	0.55	0.06	**Ser**	TCG	0.04	0.01
	CAG	**1.45**	**1.96**		TCT	1.03	1.57
Glu	GAA	0.85	**1.19**		AGC	0.91	1.59
	GAG	**1.15**	0.81		**AGT**	**2.25**	**2.01**
Gly	GGA	0.78	0.83		**ACA**	**1.19**	1.23
	GGC	0.74	1.05	**Thr**	ACC	1.19	1.22
	GGG	**1.98**	**1.88**		ACG	0.59	**1.49**
	GGT	0.53	0.24		ACT	1.05	0.04
His	**CAT**	**1.61**	0.71	**Tyr**	TAC	0.51	0.12
	CAC	0.39	**1.29**		**TAT**	**1.49**	**1.90**
Ile	ATA	1.10	0.81		GTA	0.49	0.58
	ATC	0.73	**1.29**		GTC	0.80	**1.53**
	ATT	**1.15**	0.91	**Val**	**GTG**	**1.58**	0.69
					GTT	1.15	1.20

*The preferentially used codons for each amino acid are shown in bold. The red marked amino acids show difference for preferred codon usage in both the genes.

12.2. Codon usage in IAV

The RSCU analysis of codons showed most of the amino acids demonstrated dissimilar preference of codons in both the datasets (Behura and Severson, 2012). In case of M2, amino acids Cys, His, Ile and Val preferred the codons that ended with C. The amino acid Pro showed exceptional preference towards CCT while omitting the codon CCC exclusively in case of M2. Derived from the analysis of all the synonymous codon occurrence (excluding the codons for Met, Trp, and the nonsense codons), seven codons were identified as high-frequency codons or over-represented codons (RSCU > 1.6) in M1 gene. The codons AGA (Arg), AGT (Ser) AGG (Arg), GGG (Gly), CCA (Pro), CTT (Leu) and CAT (His) were over-represented (Table 1). For M2 gene, the scenario of high-frequency codons was a little different with the highest representation of codon CCT (Pro). Ten codons including AGA, AGT, GGG and CTT were over-represented in M2 gene and the same codons also occurred at high frequency in M1 gene. Apart from these, GCT (Ala), TGC (Cys), CAG (Gln), TAT (Tyr) and TTG (Leu) codons were also over-represented. Ironically, at least seven codons were identified which were more preferred in M2 gene but less preferred or avoided in M1 gene (Table 1). These codons encode the amino acids Cys, Glu, His, Ile, Pro, Thr and Val. We performed G-test to test the hypothesis of equal usage of synonymous codons for 18 amino acids. The results revealed statistically significant difference in synonymous codon choice for all the amino acids (G-test, $p < .001$,) (S2).

Table 2
Odds ratios (R_{xy}) for specific dinucleotides in the selected genes.

	AA	AT	AG	AC	TA	TT	TG	TC	GA	GT	GG	GC	CA	CT	CG	CC
M1	0.83	0.86	1.27	0.75	0.65	0.88	**1.22**	**1.22**	1.00	0.83	0.82	**1.28**	**1.45**	**1.43**	**0.52**	0.93
M2	0.83	0.99	1.16	1.02	**0.57**	1.01	1.12	**1.38**	**1.44**	**0.70**	**0.61**	1.07	1.16	**1.34**	0.96	**0.76**

The under-represented ($R_{xy} < 0.80$) and over-represented ($R_{xy} > 1.20$) dinucleotides are marked in boldface.

12.3. Codon usage bias of M1and M2 genes in H3N2

To examine whether any difference exists between the IAV genes in terms of codon usage bias, the effective number of codons (Nc) for each cds was estimated (mean 55.9 ± 1.22 for M1 and 55.6 ± 4.63 for M2). Nc value of a cds holds an inverse relationship with the degree of codon usage bias. Looking at the high Nc values for M1 and M2 genes, we inferred that the overall codon usage bias in these genes was weak, indicating the maintenance of synonymous codon variability in the cds for each amino acid.

12.4. Analysis of tRNA (tAI) adaptation index

The tRNA adaptation index for M1 and M2 genes was 0.2404 and 0.2368, respectively which suggested that the coding sequences of M1 and M2 proteins had low adaptation to the tRNA species.

12.5. Dinucleotide analysis

The dinucleotide analysis (Table 2) showed a tendency of avoiding NCG codons (N stands for any nucleotide) in both the genes, however, M2 exceptionally preferred ACG (Thr). While all the other codons with the aforesaid composition were under-represented (RSCU < 0.60), M2 again proved to be little divergent by moderate representation of GCG (Ala) and CCG (Pro). The NTA codons were also in low frequency in both the datasets except ATA which was moderate. Overall, the highest represented dinucleotides were CpA, CpT, GpC, TpC, TpG in M1 and GpA, TpC and CpT in case of M2. The lowest dinucleotides for M1 gene were CpG, TpA and ApC, while for M2 gene the lowest frequency was observed for TpA, GpT and CpC with moderate representation of CpG. This was, in fact, a bit unusual as most of the previous works including our previous one recorded a low usage of CpG in IAV (Ahn and Son, 2010; Goni et al., 2012; Wong et al., 2010a). The CpG depletion was proposed to be linked with the viral strategy to escape the host immune machinery as unmethylated CpGs are recognized by the host immune cells as the signals of pathogen entry (Greenbaum et al., 2008; Shackelton et al., 2006; Zhong et al., 2007). We concluded that escaping the host immune system could constitute a selective pressure in these IAV genes (Goni et al., 2012).

12.6. Codon usage in relation to host cells

One of the key features of IAV life cycle is the acquisition of the host translational machinery and harnessing it for viral replication within host cells. Thus it is very significant for the virus to adapt to the host cell conditions. In quest for the extent of codon usage bias in H3N2 strains in relation to the human host cells, the codon usage frequency value for each codon was calculated for the two genes under study and compared with the standard codon usage table for humans obtained from Codon Usage Database (http://www.kazusa.or.jp/codon). The results are graphically represented in Fig. 1. The normalized codon usage values for M1 and M2 genes were plotted against the human counterparts along the primary and secondary vertical axes, respectively. Two different trends were observed for both the genes in reference to human host cells. The squared correlation coefficients (R^2) of 0.3256 and 0.1039, respectively for M1 and M2 genes were indicative of the fact that M1 genes were more closely adapted to host cells. The third codon position is crucial for the virus from translational point of view as this is the position that makes wobble pairing with the first position of the corresponding anticodon in the host cells (Goni et al., 2012). Thus it is imperative to get insights into the variations at third position in the context of viral adaptation to host conditions.

