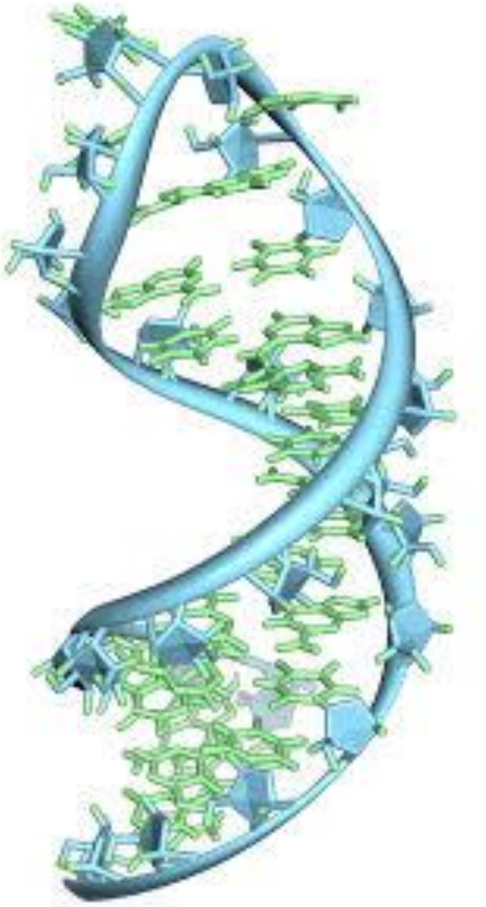
Chapter

4



RESULTS

CHAPTER 4

RESULTS

4.1 Analysis of codon usage pattern in the genomes of the IAV subtypes

4.1.1 Effective number of codons (Nc) in the IAV genes

To inspect whether IAV genes demonstrate similar codon usage pattern, the effective number of codons (Nc) for each coding sequence (cds) was estimated. The average Nc values of all the genes across the subtypes are presented in **table 4.1.1.** The values were in the range of 47.51-58.48 considering all the subtypes together. The overall values of Nc >40 indicates weak bias prevailing in the genes of IAV. Weak codon usage bias has been reported in many previous works involving IAV (Rabadan *et al.*, 2006; Dawood *et al.*, 2009; Ahn and Son, 2010; Li *et al.*, 2010; Goni *et al.*, 2012). The individual subtypes, however, showed differential magnitude of codon bias, with differences among the genes within the same subtype. Codon usage bias was inversely related to Nc value. HA gene showed the highest bias (lowest Nc) in subtypes H1N1 (Nc = 47.51 ± 1.847) and H5N1 (Nc = 49.85 ± 1.478). PB1 recorded highest bias in case of H1N2 (Nc = 49.44 ± 0.295) and H3N2 (Nc = 49.48 ± 0.437). In case of H2N2, M2 showed the highest bias with an average Nc value of 49.27 ± 3.374.

Table 4.1.1 Average Nc values of all the genes across the subtypes

Gene	H1N1	H1N2	H2N2	H3N2	H5N1
HA	**47.51 ± 1.847**	49.66 ± 0.353	49.74 ± 0.773	52.74 ± 0.640	**49.85 ± 1.478**
M1	56.76 ± 0.967	57.21 ± 2.491	56.2 ± 1.424	56.28 ± 1.706	53.60 ± 1.070
M2	57.02 ± 1.835	58.48 ± 2.818	**49.27 ± 3.374**	56.28 ± 3.922	46.81 ± 1.560
NA	52.10 ± 1.770	53.65 ± 0.413	53.15 ± 0.594	52.79 ± 0.446	50.76 ± 1.363
NP	50.72 ± 0.381	52.75 ± 0.672	53.50 ± 0.394	53.55 ± 0.373	50.83 ± 0.720
PA	54.75 ± 0.333	51.30 ± 0.734	50.75 ± 0.309	51.17 ± 0.583	52.93 ± 1.134
PB1	51.56 ± 0.801	**49.44 ± 0.295**	49.97 ± 0.464	**49.48 ± 0.437**	52.86 ± 0.706
PB2	51.28 ± 0.555	49.66 ± 0.562	50.65 ± 0.391	50.58 ± 0.737	53.16 ± 0.581

4.1.2 Relative synonymous codon usage (RSCU)

The investigation of relative synonymous codon usage (RSCU) displayed an intricate depiction of the underlying codon choice in the IAV genes across all the subtypes. Although the overall preference of codons in the genes was dissimilar, bulk of the genes preferred codons that ended with A/T. We compared the RSCU values to find out any similarity in codon preferences across the subtypes. It was observed that there was some similarity in choice of codons between H1N1 and H1N2, while H2N2, H3N2 and H5N1 showed different preference over codons (**Fig. 4.1.1**)

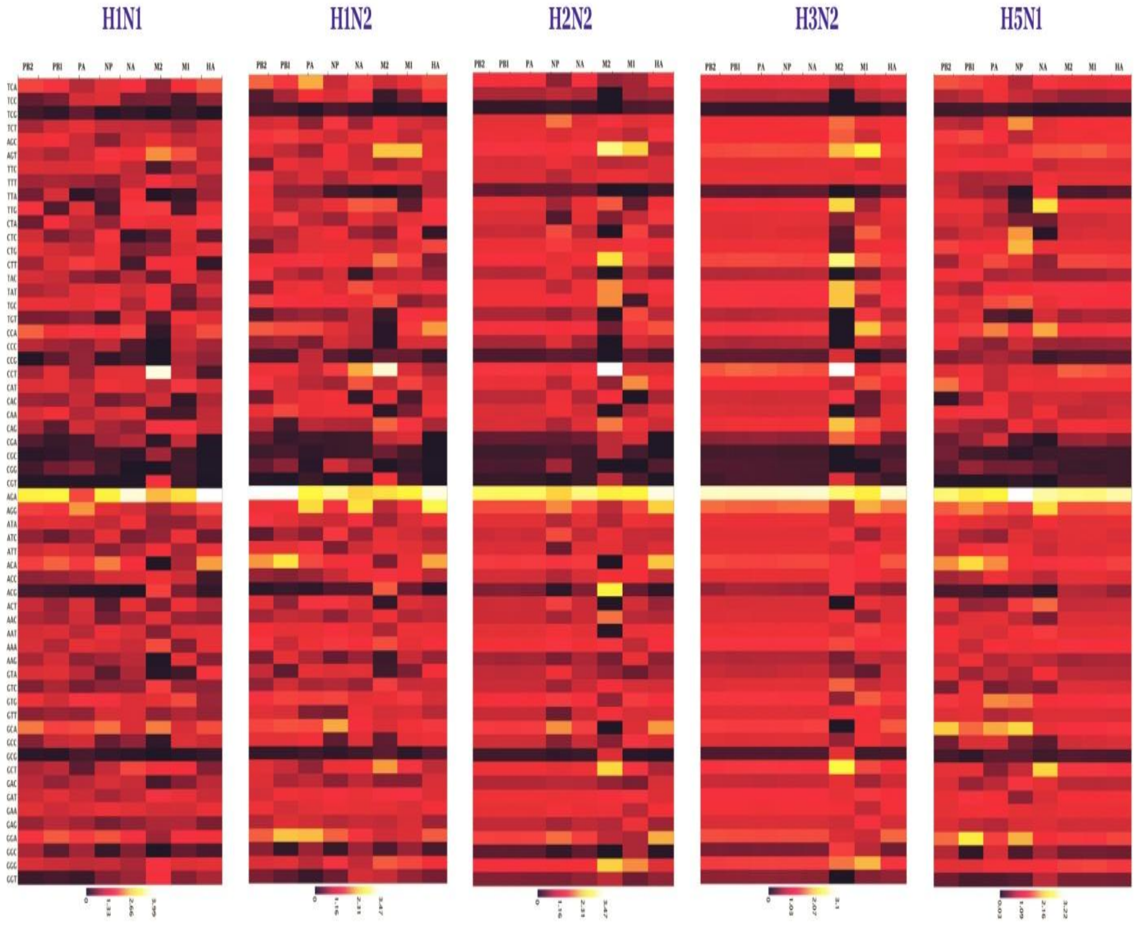


Fig. 4.1.1 Heatmap of RSCU values in the genes of the five IAV subtypes

Further, we observed dissimilar codon preference within subtypes as well with different genes opting for varied codon choice. For example, leucine in H1N1 displayed four preferred codons *i.e.* CTA (for HA and PB1), CTT (for M1 and PA), TTG (M2, NA and PB2) and CTC (for NP). Analogous pattern was noticed for rest of the subtypes too. Overall, AGA (arg), CCT (pro), ACA (thr), AGT (ser) were some of the mostly favoured codons.

We also performed the rare codons analysis using Anaconda 2.0 software (Moura *et al*., 2007). The usage of CGN and NCG type codons was drastically reduced. Low CpG usage is reported in many previous works (Karlin *et al*., 1994; Goni *et al*., 2012; Cheng *et al*., 2013). The codons CpG containing codons CGC, TCG, CGT were seldom used, while some other codons like CCG, ACG, CGA, CGG and GCG were also suppressed to a great extent, albeit non-uniformly across the subtypes (**Fig 4.1.2**).

We executed a G-test for equal usage of synonymous codons of 18 amino acids among the genes of all the subtypes. If G value is less than p<0.05, synonymous codons of that amino acid significantly differ in usage at 5%. G-test follows a chi-square distribution. Almost all the G-values in our analysis showed p value less than 0.01 meaning that the usage of synonymous codons encoding the specific amino acid is statistically different.