A comparative analysis between the virus and the host cell codon usage preferences is represented in Fig. 2. We plotted the codons ending

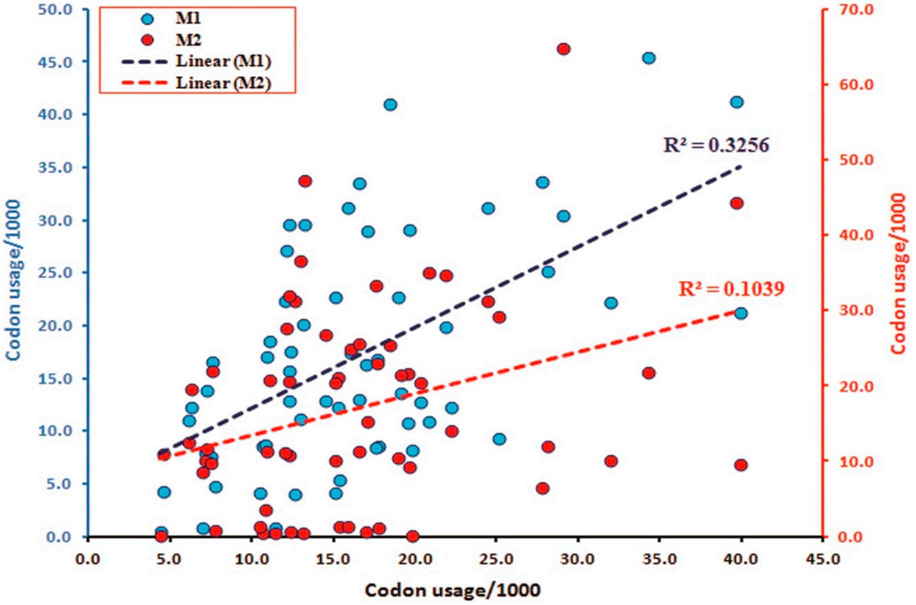

Fig. 1. Comparison between virus and human host cell codon usage. The values are normalized to 'usage per thousand'. The codon usage frequencies of the estimated virus genes were compared with the standard human codon usage values from the codon usage database.

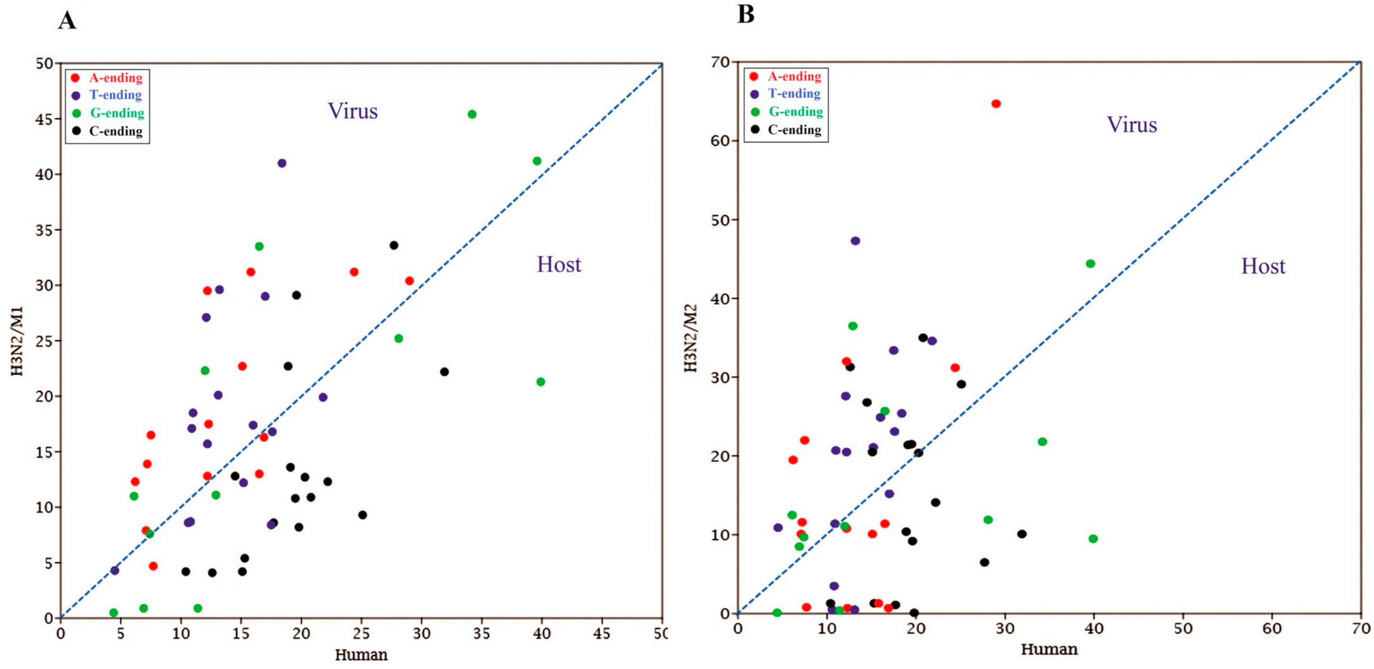

Fig. 2. Comparative representation of the preference for codons in the H3N2 matrix protein coding genes (M1 and M2) and human host cells.

with four different nucleotides in four representing colours. The position of scattered dots indicated the preference of A- and T-ending codons by the viral genes while the host cells preferred G- and C-ending codons more than the rest. However, the extent of codon preference was somewhat different for both the genes at silent position. We performed Pearson's correlation analysis between viral and host codon usage and found significant positive correlations for both M1 (r = 0.571, p < .01) and M2 genes (r = 0.322, p < .05). From the results of correlation analysis we could conclude that the M1 gene of virus was more adapted than M2 gene to the host cell during evolution.