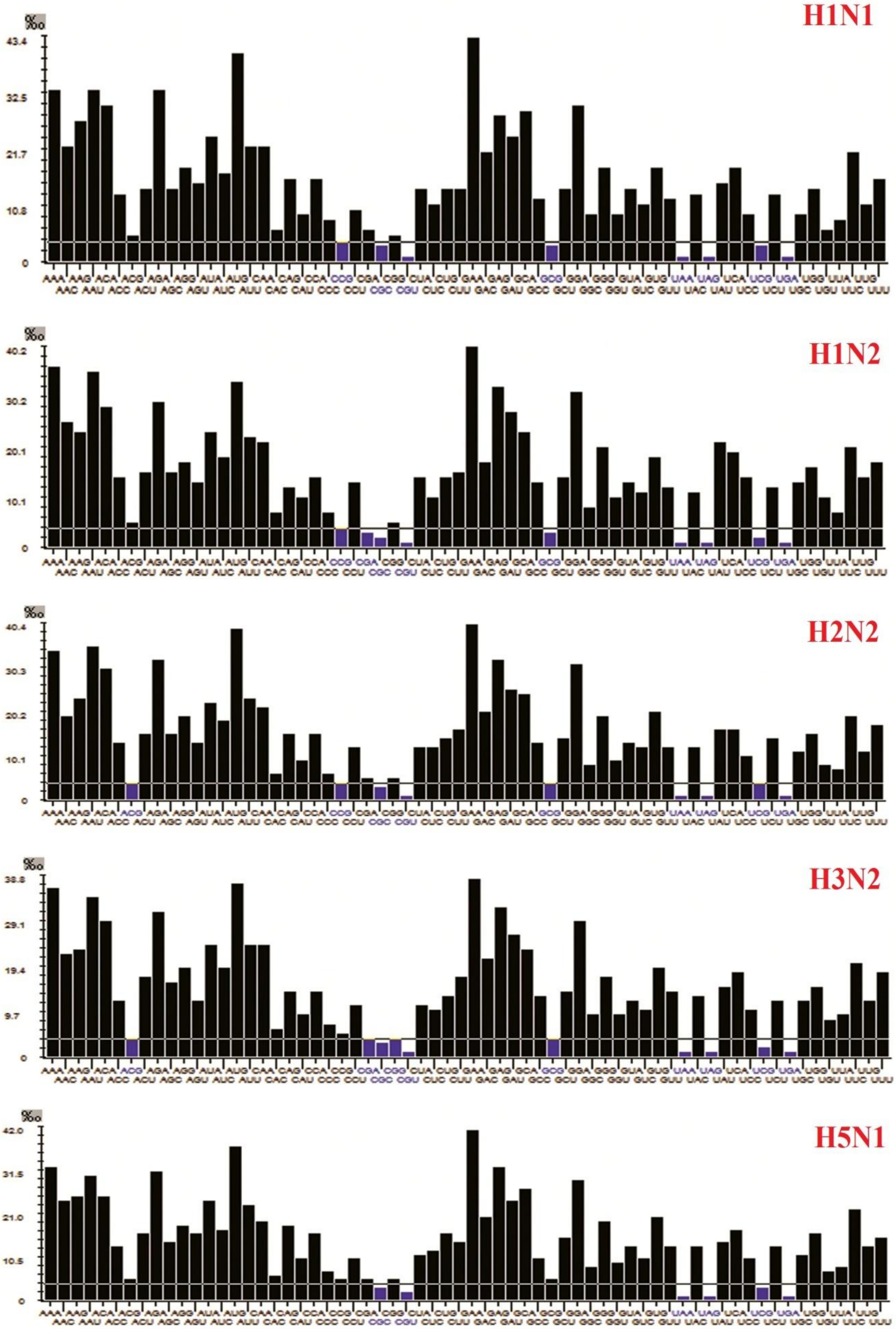


Fig. 4.1.2 Rare codons (in blue color) in 5 IAV subtypes enrolled in the study

4.1.3 Frequency of optimal codons (Fop)

The Fop is the ratio of the number of optimal codons used codons in a gene to the total count of synonymous codons in that gene. If the synonymous codon usage is random the Fop values will be close to unity. Our findings suggested that the Fop average values are below in the range of 0.22-0.40, meaning that there exists certain level of bias in these genes and the synonymous codons are not randomly used (**table 4.1.2**). The highest Fop value (0.40) was recorded for M1 gene of H5N1, while PB2 genes of all the H1N2, H2N2 and H3N2 subtypes recorded the lowest value of Fop (0.22-0.23)

Table 4.1.2 Frequency of optimal codons (Fop) in the selected IAV genes

Fop	H1N1	H1N2	H2N2	H3N2	H5N1
HA	0.28	0.30	0.28	0.29	0.31
M1	0.35	0.33	0.35	0.33	0.40
M2	0.33	0.33	0.31	0.29	0.34
NA	0.27	0.27	0.25	0.26	0.24
NP	0.28	0.27	0.29	0.29	0.30
PA	0.31	0.31	0.32	0.32	0.31
PB1	0.29	0.29	0.31	0.31	0.32
PB2	0.27	0.22	0.23	0.23	0.27

We performed correlation analysis among the codon usage parameters and Fop. Effective number of codons (Nc) showed a significant positive correlation with Fop ($r = 0.350$, $p < 0.01$). However, the rest of the codon bias parameters and gene expression measures did not show any statistically significant correlation with Fop. These results suggest that the codons in IAV subtypes are not selected for optimal translation.

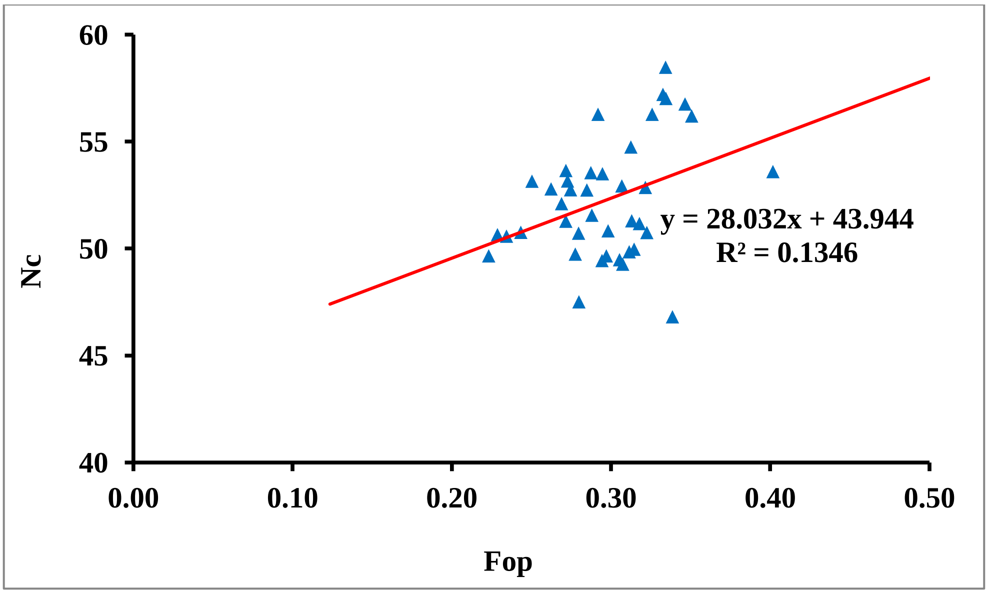


Fig. 4.1.3 Relationship between Fop and Nc values in IAV genes

4.2 Compositional features in the coding sequences of IAV genes

4.2.1 Nucleotide composition

The complete coding sequences (cds) were scrutinized for the nucleobase composition which did not reveal much deviation in compositional features among the five selected subtypes (**table 4.2.1)**. The mononucleotide usage followed the decreasing order of A>G>T>C in almost all the subtypes and across all the genes, however with varying magnitudes.

We further investigated the nucleobase choice at different codon positions. The results revealed that the 1st and 2nd codon positions followed more or less similar pattern in most of the genes among all the subtypes. However, the wobble position (3rd codon position) showed variation among the genes as well as among the subtypes. While the general order of preference for nucleobase at 3rd codon position was A3>T3>C3>G3, the gene M1 in all the subtypes except H5N1 preferred G3 to A3. M2 preferred T3 in all the subtypes except H5N1. Similarly, NA gene from H2N2 came out as exception by preferring A3 when the rest of the subtypes chose T3. The rest of the genes followed more or less similar pattern across the subtypes (**Fig 4.2.1**).

Table 4.2.1 Compositional features of the genes in the IAV subtypes

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

Note: The values in boldface indicate deviations in magnitude as compared to the values from the rest of the members in the concerned group

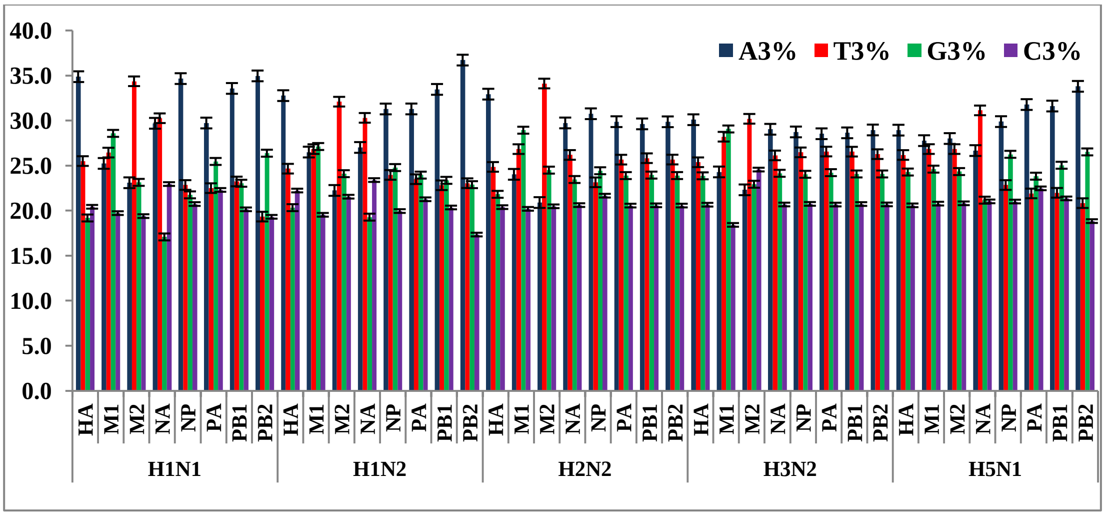


Fig. 4.2.1 Nucleobase composition at synonymous positions in the IAV genes

The genes in the IAV subtypes were found to be low in overall GC content (Mean ± SD = 44.5 ± 1.85). The overall GC content in M1 was the highest across all subtypes. The subtypes H3N2 and H5N1 recorded more or less similar usage of GC content, both overall as well as at wobble position.

4.2.2 Amino acid usage

The amino acid usage analysis revealed differential usage of amino acids in different genes. Overall most frequent amino acids include Ala, Leu, Glu and Ser which were abundant throughout the genes. Cys, Gln, His, Met, Trp and Tyr were found less frequently in the encoded proteins. Variations in amino acid usage were observed among the genes even within the same subtype.

The frequency of amino acids in H3N2 and H5N1 closely resembled each other in terms of preference as well as magnitude. However, the rest of the subtypes showed difference in usage of the amino acids with most of the variations occurring in the genes HA, M1, M2 and NA. The amino acid usage profiles of the all the IAV subtypes used in the study are summarised in **table 4.2.2.**

Table 4.2.2 Amino acid usage in the major genes of the IAV subtypes

	Gene	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
H1N1	HA	5.6	3.5	7.4	4.2	2.8	2.8	6.3	7.3	2.8	6.2	8.3	7.3	1.3	3.5	3.6	8.5	6.0	1.9	4.7	6.4
	M1	10.0	6.7	4.4	2.4	1.2	5.7	6.7	6.3	2.3	4.2	10.2	5.2	5.6	2.8	3.2	7.1	8.0	0.4	2.0	5.6
	M2	3.4	7.2	2.6	5.9	3.0	3.6	11.3	5.7	2.3	8.6	9.1	3.2	2.2	4.0	4.1	8.1	4.7	2.1	2.9	6.2
	NA	3.7	3.5	8.5	4.5	4.0	3.0	4.2	9.6	1.5	9.6	4.0	4.4	1.6	3.7	4.6	11.3	6.0	3.4	3.0	6.0
	NP	7.9	9.9	5.2	4.6	1.2	4.2	7.2	8.4	1.4	5.8	6.5	4.3	4.6	3.4	3.3	8.1	5.1	1.2	2.6	5.0
	PA	5.3	5.8	4.4	5.2	2.3	2.6	10.6	5.0	2.0	6.8	8.8	7.5	3.5	5.0	4.5	6.8	5.0	1.8	2.8	4.4
	PB1	5.3	6.6	6.5	4.5	1.3	4.1	6.2	6.2	1.3	6.3	7.4	7.0	4.9	4.4	4.2	6.6	7.8	1.2	3.3	4.9
	PB2	5.7	7.9	3.9	4.2	1.1	5.0	6.8	6.3	1.6	6.3	8.5	6.2	4.3	3.1	3.8	7.4	6.8	1.6	2.1	7.6
H1N2	HA	5.2	3.4	8.2	3.5	2.8	2.8	6.9	7.6	2.5	5.9	9.1	6.3	1.6	3.5	3.7	8.5	5.7	1.9	4.8	6.2
	M1	10.3	6.7	4.1	2.7	1.2	5.9	6.6	6.3	2.1	4.1	10.3	5.2	5.6	2.8	3.2	6.8	7.6	0.4	2.0	6.1
	M2	3.7	7.4	2.9	6.1	2.7	2.9	10.7	5.1	2.5	8.4	10.6	3.9	2.1	4.1	4.0	8.0	3.7	2.1	3.1	6.0
	NA	3.2	4.6	6.6	5.2	4.7	2.8	5.2	8.4	2.1	7.9	5.7	5.0	1.4	2.8	4.1	9.8	7.6	2.3	3.2	7.5
	NP	7.5	9.4	5.4	4.6	1.2	4.0	7.2	8.5	1.2	5.4	6.6	4.6	4.8	3.4	3.4	8.1	5.4	1.2	3.0	4.7
	PA	5.3	5.5	4.9	5.1	2.4	2.7	10.8	4.9	1.8	7.1	8.9	7.4	3.4	5.0	4.2	7.0	4.7	1.7	3.0	4.2
	PB1	5.4	6.6	6.3	4.4	1.3	4.1	6.5	6.2	1.3	6.2	7.5	7.0	4.9	4.4	4.2	6.6	7.9	1.2	3.3	4.8
	PB2	5.4	8.2	4.2	4.3	0.7	4.6	6.7	6.2	1.4	6.6	7.8	5.8	4.7	3.2	3.7	7.6	6.9	1.3	2.1	8.4
H2N2	HA	4.7	3.7	7.0	4.1	2.7	2.5	7.6	8.5	2.5	5.8	9.2	7.2	2.9	3.4	3.5	6.9	6.2	2.0	4.0	5.8
	M1	11.1	7.2	4.0	2.8	1.2	6.0	6.7	6.3	1.9	4.1	10.3	4.8	5.6	2.8	3.2	6.7	6.7	0.4	2.0	6.3
	M2	4.0	7.2	2.3	6.1	3.1	2.1	10.6	5.3	3.1	8.0	10.5	3.8	2.1	5.1	4.0	9.3	3.2	2.1	2.1	6.2
	NA	5.9	6.4	5.2	4.4	2.1	3.8	7.6	6.7	2.0	6.5	8.4	5.6	3.5	3.7	3.8	7.8	6.0	1.6	3.0	5.9
	NP	7.8	9.3	5.5	4.7	1.2	4.2	7.2	8.2	1.0	5.4	6.3	4.8	5.2	3.4	3.6	7.7	5.4	1.2	3.0	4.5
	PA	6.0	6.7	5.1	4.5	2.0	3.8	7.6	6.7	1.9	6.3	8.5	5.6	3.8	3.7	3.8	7.7	6.1	1.5	2.9	5.9
	PB1	6.1	6.7	5.0	4.5	2.0	3.8	7.7	6.8	1.9	6.3	8.5	5.5	3.7	3.6	3.8	7.7	6.0	1.5	2.9	6.0
	PB2	6.0	6.7	5.1	4.5	2.0	3.8	7.6	6.7	1.9	6.3	8.5	5.6	3.8	3.7	3.8	7.7	6.1	1.5	2.9	5.9
H3N2	HA	6.1	6.8	5.1	4.5	2.0	3.8	7.7	6.7	1.9	6.3	8.4	5.7	3.8	3.7	3.8	7.6	6.0	1.5	2.9	5.8
	M1	6.3	6.9	4.9	4.4	1.9	3.8	7.7	6.8	1.9	6.2	8.6	5.4	3.9	3.6	3.7	7.7	5.9	1.5	2.8	6.1
	M2	6.4	6.7	4.9	4.4	2.0	3.8	7.8	6.9	2.0	6.1	8.7	5.4	3.8	3.7	3.7	7.7	5.8	1.5	2.9	5.8
	NA	6.1	6.7	5.0	4.5	2.0	3.8	7.7	6.8	1.9	6.3	8.5	5.5	3.7	3.6	3.8	7.7	6.0	1.5	2.9	6.0
	NP	6.1	6.6	5.1	4.5	2.1	3.8	7.7	6.8	1.9	6.3	8.5	5.5	3.7	3.7	3.7	7.7	5.9	1.5	2.9	5.9
	PA	6.2	6.8	4.9	4.5	2.0	3.8	7.7	6.7	1.9	6.3	8.5	5.5	3.8	3.7	3.8	7.7	6.0	1.5	2.8	5.9
	PB1	6.0	6.8	5.0	4.6	2.1	3.7	7.8	6.8	1.9	6.4	8.4	5.5	3.7	3.7	3.8	7.7	5.9	1.5	2.9	5.9
	PB2	6.1	6.8	5.1	4.5	2.0	3.7	7.7	6.8	1.9	6.3	8.3	5.6	3.8	3.7	3.8	7.7	6.0	1.5	2.9	5.9
H5N1	HA	6.2	6.8	5.0	4.5	1.9	3.8	7.8	6.7	1.9	6.2	8.5	5.6	3.9	3.8	3.8	7.6	5.9	1.5	2.9	5.8
	M1	6.4	6.9	4.8	4.5	2.0	3.8	7.8	6.7	1.9	6.2	8.6	5.3	3.9	3.7	3.7	7.7	5.8	1.4	2.8	5.9
	M2	6.1	6.9	4.9	4.6	2.0	3.7	7.8	6.8	1.9	6.3	8.5	5.4	3.8	3.8	3.8	7.8	5.7	1.5	2.8	5.8
	NA	6.3	6.8	4.9	4.5	1.9	3.8	7.7	6.8	1.9	6.2	8.6	5.5	3.9	3.7	3.7	7.7	5.9	1.5	2.8	6.0
	NP	6.1	6.9	4.9	4.6	2.0	3.7	7.8	6.8	1.9	6.3	8.5	5.4	3.8	3.7	3.8	7.8	5.7	1.5	2.8	5.8
	PA	6.2	6.8	5.0	4.5	1.9	3.8	7.7	6.8	1.9	6.2	8.5	5.5	3.9	3.7	3.7	7.6	6.0	1.5	2.8	5.9
	PB1	6.2	6.8	5.0	4.5	2.0	3.8	7.7	6.8	1.9	6.2	8.5	5.5	3.8	3.7	3.8	7.7	5.9	1.5	2.9	5.9
	PB2	6.1	6.9	4.9	4.6	2.0	3.7	7.8	6.8	1.9	6.3	8.5	5.4	3.8	3.7	3.8	7.8	5.8	1.5	2.8	5.9

Note: The red blocks represent percentage of amino acids while blue color depicts the lower frequency of the amino acids. The lighter shades represent rest of the amino acid usage in between the extremes depending on their magnitude

4.2.3 Dinucleotide analysis and CpG usage

The analysis of dinucleotide usage clearly indicated a sharp diminution of the dinucleotide CpG supporting previous findings (Karlin *et al*., 1994; Goni *et al*., 2012; Cheng *et al*., 2013). The inspection of overall odds ratio values taking all the genes as a whole revealed that TpG (mean ± SD = 1.45 ± 0.084) was the most over-represented dinucleotide while CpG (0.53±0.155) was the least frequent one.