12.7. Neutrality analysis

We performed correlation analysis between GC12 on GC3 and observed highly significant correlation for M1 (r=0.136**, p < .01) and M2 (r = −0.376, p < .01) genes, which suggested that directional mutation pressure acted on all codon positions. Further, we observed points in the neutrality plot were not diagonally located and the values of GC3 values were in a narrow distribution (Fig. 3), suggesting that natural selection might also have influenced the CUB for M1 and M2 genes. Moreover, the contribution of mutation pressure was only 7% for M1 while, in M2 mutation contributed approximately 12.5% as evident from regression coefficients (Fig. 3**)**. Thus, in both the datasets, we could conclude that selection pressure had a more pronounced effect than mutation pressure in determining the codon usage bias of these two genes.

12.8. PR2 bias analysis

The Parity rule 2 (PR2) plot is a measure of intra-strand bias which was performed to investigate the influence of mutation and selection stress on the codon usage of genes (Sueoka, 1995). It is a plot where AT-bias *i.e.* [A3/(A3 + T3)] is plotted as the ordinate and GC-bias *i.e.* [G3/(G3 + C3)] is plotted as the abscissa. According to parity rule 2, the centre of the plot, where both the coordinates are equal to 0.5, is the place where A equals T and G equals C, in absence of the influence exerted by mutation and selection (Sueoka, 2002).

In most of the previous works on PR2 analyses, only the 4- and 6-fold degenerate codons were considered (Chen et al., 2014; Wei and Guo, 2010; Yang et al., 2014). But we included 2-fold codons as well in

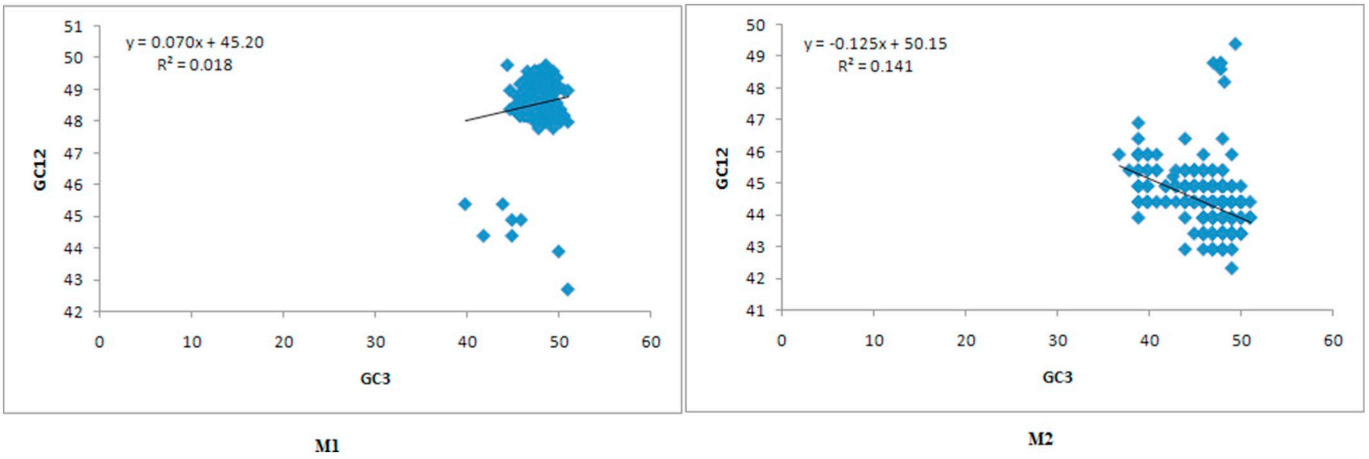

Fig. 3. Neutrality analyses for the M1 and M2 datasets. The regression line for M1 and M2 were y = 0.070× + 45.20 and y = −0.125× + 50.15.

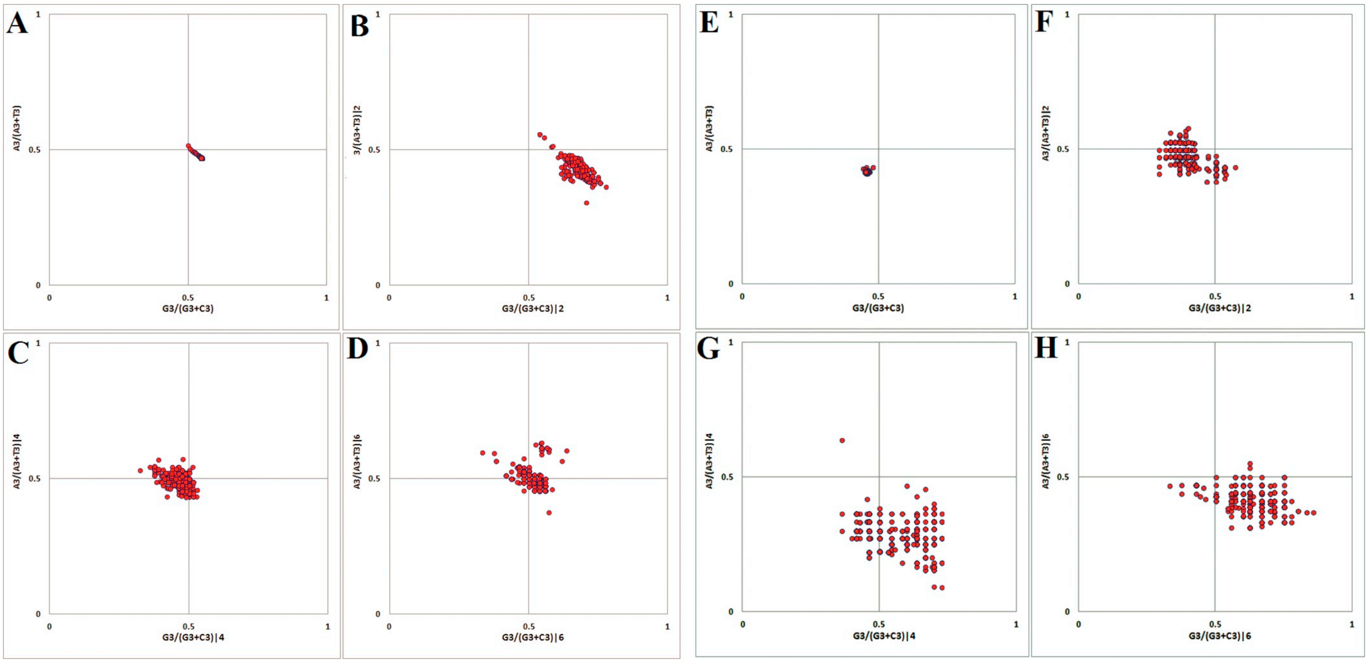

Fig. 4. PR2 analyses for the M1 and M2 genes considering different degeneracy levels. Blocks A and E represent the PR2 fingerprints for the 58 codons for M1 and M2, respectively; 2-fold degenerate codons are shown in blocks B (M1) and F (M2); the 4-fold degenerate codons in blocks C (M1) and G (M2); the 6-fold degenerate codons are in blocks D (M1) and H (M2).

PR2 analysis. Excluding Met and Trp, half the amino acids were encoded by 2-fold degenerate codons and hence we did not ignore 2-fold codons in PR2 analysis. Therefore, to see if there exists any significant difference between PR2 biases based on degeneracy level, we divided the dataset for PR2 analysis into four categories. The first one comprised of all the codons except the stop codons, the codons for Met and Trp and the codon ATA of Ile amino acid (hereafter referred to as case I). The second category composed of the 2-fold degenerate codons (case II), while the 3rd and 4th categories were for the 4-fold (case III) and 6-fold degenerate codons (case IV). The results are summarised in Fig. 4. In M1 gene, PR2 bias was less pronounced except for the 2-fold degenerate codons where there was visible inclination towards G/T at the synonymous sites. The PR2 fingerprints of M2 however showed marked PR2 bias in all the cases. Overall a slight preference was observed for T and C at silent position (case I). For 2-fold degenerate codons, the preference was for T and C with a few codons choosing A at silent sites. Different preference was, however, noticed for case III and IV, where the deviation from PR2 was markedly higher. There was a sharp inclination towards T and G in the synonymous codon positions. To sum up, barring the 2-fold degenerate codons for M1 gene, the rest of the cases showed very little PR2 bias. On the other hand PR2 bias was very much noticeable in case of M2 dataset.

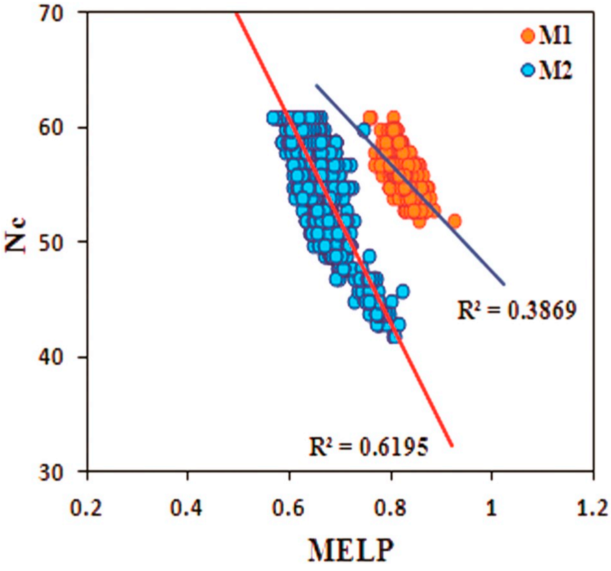

Fig. 5. Correlation between codon usage (Nc) and gene expression (MELP). Both M1 and M2 showed significant negative correlation between these two indices.

12.9. Analysis of mutational responsive index and translational selection (P2)

The MRI for M1 and M2 was 0.43 and 0.58 respectively, which suggested that directional mutation pressure acted on M1 and M2 genes. The P2 values for M1 and M2 were 0.10 and 0.34 respectively, which suggested that the extent of translational efficiency was small (Sur et al., 2007).

12.10. Effect of gene expression level (MELP) on codon usage bias for M1 and M2

It has been well documented that gene expressivity has a profound influence on the underlying codon usage of a particular organism (Coghlan and Wolfe, 2000; Kliman et al., 2003; Sharp and Li, 1987). To unravel any bias occurring in the context of gene expression, we estimated MELP value for the genes in this study. The mean MELP values for M1 and M2 were 0.817 and 0.657, respectively, which suggested that the expression level of the genes was relatively high. Further, in comparison to M2, the gene expression level for M1 was high. In addition we performed correlation analysis between Nc and MELP to understand the effect of expression level for codon usage bias. Highly significant correlation ($r = -0.741^{**}, p < .01$) was found between Nc and MELP for M2 gene (Fig. 5) but no significant relation was found for M1. These results suggested that the expression level of M2 gene might be associated with its codon usage bias unlike M1 gene.