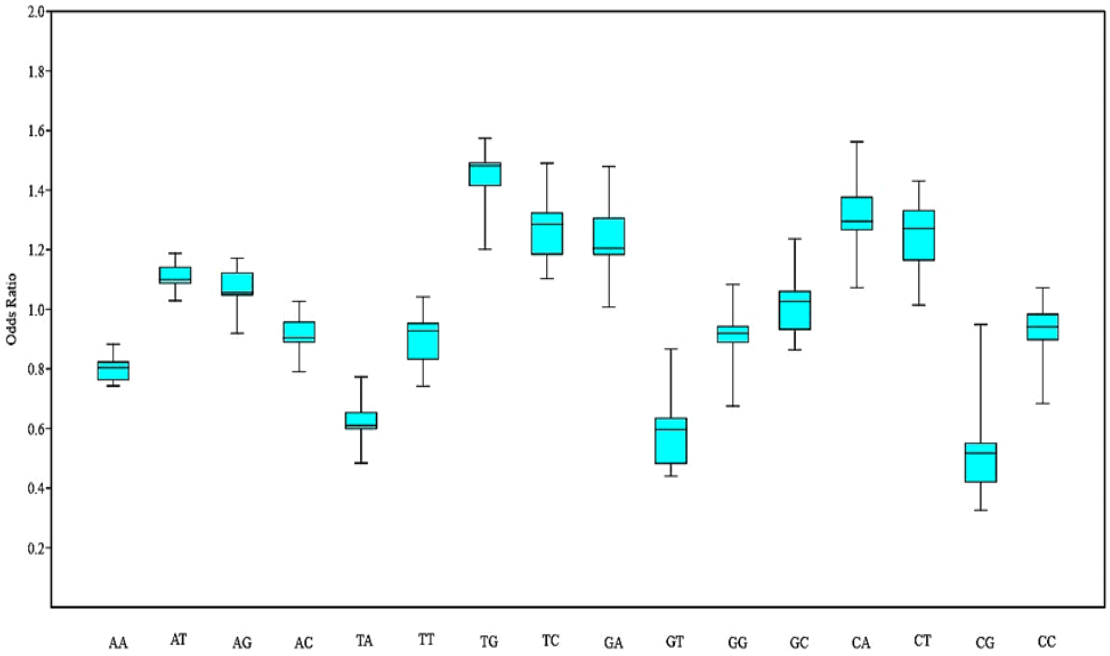


Fig. 4.2.2 Overall dinucleotide usage in the IAV genes

The dinucleotides TpC, CpA, CpT and GpA were also found in high magnitude (mean odds ratio>1.25) as opposed to the rest. GpT and TpA dinucleotides were among the under-represented (mean odds ratio<0.80) ones preceding the least CpG. Nonetheless, this observation was by and large a collective one; thus it did not display absolute uniformity *per se*, with slender deviations amid a few representative genes. To point out a few, TpG was over-represented in all genes except M2 in H1N1, H1N2 and H2N2. The dinucleotide CpA was found in high frequency in most of the genes with a few exceptions *i.e.* M2 in all the genes, NA in H1N1 and PA in H5N1. CpG was the most depleted in all but H5N1. Again M2 gene proved to be an exception with moderate usage of CpG. TpA was uniformly low in all the genes across the subtypes. Barring M2 (H1N1 And H2N2), and NA (H1N2, H5N1) genes, GpT was at a low level in all other genes.

Table 4.2.3 Dinucleotide usage in the IAV genes

Subtype	Gene	AA	AT	AG	AC	TA	TT	TG	TC	GA	GT	GG	GC	CA	CT	CG	CC
H1N1	HA	0.81	1.12	1.02	0.99	0.77	0.85	1.44	1.17	1.04	0.66	1.06	0.98	1.41	1.11	0.39	0.99
	M1	0.79	1.03	1.15	0.97	0.67	0.77	1.48	1.14	1.12	0.63	0.89	1.16	1.37	1.41	0.48	0.78
	M2	0.78	1.12	1.16	0.82	0.66	0.82	1.20	1.40	1.36	0.87	0.68	1.09	1.25	1.10	0.93	0.89
	NA	0.88	1.15	0.92	0.99	0.72	0.83	1.43	1.24	1.02	0.78	1.05	0.91	1.46	1.08	0.43	0.93
	NP	0.83	1.19	1.12	0.79	0.49	0.87	1.48	1.38	1.26	0.48	0.90	1.03	1.35	1.17	0.49	1.01
	PA	0.84	1.12	1.06	0.91	0.56	0.97	1.52	1.16	1.34	0.48	0.89	1.01	1.22	1.27	0.49	1.04
	PB1	0.78	1.14	1.08	1.02	0.68	0.93	1.46	1.18	1.24	0.47	1.00	0.91	1.35	1.19	0.37	0.98
	PB2	0.82	1.07	1.17	0.93	0.61	0.95	1.42	1.25	1.20	0.60	0.88	0.93	1.42	1.09	0.42	0.94
H1N2	HA	0.84	1.13	0.97	0.97	0.70	0.82	1.54	1.21	1.10	0.68	1.08	0.86	1.36	1.17	0.34	1.04
	M1	0.81	1.06	1.12	0.92	0.66	0.81	1.48	1.16	1.06	0.63	0.92	1.18	1.41	1.35	0.46	0.84
	M2	0.77	1.11	1.10	0.89	0.60	0.87	1.25	1.40	1.44	0.79	0.69	1.06	1.17	1.22	0.94	0.78
	NA	0.85	1.15	0.95	0.95	0.74	0.78	1.54	1.10	1.01	0.83	1.03	1.01	1.43	1.12	0.38	1.04
	NP	0.85	1.19	1.07	0.86	0.53	0.82	1.41	1.49	1.31	0.47	0.88	0.94	1.26	1.38	0.55	0.90
	PA	0.82	1.08	1.12	0.90	0.55	1.04	1.43	1.22	1.34	0.48	0.90	1.01	1.28	1.21	0.47	1.07
	PB1	0.81	1.11	1.08	1.00	0.65	0.94	1.46	1.21	1.24	0.47	1.01	0.91	1.34	1.20	0.39	0.97
	PB2	0.80	1.05	1.13	1.03	0.65	0.99	1.35	1.20	1.21	0.61	0.91	0.87	1.43	1.01	0.48	1.00
H2N2	HA	0.82	1.17	0.98	0.96	0.62	0.86	1.57	1.23	1.19	0.57	1.03	0.90	1.40	1.18	0.32	1.05
	M1	0.82	1.12	1.15	0.83	0.65	0.74	1.51	1.17	1.09	0.62	0.88	1.20	1.38	1.43	0.41	0.90
	M2	0.81	1.04	1.14	0.89	0.61	0.91	1.21	1.39	1.48	0.83	0.67	1.09	1.07	1.35	0.95	0.68
	NA	0.77	1.10	1.06	0.90	0.63	0.92	1.48	1.28	1.19	0.61	0.94	1.04	1.30	1.25	0.53	0.95
	NP	0.82	1.17	1.05	0.92	0.51	0.80	1.47	1.48	1.30	0.46	0.91	0.93	1.29	1.42	0.53	0.82
	PA	0.76	1.09	1.06	0.91	0.62	0.93	1.49	1.30	1.20	0.59	0.94	1.03	1.30	1.28	0.52	0.95
	PB1	0.77	1.09	1.05	0.90	0.61	0.93	1.49	1.31	1.20	0.60	0.93	1.03	1.29	1.28	0.53	0.95
	PB2	0.76	1.09	1.06	0.91	0.62	0.93	1.49	1.30	1.20	0.59	0.94	1.03	1.30	1.28	0.52	0.95
H3N2	HA	0.77	1.10	1.06	0.90	0.61	0.93	1.49	1.30	1.21	0.58	0.94	1.03	1.29	1.27	0.52	0.96
	M1	0.81	1.12	1.15	0.83	0.63	0.83	1.47	1.15	1.06	0.62	0.91	1.24	1.46	1.34	0.39	0.88
	M2	0.82	1.06	1.04	0.97	0.56	0.98	1.29	1.32	1.48	0.76	0.71	1.02	1.09	1.27	0.95	0.72
	NA	0.76	1.09	1.05	0.90	0.61	0.94	1.49	1.31	1.20	0.60	0.93	1.05	1.28	1.30	0.54	0.94
	NP	0.75	1.09	1.05	0.90	0.61	0.93	1.49	1.31	1.20	0.60	0.93	1.06	1.28	1.30	0.55	0.94
	PA	0.75	1.09	1.05	0.89	0.60	0.94	1.49	1.32	1.20	0.60	0.92	1.06	1.27	1.31	0.56	0.93
	PB1	0.75	1.09	1.05	0.89	0.60	0.95	1.49	1.32	1.21	0.60	0.92	1.05	1.27	1.31	0.56	0.93
	PB2	0.76	1.09	1.05	0.90	0.60	0.94	1.49	1.31	1.21	0.59	0.93	1.05	1.28	1.30	0.54	0.94
H5N1	HA	0.75	1.08	1.06	0.89	0.60	0.95	1.49	1.31	1.21	0.59	0.93	1.06	1.27	1.31	0.55	0.94
	M1	0.74	1.09	1.05	0.88	0.60	0.95	1.49	1.33	1.21	0.60	0.91	1.08	1.25	1.34	0.58	0.91
	M2	0.74	1.09	1.05	0.88	0.60	0.97	1.49	1.34	1.21	0.60	0.91	1.06	1.25	1.33	0.59	0.90
	NA	0.87	1.17	1.00	0.96	0.69	0.84	1.40	1.18	1.04	0.81	1.00	0.97	1.56	1.08	0.37	0.98
	NP	0.77	1.16	1.17	0.91	0.48	0.84	1.39	1.45	1.39	0.44	0.88	0.90	1.30	1.39	0.51	0.90
	PA	0.85	1.15	1.04	0.89	0.52	1.04	1.52	1.21	1.35	0.44	0.88	1.01	1.22	1.20	0.55	1.06
	PB1	0.79	1.14	1.06	1.01	0.61	0.97	1.46	1.18	1.31	0.45	0.96	0.90	1.31	1.16	0.50	0.97
	PB2	0.81	1.13	1.09	0.95	0.64	0.93	1.39	1.25	1.18	0.57	0.91	0.94	1.37	1.10	0.52	0.98

Note: The light red shaded units represent the over-represented (odds ratio>1.25) and under-represented (odds ratio<0.8) dinucleotides respectively; while the green shaded units show the dinucleotides in normal range (0.8>odds ratio<1.25)

4.2.4 Neutrality plot analysis in the IAV genes

A scatter plot of average GC content in the first two codon positions (GC12) along the ordinates and the GC3 content along the abscissa is used as a predictor of the mutation and selection equilibrium in codon usage bias analysis (Sueoka, 1988).

The neutrality plot is useful in getting insights into the possible interplay between the mutational and selective forces exerting in an organism’s codon usage pattern. It has been proposed that, a correlation of statistical significance between GC12 and GC3 plus a linear regression line with a slope close to unity is suggestive of mutational pressure being the principal evolutionary force. Selection in opposition to mutation is believed to be operational in case of weak correlation (Sueoka, 1988; Sueoka, 1995). A slope below 1 in the regression line indicates a propensity of non-neutral mutational pressure.

To divulge any links among the three codon positions, we created neutrality plots (GC12 vs. GC3) for each IAV subtypes (**Fig. 4.2.3**). We observed statistically significant positive correlation (Karl Pearson) between GC12 and GC3 contents of genes in all the subtypes except H1N1**.** The slopes however varied in magnitude. The results suggest that mutational constraint was the prominent evolutionary force inflicting codon usage bias in most of the IAV genes. The codon bias of H1N1 subtype however seems to be under selective forces as neutrality analysis suggested relative effect of mutational pressure is very weak.