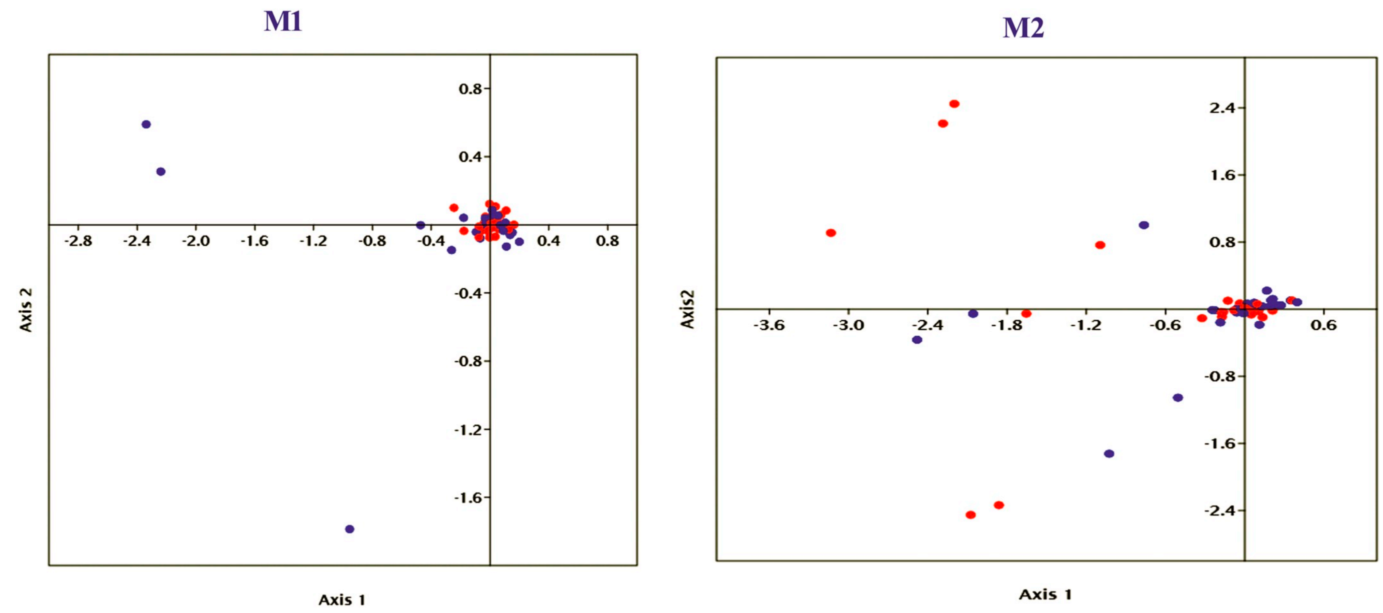

Fig. 6. Correspondence analysis of M1and M2 genes. The red and blue dots represent AT- and GC-ending codons, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

12.11. Correspondence analysis

Correspondence analysis is a multivariate dimension reduction technique to evaluate bulk data in a two-dimensional scatter plot. In this technique, rows and columns represent the variable types which are displayed in a scatter diagram of lower dimension thus making the original data simple to visualize and interpret (Grantham et al., 1980; Greenacre and Hastie, 1987). To map the variability of synonymous codon usage among the genes encoding the matrix proteins, correspondence analysis was performed using the RSCU values of codons for this study. The coordinates of each of the 59 synonymous codons on the two prime axes (axes 1 and 2) are plotted in Fig. 6**.** These two axes together accounted for 60.7% and 43.1% of total variations in M1 and M2 genes, respectively. While most of the codons, in case of M1, were placed near the axes, three codons for Ser (TCA, TCC and TCG) scattered apart from the axes. Similarly, for M2, all the Ser codons along with the TTN type codons for Leu scattered away from the origin around which most of the points were seen clustered. It was observed that in the 6-fold degenerate codon families, one particular codon was more favoured than others. For amino acid Ser, the codon AGT was mostly used while TCG was almost omitted. Similar was the case for Leu where CTT was extremely high and TTA was equally low. Even in Arg too, AGA was favoured at the expense of CGG, the latter being almost absent. The remaining codons were used in very low to moderate frequencies for these three amino acids. In general most of the codons were more close to the axes, which suggested that compositional properties under mutation pressure might affect the CUB. However, some codons were distantly located suggesting other factors such as natural selection might influence the CUB in M1 and M2 genes.

The axis1 generated from correspondence analysis was tested for correlation with other codon usage indices. For M1 gene, significant negative correlations ($p < .01$) of axis 1 were found with GC3 ($r = -0.328$), GC ($r = -0.383$) and Nc ($r = 0.265$). However, for M2 all correlations of axis 1 were significant (p $<$ .01) and positive with Nc ($r = 0.394$), GC ($r = 0.602$) and GC3 ($r = 0.781$). This revealed that M1 and M2 genes showed differential trends in codon usage.

13. Discussion

Studies on codon usage bias facilitate greater insights into the genomic evolution of viruses as well as their approach to homeostasis. Antigenic shift and antigenic drift are two noteworthy features of evolution in influenza viruses. While deceiving the host immune system, these viruses undergo faster evolution and overcome the selection pressure leading to invasion of host cellular machinery. The codon usage preferences in influenza viruses play a key role in adaptation of the viruses and exploitation of the host cellular components for their own replication. The genetic material of viruses, either DNA or RNA and double stranded or single stranded, varies among them and from their hosts in terms of nucleotide composition. Due to selection pressure, the directional changes resulted on codon usage profile of H1N1 viral genes, as observed by Wong et al. (2010b) (Kumar et al., 2016; Li et al., 2018; Wong et al., 2010a). In the context of nucleotide content, the coding sequences of M1 and M2 proteins could be distinguished. The nucleotide A was the most abundant in M1 dataset and T was the minimal whereas, in M2 dataset, nucleotide A was again in most abundance followed by T.

Analysing the third codon positions in M1 and M2 datasets, it was revealed that the A/T ending codons were mostly preferred. Variation in overall GC content and GC3 content between M1 dataset (49.1% and 44.5%) and M2 dataset (45.4% and 44.2%) were observed. In RNA viruses, GC content is correlated to genome polarity, codon usage bias which ultimately influences the level of gene expressivity. Auewarakul (2005) studied the correlation of GC content in RNA viruses with amino acid usage in order to comprehend the influence of GC bias on amino acid usage. The GC rich codons encoding amino acids namely glycine, alanine, arginine and proline (GARP) exhibited high Pearson correlation coefficient ($r = 0.959$, $p < .01$) with GC content (Auewarakul, 2005). Since the viral genome is AT rich, it can be explained as an activity of host selection pressure resulting in constriction on overall GC usage and G content at synonymous sites (Wong et al., 2010b). The one-way analysis of variance (ANOVA) results indicated that each coding strand of M1 and M2 datasets was significantly distinct.

Considering the third codon positions, it was mentioned that A/T ending codons were mostly preferred in M1 and M2 genes and it was further supported by the RSCU values of codons. Goni et al. (2012) found similar over-represented codons encoding the amino acids (Arg, Ser, Gly, Pro, Leu, His, Ala, Gln, Tyr) while analysing codon usage bias of 310 IAV strains. Interestingly, in the present study, the biased codons were deficient in CpG dinucleotides in M1 dataset but moderately represented in M2 dataset. Unmethylated CpG can act as an immunostimulant that activates Toll-like receptor 9 (TLR9) and subsequently triggers downstream pathways of innate immune system. Under strong selection pressure, the IAV strains might have decreased the usage of genomic CpG dinucleotides. Substitution of natural codons with synonymous ones possessing CpG dinucleotides was utilized to diminish the virulence in Poliovirus (Goñi et al., 2012). Overrepresentation of CpA and TpG associated with the underrepresentation of CpG and TpA were observed in case of a few virus families. Similar taxonomic group infected by viruses might contain heterogeneous composition of dinucleotides but viruses belonging to same family match in their dinucleotide composition (Di Giallonardo et al., 2017). To examine the hypothesis of equivalence in synonymous codon usage for 18 amino acids, G-test was performed and it indicated difference in choice of codons was statistically significant.

Codon usage bias prevails at a lower scale in both M1 and M2 genes. Since H1N1 viruses are solely dependent on host cellular system for their multiplication, the codon usage bias of viruses is influenced by the codon bias profile of their hosts (Baker et al., 2015). Similar observations were reported in previous works on IAV (Goni et al., 2012; Jenkins et al., 2001; Li et al., 2010; Zhou et al., 2005). The Nc values correlated significantly with the corresponding GC3 values (Pearson $r = 0.211$ for M1 and $r = 0.249$ for M2; $p < .01$). In M1 gene the Nc value showed significant positive correlation with overall GC composition ($r = 0.409$, p $<$ .01) and not in M2 gene. Thus, we concluded that overall GC composition plays a vital role in codon usage of the M1 genes (Li et al., 2016; Ohama et al., 1990). Viruses with similar infection symptoms might fall under a narrow range of ENC values (Zhong et al., 2012). High ENC value is also related to fidelity in replication as well as conserved genomic morphology. The contest for replication machinery might be declined due to low biasness, which is advantageous for RNA viruses (Butt et al., 2016a). Significant correlation between ENC and GC3 in both M1 and M2 datasets suggests that compositional characteristics impacted the synonymous codon preference (Alnazawi et al., 2017).

To understand the adaptation of viruses to their host cells, the codon usage distances between the viruses and their hosts were estimated in terms of squared correlation coefficients. Further, a comparison of codon usage between host cell and virus was done with graphical representation. The host cell preferred C and G ending codons while the H3N2 virus preferred A and T ending codons. A significant positive correlation of host cell with M1 and M2 genes was observed in terms of their codon usage pattern. However, considering the squared correlation coefficients, the M1 genes were found to be better adapted to their hosts. The mammalian viruses exhibited similar codon usages with respect to their hosts but showed divergent pattern from any other mammalian virus. The human infecting viruses showed resemblance with a wide array of taxonomic host groups *viz.* mammals, birds, majority of insects and plants. However, only human viruses and rat viruses are closely related in terms of their codon usage pattern (Bahir et al., 2009).

A neutrality plot was generated to analyze the magnitude of evolutionary forces acting on both the genes and to identify the dominant evolutionary force. In M2 gene significant correlation between GC12 and GC3 in M2 gene revealed the major influence of mutational pressure unlike M1 gene. However, both M1 and M2 genes were substantially driven by natural selection. Similar results were observed in ZIKV and NPV where natural selection acted as a major factor in their evolution (Zhao et al., 2016) (Butt et al., 2016b).

PR2 plot was drawn to figure out the consequences of mutation and translational selection on M1 and M2 genes. Weakly biased genes (M1 and M2) were investigated for their choice of codons with PR2 analysis. The 2-fold degenerate codons for M1 and the overall codons for M2 manifested biased trend in codon usage. Inconsistent occurrence of codons in both the genes indicated that mutational pressure and natural selection might have co-existed in establishing the codon usage profile of these genes (Butt et al., 2016a; Nasrullah et al., 2015).

MELP values revealed high degree of gene expression and inverse relationship to codon usage bias (Nc). It showed codon usage bias of M1 and M2 genes negatively correlated with codon usage bias. However, MELP did not show significant correlation with GC3 and axis1 of correspondence analysis (COA).

COA was used to gain insights into the pattern and variation of synonymous codon usage in M1 and M2 datasets. In this study, the major variation in codon usage of M1 and M2 datasets was accounted for by first axis. High preference for a single codon in case of amino acids serine, leucine and arginine was noticed. It was earlier reported that selection pressure might enhance with the level of codon degeneracy level (Whittle and Extavour, 2015). Further, the inclination towards the use of a fewer codons could be linked with the nucleotide composition of the whole genome of viruses. Earlier study revealed that AT rich genomes preferred the use of codons ending with A/U while GC rich genomes had greater preference towards the use of G/C ending codons (Roychoudhury and Mukherjee, 2013). Correlation analysis of axis1 with GC3, GC, and Nc indicated differential codon usage patterns in M1 and M2 genes, respectively. The results of the present study implied that compositional features and mutation pressure might have roles in shaping the codon usage pattern of M1 and M2 genes (D'Andrea et al., 2011; Zhao et al., 2016). Similar kind of correlations were reported in case of HAV and other RNA viruses (Chen, 2013; D'Andrea et al., 2011).

14. Conclusions

In the present study, we presented an overview of the codon usage profile for the matrix protein genes M1 and M2 of influenza A virus subtype H3N2. In continuation of our previous work (Deka and Chakraborty, 2016), here also we observed anomalous codon usage behaviour in M2 genes. Both M1 and M2 genes showed slightly different preference in composition as well as in codon choices. The dinucleotide usage differed between them, with moderately high CpG usage in M2 which is quite unusual for IAV. Extensive research on codon usage revealed low CpG usage could be a probable strategy of the viruses to escape host immune mechanism (Cheng et al., 2013; Greenbaum et al., 2008; Karlin et al., 1994). The codon usage in M1 genes seemed to be relatively better adapted to that of host conditions as evident from the coefficient of determination (squared correlation coefficient). The overall codon usage bias in present study for M1 and M2 genes was low (Nc $>$ 55) which is in accordance with the previous findings (Goni et al., 2012; Greenbaum et al., 2008; Li et al., 2010).

The correspondence analysis reflected differential codon preference in both the datasets. PR2 analysis showed that variation prevailed in the composition of amino acids encoded by synonymous codons based on the degeneracy level. This variation could be linked to the function and localization of these proteins in the virus. In M1 gene significant correlation between Nc and overall GC ($r = 0.409$, $p < .01$) indicated the possible role of GC-synonymous bias in determining the magnitude of codon bias in this gene. However, M2 gene did not exhibit such relationship. It was, thus concluded, that apart from compositional constraint, other selective forces might have played significant role in shaping the codon usage bias in M1 and M2 genes of influenza A virus.