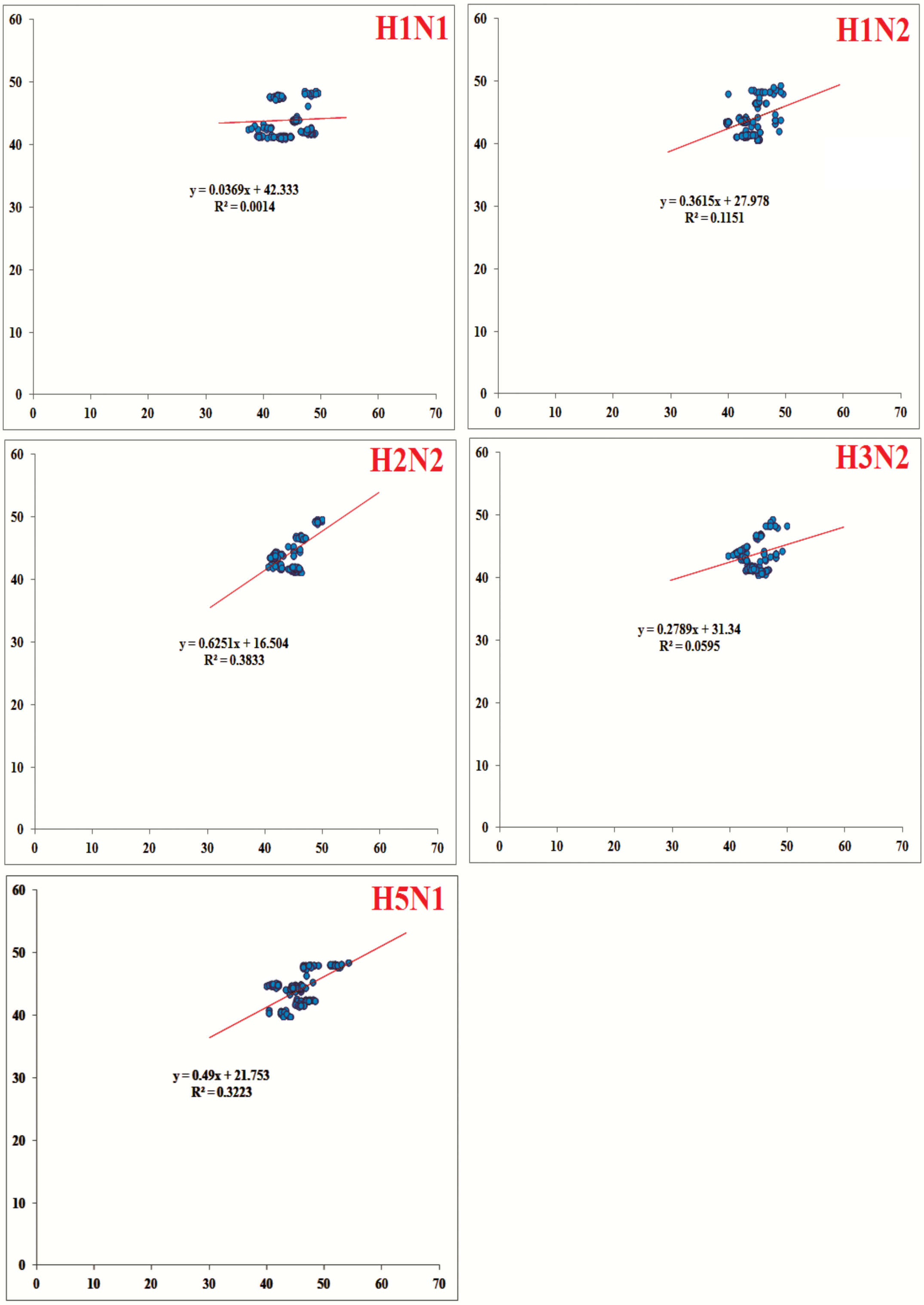


Fig. 4.2.3 Neutrality plot (GC12 vs GC3) analysis in the five IAV subtypes

4.3 Comparison of the codon usage patterns across the IAV subtypes

4.3.1 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 analysis and examined the alliance between purines (A and G) and pyrimidines (C and T). To avoid asymmetry of data in our analysis, we left out the three stop codons, codons for Met, 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 (Chen, 2013).

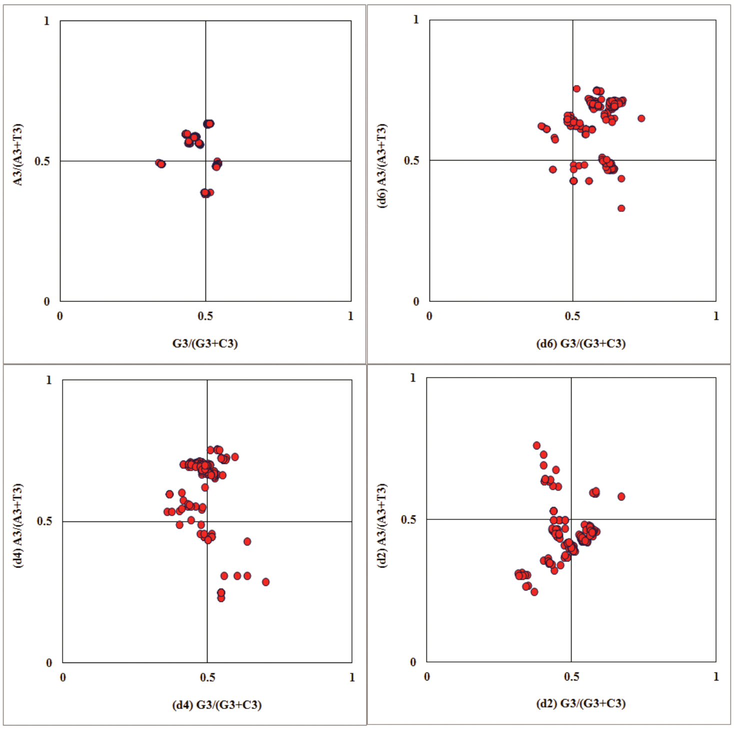


Fig. 4.3.1.1 PR2 bias analysis in H1N1 showing the overall as well as degeneracy level-specific PR2-fingerprints

Generally the PR2 analyses in most of previous reports considered only the tetra-fold (including four codons from the hexa-fold amino acids) degenerate codons (Wei and Guo, 2010; Chen *et al*., 2014; Yang *et al*., 2014). However, the quantity of 2-fold codons is quite a big figure to be ignored from the analysis. Not including Met and Trp, the 2-fold degenerate codons encode half the amino acids. Therefore, to find out any significant difference between PR2 bias fingerprints based on degeneracy level, we separated the dataset for PR2 analysis into four groups. All the subtypes demonstrated certain level of bias. The allocation of the points around the midpoint in the plot hinted towards a weak PR2-bias towards the purines at synonymous sites (3rd codon position) when we considered 58 codons as a single group. However, the scenario changed more or less when we considered the genes based on degeneracy level (**Fig. 4.3.1.1 – 4.3.1.5)**.

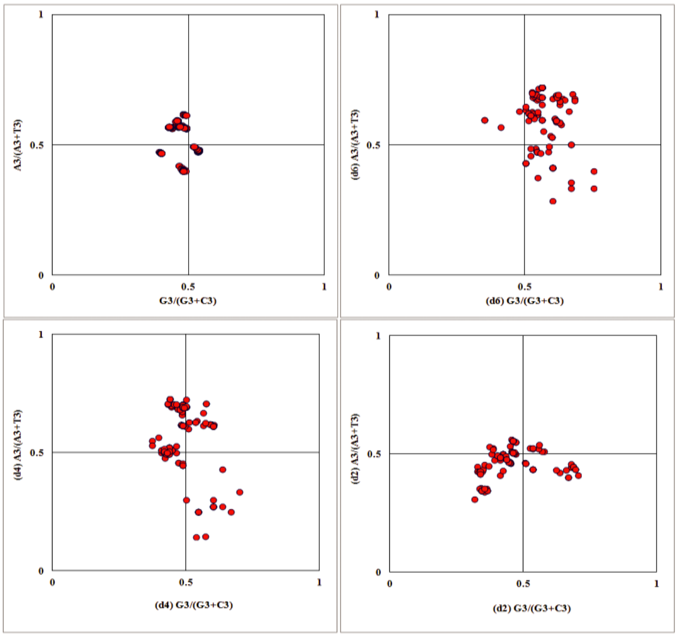


Fig. 4.3.1.2 PR2 bias analysis in H1N2 showing the overall as well as degeneracy level-specific PR2-fingerprints

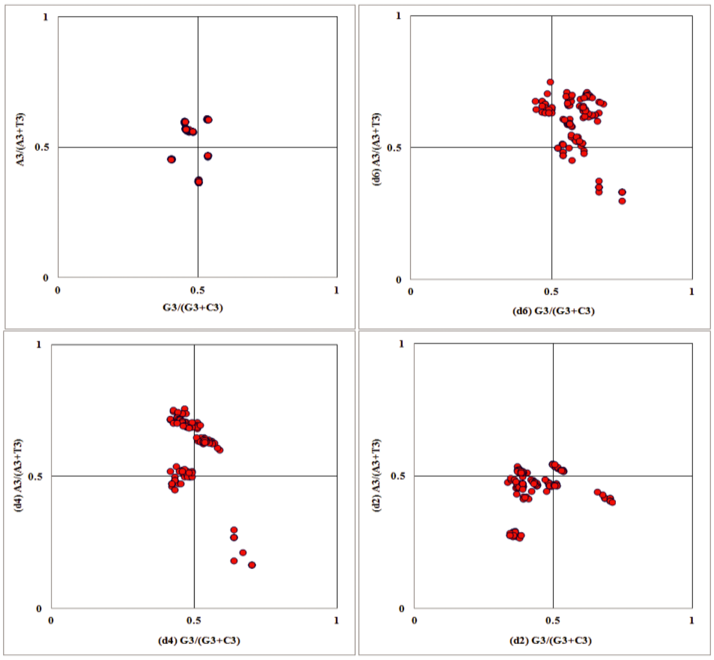


Fig. 4.3.1.3 PR2 bias analysis in H2N2 showing the overall as well as degeneracy level-specific PR2-fingerprints