Conflict of interest

The authors declare that no competing interest exists in this work.

Acknowledgements

The authors are grateful to Assam University, Silchar-788011, Assam, India for providing the necessary research facility. They are thankful to Mr. Arup Kr. Malakar, and Mr. Prosenjit Paul, Department of Biotechnology, Assam University for their kind support and critical comments on analysis and interpretations of results.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2018.10.015.

References

Ahn, I., Son, H.S., 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.

Ahn, I., Son, H.S., 2012. Evolutionary analysis of human-origin influenza A virus (H3N2) genes associated with the codon usage patterns since 1993. Virus Genes 44, 198–206.

Alnazawi, M., Altaher, A., Kandeel, M., 2017. Comparative Genomic Analysis MERS CoV Isolated from Humans and Camels with special Reference to Virus Encoded Helicase. Biol. Pharm. Bull. 40, 1289–1298.

Antrobus, R.D., Berthoud, T.K., Mullarkey, C.E., Hoschler, K., Coughlan, L., Zambon, M., Hill, A.V., Gilbert, S.C., 2014. Coadministration of seasonal influenza vaccine and MVA-NP + M1 simultaneously achieves potent humoral and cell-mediated responses. Mol. Ther. 22, 233–238.

Auewarakul, P., 2005. Composition bias and genome polarity of RNA viruses. Virus Res. 109, 33–37.

Bahir, I., Fromer, M., Prat, Y., Linial, M., 2009. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Mol. Syst. Biol. 5, 311.

Baker, S.F., Nogales, A., Martinez-Sobrido, L., 2015. Downregulating viral gene expression: codon usage bias manipulation for the generation of novel influenza A virus vaccines. Future Virol 10, 715–730.

Behura, S.K., Severson, D.W., 2012. Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. PLoS One 7, e43111.

Bera, B.C., Virmani, N., Kumar, N., Anand, T., Pavulraj, S., Rash, A., Elton, D., Rash, N., Bhatia, S., Sood, R., 2017. Genetic and codon usage bias analyses of polymerase genes of equine influenza virus and its relation to evolution. BMC genomics 18, 652.

Berlanda Scorza, F., Tsvetnitsky, V., Donnelly, J.J., 2016. Universal influenza vaccines: Shifting to better vaccines. Vaccine 34, 2926–2933.

Butt, A.M., Nasrullah, I., Tong, Y., 2014. Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. PLoS One 9, e90905.

Butt, A.M., Nasrullah, I., Qamar, R., Tong, Y., 2016a. Evolution of codon usage in Zika virus genomes is host and vector specific. Emerging Microbes Infections 5, e107.

Butt, A.M., Nasrullah, I., Qamar, R., Tong, Y., 2016b. Evolution of codon usage in Zika virus genomes is host and vector specific. Emerg Microbes Infect 5, e107.

Campitelli, L., Donatelli, I., Foni, E., Castrucci, M.R., Fabiani, C., Kawaoka, Y., Krauss, S., Webster, R.G., 1997. Continued evolution of H1N1 and H3N2 influenza viruses in pigs in Italy. Virology 232, 310–318.

Chan, C.-H., Lin, K.-L., Chan, Y., Wang, Y.-L., Chi, Y.-T., et al., 2006. Amplification of the entire genome of influenza A virus H1N1 and H3N2 subtypes by reverse-transcription polymerase chain reaction. J. Virol. Methods 136, 38–43.

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.

Chen, H., Sun, S., Norenburg, J.L., Sundberg, P., 2014. Mutation and selection cause codon usage and bias in mitochondrial genomes of ribbon worms (Nemertea). PLoS One 9, e85631.

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.

Coghlan, A., Wolfe, K.H., 2000. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. Yeast 16, 1131–1145.

D'Andrea, L., Pintó, R.M., Bosch, A., Musto, H., Cristina, J., 2011. A detailed comparative analysis on the overall codon usage patterns in hepatitis a virus. Virus Res. 157, 19–24.

Deka, H., 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), 1972–1982. https://doi.org/10.4014/jmb.1605.05016.

Di Giallonardo, F., Schlub, T.E., Shi, M., Holmes, E.C., 2017. Dinucleotide composition in animal RNA viruses is shaped more by virus family than host species. J. Virol. 02381–02386.

dos Reis, M., Savva, R., Wernisch, L., 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 32, 5036–5044.

Gatherer, D., McEwan, N.R., 1997. Small regions of preferential codon usage and their effect on overall codon bias-The case of the plp gene. IUBMB Life 43, 107–114.

Goni, N., Iriarte, A., Comas, V., Sonora, M., Moreno, P., Moratorio, G., Musto, H., Cristina, J., 2012. Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development. Virol. J. 9, 263.

Goñi, N., Iriarte, A., Comas, V., Soñora, M., Moreno, P., Moratorio, G., Musto, H., Cristina, J., 2012. Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development. Virol. J. 9, 263.

Gouy, M., Gautier, C., 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10, 7055–7074.

Grantham, R., Gautier, C., Gouy, M., 1980. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res. 8, 1893–1912.

Greenacre, M., Hastie, T., 1987. The geometric interpretation of correspondence analysis. J. Am. Stat. Assoc. 82, 437–447.

Greenbaum, B.D., Levine, A.J., Bhanot, G., Rabadan, R., 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. 4, e1000079.

Hammer, Ø., Harper, D., Ryan, P., 2001. PAST-PAlaeontological STatistics, ver. 1.89. Palaeontol. Electron. 4, 1–9.

Holmes, E.C., Ghedin, E., Miller, N., Taylor, J., Bao, Y., St George, K., Grenfell, B.T., Salzberg, S.L., Fraser, C.M., Lipman, D.J., 2005. Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS Biol. 3, e300.

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

Jenkins, G.M., Pagel, M., Gould, E.A., De, A.Z.P.M., Holmes, E.C., 2001. Evolution of base composition and codon usage bias in the genus Flavivirus. J. Mol. Evol. 52, 383–390.

Karlin, S., Doerfler, W., Cardon, L.R., 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.

Kliman, R.M., Irving, N., Santiago, M., 2003. Selection conflicts, gene expression, and codon usage trends in yeast. J. Mol. Evol. 57, 98–109.

Kumar, N., Bera, B.C., Greenbaum, B.D., Bhatia, S., Sood, R., Selvaraj, P., Anand, T., Tripathi, B.N., Virmani, N., 2016. Revelation of Influencing Factors in overall Codon Usage Bias of Equine Influenza Viruses. PLoS One 11, e0154376.

Lekcharoensuk, P., Nanakorn, J., Wajjwalku, W., Webby, R., Chumsing, W., 2010. First whole genome characterization of swine influenza virus subtype H3N2 in Thailand. Vet. Microbiol. 145, 230–244.

Li, Z.P., Ying, D.Q., Li, P., Li, F., Bo, X.C., Wang, S.Q., 2010. Analysis of synonymous codon usage bias in 09H1N1. Virol. Sin. 25, 329–340.

Li, X., Song, H., Kuang, Y., Chen, S., Tian, P., Li, C., Nan, Z., 2016. Genome-Wide Analysis of Codon Usage Bias in Epichloe festucae. Int. J. Mol. Sci. 17.

Li, G., Wang, R., Zhang, C., Wang, S., He, W., Zhang, J., Liu, J., Cai, Y., Zhou, J., Su, S., 2018. Genetic and evolutionary analysis of emerging H3N2 canine influenza virus. Emerg Microbes Infect 7, 73.

Lindstrom, S.E., Hiromoto, Y., Nerome, R., Omoe, K., Sugita, S., Yamazaki, Y., Takahashi, T., Nerome, K., 1998. Phylogenetic analysis of the entire genome of influenza A (H3N2) viruses from Japan: evidence for genetic reassortment of the six internal genes. J. Virol. 72, 8021–8031.

Mosier, P.D., Chiang, M.J., Lin, Z., Gao, Y., Althufairi, B., Zhou, Q., Musayev, F., Safo, M.K., Xie, H., Desai, U.R., 2016. Broad Spectrum Anti-Influenza Agents by Inhibiting Self-Association of Matrix Protein 1. Sci. Rep. 6, 32340.

Nasrullah, I., Butt, A.M., Tahir, S., Idrees, M., Tong, Y., 2015. Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. BMC Evol. Biol. 15, 174.

Novembre, J.A., 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol. Biol. Evol. 19, 1390–1394.

Ohama, T., Muto, A., Osawa, S., 1990. Role of GC-biased mutation pressure on synonymous codon choice in Micrococcus luteus, a bacterium with a high genomic GC-content. Nucleic Acids Res. 18, 1565–1569.

Palese, P., Schulman, J., 1976. Differences in RNA patterns of influenza A viruses. J. Virol. 17, 876–884.

Peiris, J.S., Guan, Y., Markwell, D., Ghose, P., Webster, R.G., Shortridge, K.F., 2001. Cocirculation of avian H9N2 and contemporary "human" H3N2 influenza A viruses in pigs in southeastern China: potential for genetic reassortment? J. Virol. 75, 9679–9686.

Rambaut, A., Pybus, O.G., Nelson, M.I., Viboud, C., Taubenberger, J.K., et al., 2008. The genomic and epidemiological dynamics of human influenza A virus. Nature 453, 615.

Roychoudhury, S., Mukherjee, D., 2013. Complex codon usage pattern and compositional features of retroviruses. Comput Math Methods Med 2013, 848123.

Shackelton, L.A., Parrish, C.R., Holmes, E.C., 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J. Mol. Evol. 62, 551–563.

Sharp, P.M., Li, W.H., 1986. Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons. Nucleic Acids Res. 14, 7737–7749.

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

Smith, B.L., Chen, G., Wilke, C.O., Krug, R.M., 2018. Avian influenza virus pb1 gene in h3n2 viruses evolved in humans to reduce interferon inhibition by skewing codon usage toward interferon-altered tRNA pools. mBio 9, e01218–e01222.

Sueoka, N., 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. 40, 318–325.

Sueoka, N., 2002. Wide intra-genomic G + C heterogeneity in human and chicken is mainly due to strand-symmetric directional mutation pressures: dGTP-oxidation and symmetric cytosine-deamination hypotheses. Gene 300, 141–154.

Supek, F., Vlahovicek, K., 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics 6, 182.

Sur, S., Sen, A., Bothra, A.K., 2007. Mutational drift prevails over translational efficiency in Frankia nif operons. Indian J. Biotechnol. 6, 321–328.

Wei, W., Guo, F.B., 2010. Strong Strand Composition Bias in the Genome of Ehrlichia canis Revealed by Multiple Methods. Open Microbiol J 4, 98–102.

Whittle, C.A., Extavour, C.G., 2015. Codon and amino acid usage are shaped by selection across divergent model organisms of the Pancrustacea. G3: Genes, Genomes. Genetics 021402, 115.

Wong, E.H., Smith, D.K., Rabadan, R., Peiris, M., Poon, L.L., 2010a. Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. BMC Evol. Biol. 10, 253.

Wong, E.H., Smith, D.K., Rabadan, R., Peiris, M., Poon, L.L., 2010b. 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, 23–29.

Wu, N.C., et al., 2014. High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. Sci. Rep. 4, 4942.

Xiong, X., et al., 2015. Structures of complexes formed by H5 influenza hemagglutinin with a potent broadly neutralizing human monoclonal antibody. Proc. Natl. Acad. Sci. U. S. A. 112, 9430–9435.

Yang, X., Luo, X., Cai, X., 2014. Analysis of codon usage pattern in Taenia saginata based on a transcriptome dataset. Parasit. Vectors 7, 527.

Zhao, Y., Zheng, H., Xu, A., Yan, D., Jiang, Z., Qi, Q., Sun, J., 2016. Analysis of codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus (NPV) and its relation to evolution. BMC Genomics 17, 677.

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.

Zhong, Q., Xu, W., Wu, Y., Xu, H., 2012. Patterns of synonymous codon usage on human metapneumovirus and its influencing factors. Biomed. Res. Int. 2012.

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.