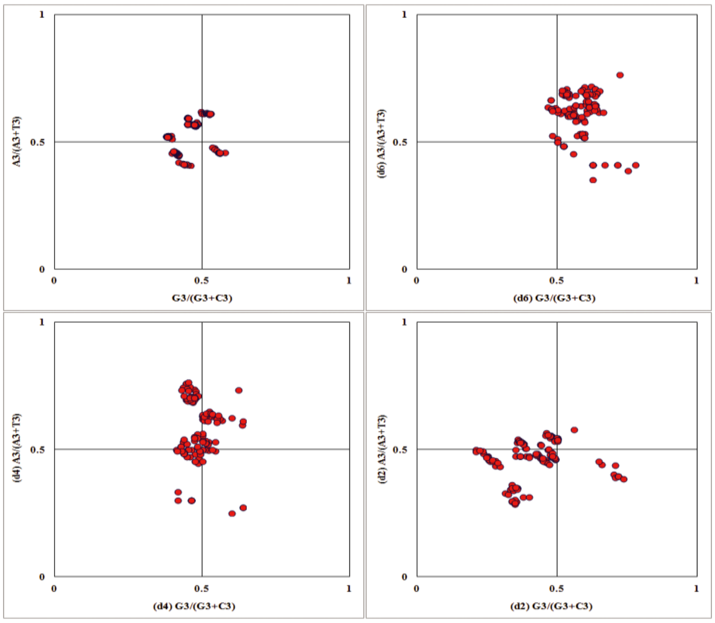


Fig. 4.3.1.4 PR2 bias analysis in H3N2 showing the overall as well as degeneracy level-specific PR2-fingerprints

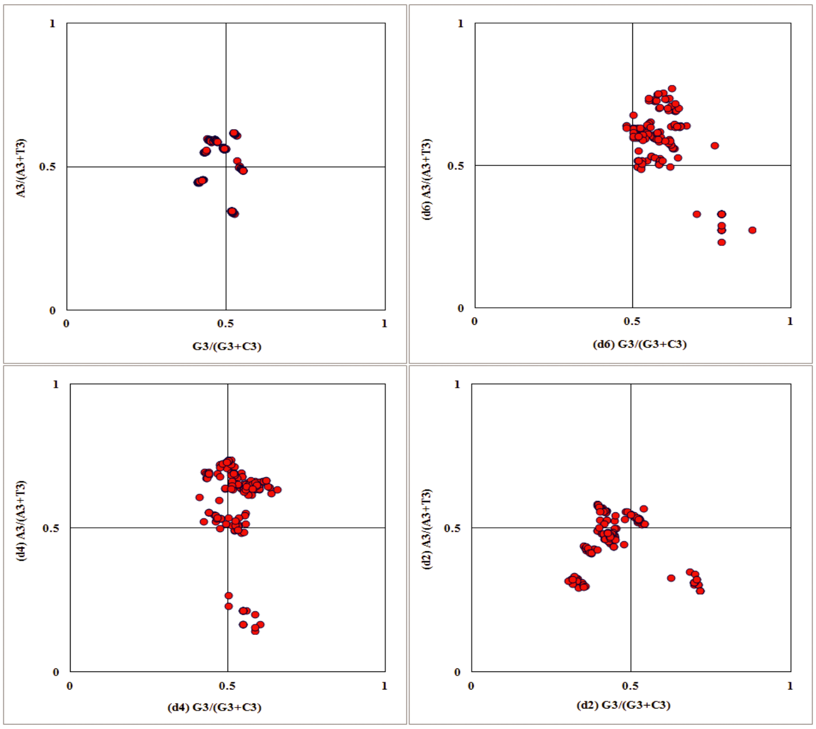


Fig. 4.3.1.5 PR2 bias analysis in H5N1 showing the overall as well as degeneracy level-specific PR2-fingerprints

The purines (A and G) seem to be preferred over the pyrimidines (C and T) at the silent sites in all the subtypes, nevertheless with varying magnitude. The tetra-fold codons in all subtypes except H5N1 showed an inclination towards A/C and to some extent G as well. H5N1 clearly favoured usage of G/A at silent sites of 4-fold codons. The 2-fold codons mostly preferred T/C in all the subtypes.

4.3.2 Correspondence analysis

Correspondence analysis (CA) is a multivariate ordination technique frequently used for its efficient way of reducing multi-dimensional data in planar form (Greenacre and Hastie, 1987). The CA allows the distribution of genes in the scatter plot based on their corresponding selection of codons, which help reveal the underlying influence on CUB. To determine the trend of codon usage variation among the IAV genes we employed CA on RSCU values. All gene data were taken as a single dataset and the two central axes were plotted in a two-dimensional scatter plot.

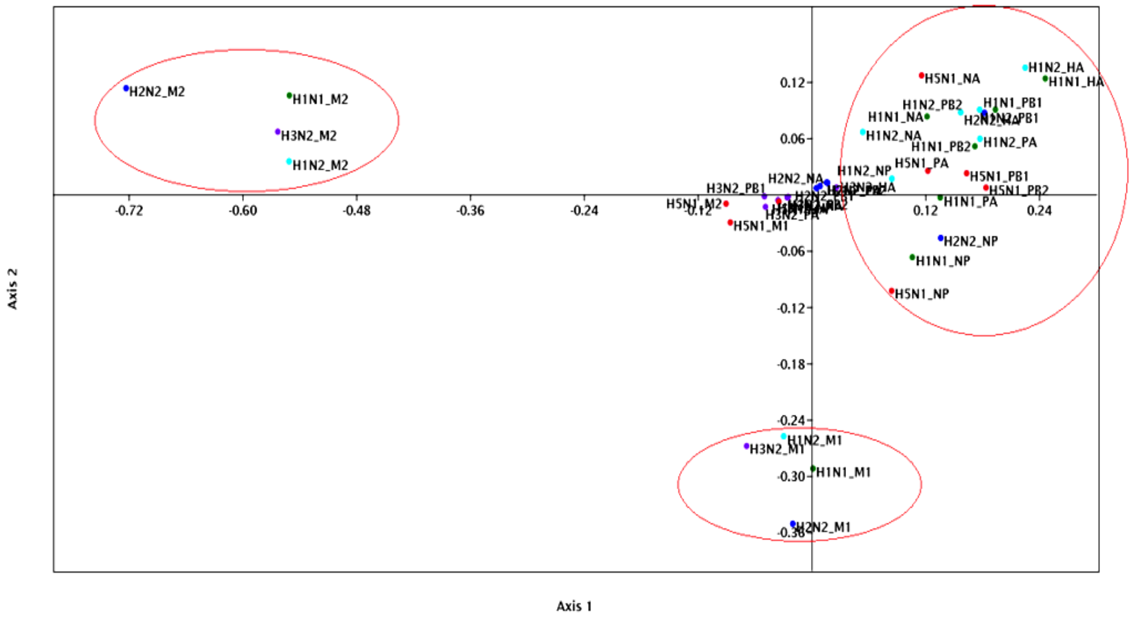


Fig. 4.3.2.1 Row plot of correspondence analysis showing allocation of genes based on their corresponding codon choices

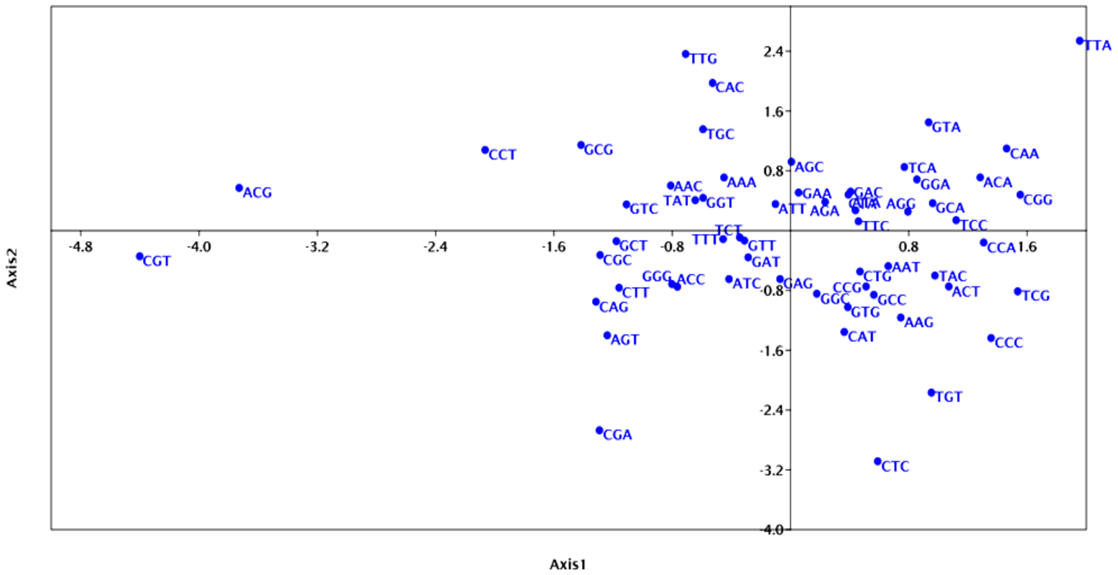


Fig. 4.3.2.2 Column plot of correspondence analysis showing allocation of the 59 codons

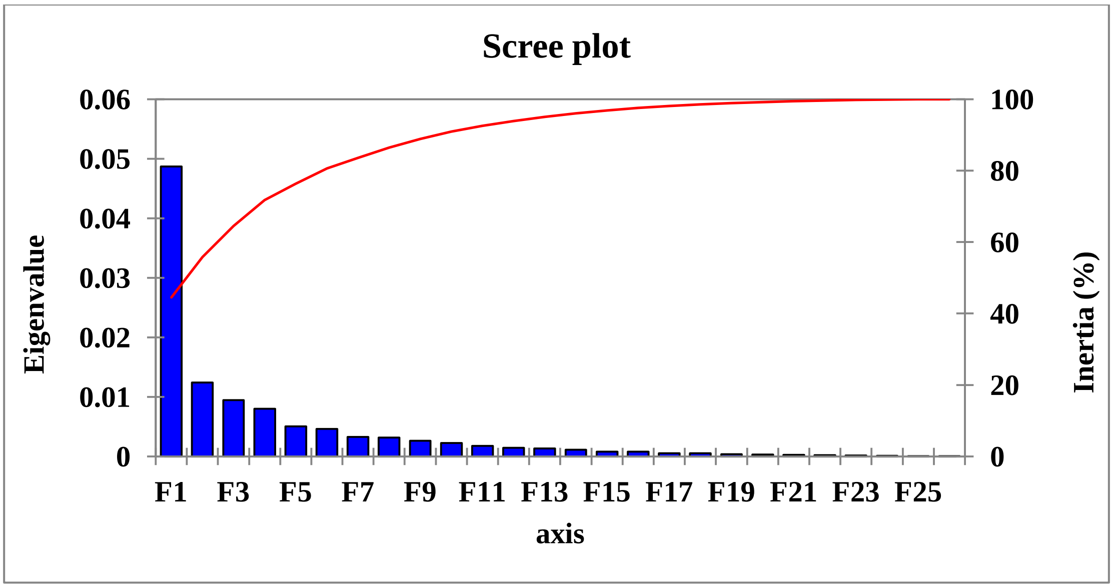


Fig. 4.3.2.3 Scree plot from correspondence analysis showing the contribution of the individual axes

The first two major axes elucidated 55.7% of the total variations with individual contributions of 44.5% and 11.3% by axis1 and axis2 respectively. The distribution of the genes in the CA plot shows the presence of at least three clusters marked in circles (**Fig. 4.3.2.1**). Genes M1 and M2 for all subtypes except H5N1 clustered separately from the rest of the genes.

4.3.3 Codon-pair context analysis

A crucial but not very much expansively studied aspect in CUB studies is the analysis of codon-pair context in the genes. Codon usage and codon-pair context are subject to the selective evolutionary forces. At the translational level, codon contexts play important role in speed and precision in the decoding fidelity of mRNA (Boycheva *et al*., 2003; Ogle and Ramakrishnan, 2005). Here, in quest for possible bias in the usage of codon-context, Anaconda v2.0 software was employed. We executed a comprehensive analysis to compare codon-pair associations using a 64x64 codon-pair contingency chart (Moura *et al*., 2005).

As per the results, the individual contexts displayed variations among the IAV subtypes. The average matrix plot of 5'context, taking into account all the genes together, showed patches of good as well as bad contexts, marked in yellow and blue circles in the plot (**Fig. 4.3.3.1)**.

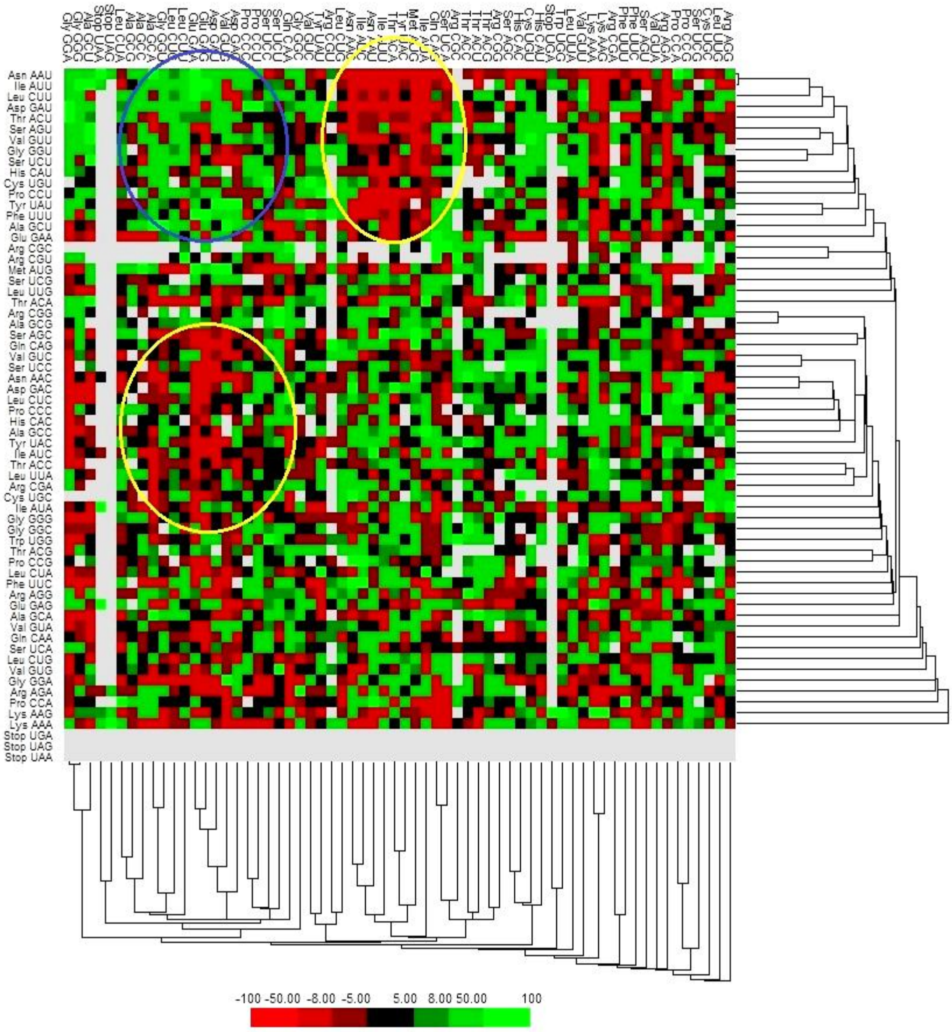


Fig. 4.3.3.1 Average matrix plot of 5' codon-pair contexts (64x64) in the IAV genes

The plot suggested the prevalence of varying preferences for codon contexts of dissimilar make-up in the genes. Contexts of NNC-GNN and NNT-ANN type were used very less frequently (**table 4.3.3.1**). Amino acid pairs like Arg-Gly, Glu-Lys, Ser-Gly, Ser-Ser were found to be used preferentially across the subtypes, however, with differential magnitudes. There was high occurrence of tandem repeats of amino acids. **Table 4.3.3.2** displays the tandem repeats (double and triple) of amino acids across the various IAV subtypes. Leu, Arg and Ala were some of the highest occurred amino acids in tandem repeats.

Table 4.3.3.1 Ten most preferred contexts in the IAV subtypes (The common contexts across the subtypes are marked bold in red)

H1N1		H1N2		H2N2		H3N2		H5N1	
Arg -> Gly	**623**	Leu -> Leu	479	Gly -> Thr	489	**Ser -> Ser**	**597**	**Ser -> Ser**	**556**
Glu -> Ser	566	**Ser -> Ser**	**477**	Ile -> Glu	475	Ile -> Glu	583	Glu -> Ser	506
Ser -> Ser	**547**	Ile -> Glu	423	**Arg -> Gly**	**473**	Gly -> Thr	572	Leu -> Leu	505
Arg -> Thr	545	Lys -> Leu	409	Arg -> Asn	469	**Arg -> Gly**	**522**	Leu -> Thr	485
Ile -> Glu	533	**Glu -> Lys**	**403**	Glu -> Leu	457	**Glu -> Lys**	**522**	Glu -> Leu	478
Ser -> Gly	**513**	**Ser -> Gly**	**365**	**Ser -> Ser**	**455**	Glu -> Leu	506	**Glu -> Lys**	**473**
Lys -> Leu	513	Glu -> Leu	360	**Glu -> Lys**	**450**	Leu -> Leu	506	Arg -> Arg	472
Leu -> Glu	509	**Arg -> Gly**	**353**	**Ser -> Gly**	**450**	**Ser -> Gly**	**499**	**Ser -> Gly**	**467**
Arg -> Arg	509	Leu -> Glu	353	Glu -> Glu	446	Ile -> Arg	483	Arg -> Thr	463
Glu -> Lys	**506**	Gly -> Lys	353	Leu -> Leu	433	Ser -> Ile	479	**Arg -> Gly**	**458**

Table 4.3.3.2 Frequency of amino acids as tandem repeats in the IAV subtypes

	H1N1		H1N2		H2N2		H3N2		H5N1	
AA	**Rep2**	**Rep3**	**Rep2**	**Rep3**	**Rep2**	**Rep3**	**Rep2**	**Rep3**	**Rep2**	**Rep3**
Lys	309	85	169	35	189	30	210	38	230	34
Asn	222	0	212	1	160	5	209	2	179	0
Thr	249	16	216	6	278	24	301	0	198	1
Arg	445	32	261	12	276	28	330	29	325	70
Ser	367	90	270	29	274	62	312	88	441	54
Ile	220	9	146	1	206	18	226	10	325	0
Met	128	30	52	16	126	15	115	19	107	14
Phe	149	0	92	0	123	0	140	0	156	0
Tyr	30	0	35	0	15	0	20	0	20	0
Cys	30	0	17	0	15	0	19	0	14	0
Trp	48	0	17	0	50	0	48	0	34	0
Leu	447	17	451	14	403	15	468	19	445	30
Pro	69	0	79	0	78	0	90	0	75	0
His	26	0	36	0	32	0	41	0	34	0
Gln	132	0	70	0	89	0	107	0	89	0
Val	222	18	168	0	198	15	240	18	183	18
Ala	347	18	185	1	258	15	292	19	271	27
Gly	286	0	150	0	227	0	220	0	210	0
Asp	119	21	137	0	186	0	152	0	73	0
Glu	457	18	347	0	380	33	439	19	415	20

Note: AA = Amino acids, Rep2 = double tandem repeats of AA, Rep3 = triple tandem repeats of AA

We analyzed the preference of nucleotides based on the dinucleotides consisting of nucleotide at 3rd codon position and nucleotide at 1st codon position of any adjacent codons. Dinucleotide analysis revealed that TpG (UpG) and CpA were the most favoured dinucleotides, while CpG and GpT (GpU) were the least preferred ones. We tried to find out whether these dinucleotides at the junction of two codons show any particular preference of nucleotides upstream and downstream of the respective dinucleotides. There was a clear preference for nucleobase A at two positions downstream and two upstream positions of the dinucleotide TpG (UpG) as well as CpA. For dinucleotides CpG and GpT (GpU) the nucleobase preference varied among the different subtypes (**Fig. 4.3.3.2**).

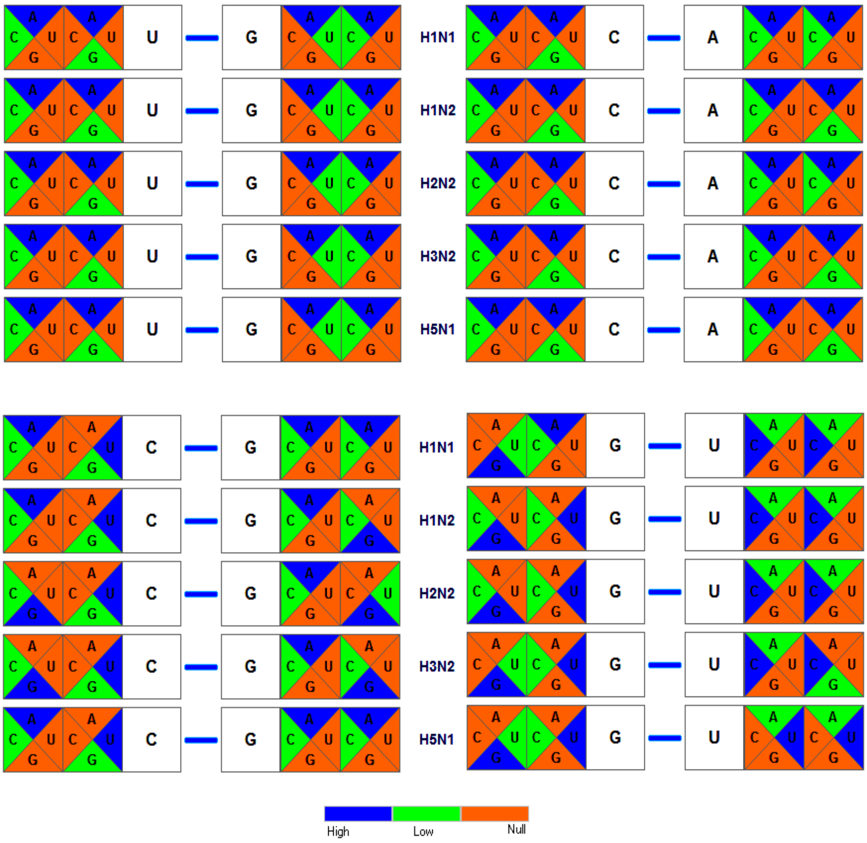


Fig. 4.3.3.2 Codon pair context preferences based on the dinucleotides at the junction of two adjacent codons

We also compared the subtypes against each other for codon context patterns which presented more or less similar pattern in all the cases **(Fig. 4.3.3.3)**.

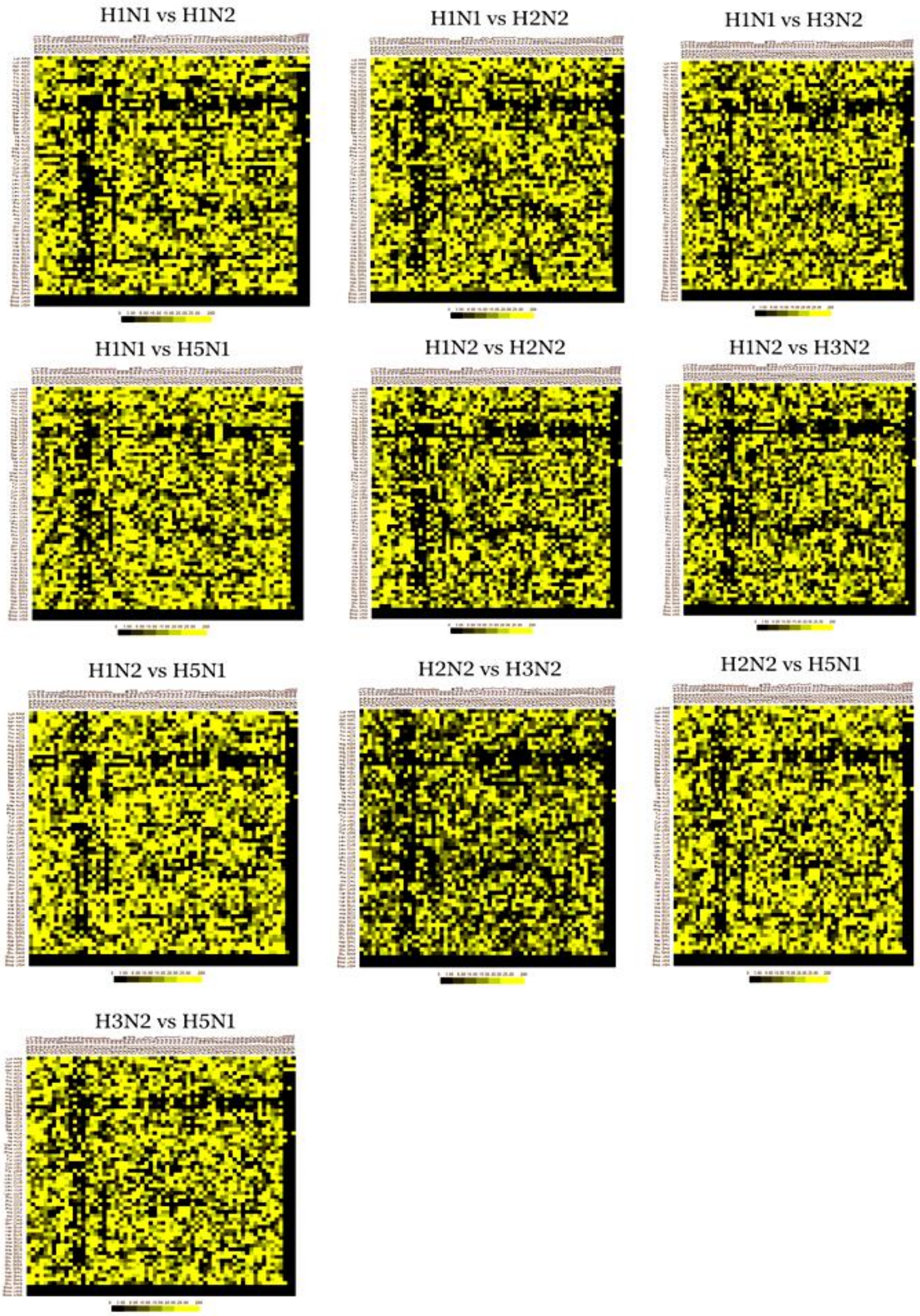


Fig. 4.3.3.3 Comparison of the codon-pair contexts between the IAV subtypes

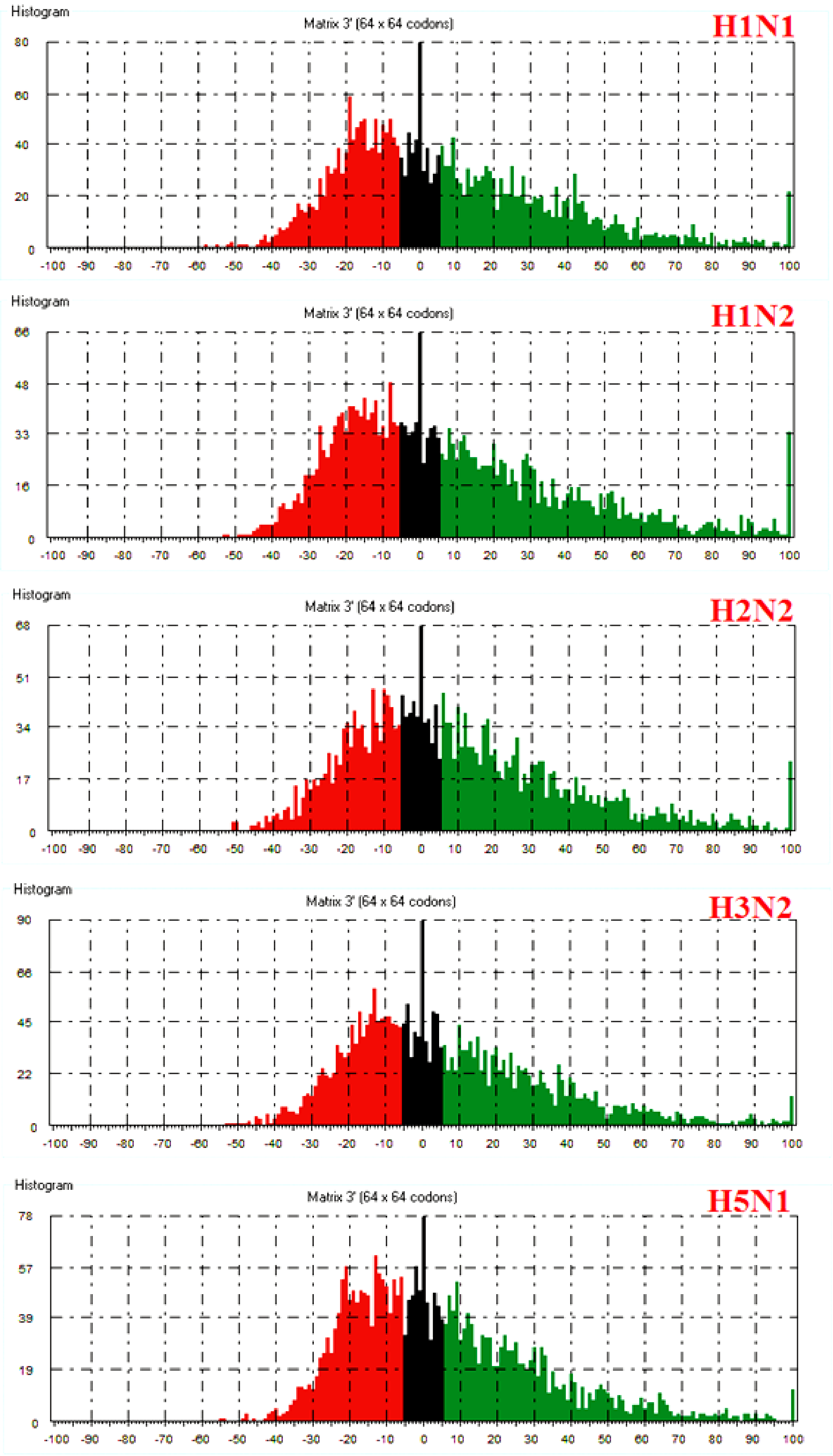


Fig. 4.3.3.4 Histogram of residuals (3' contexts) of the preferred (green) as well as rejected (red) contexts among the subtypes of IAV

4.4 Prediction of the expression of the IAV genes

4.4.1 Codon adaptation index

It is envisaged that the choice for particular codon has association with the expressivity of the genes (Sharp and Li, 1987). To find out such bias and to carry out a predictive estimation of gene expressivity, the measure codon adaptation index (CAI) was employed for each gene. CAI analysis showed a mean value of 0.83 with a standard deviation of 0.154. Interestingly, M2 gene in each subtype showed a stern decline in expressivity as reflected by mean CAI value of 0.48±0.158. H5N1 subtype differed for M2 expressivity (Mean CAI of 0.77) similar to the mean CAI values of the rest of the genes.

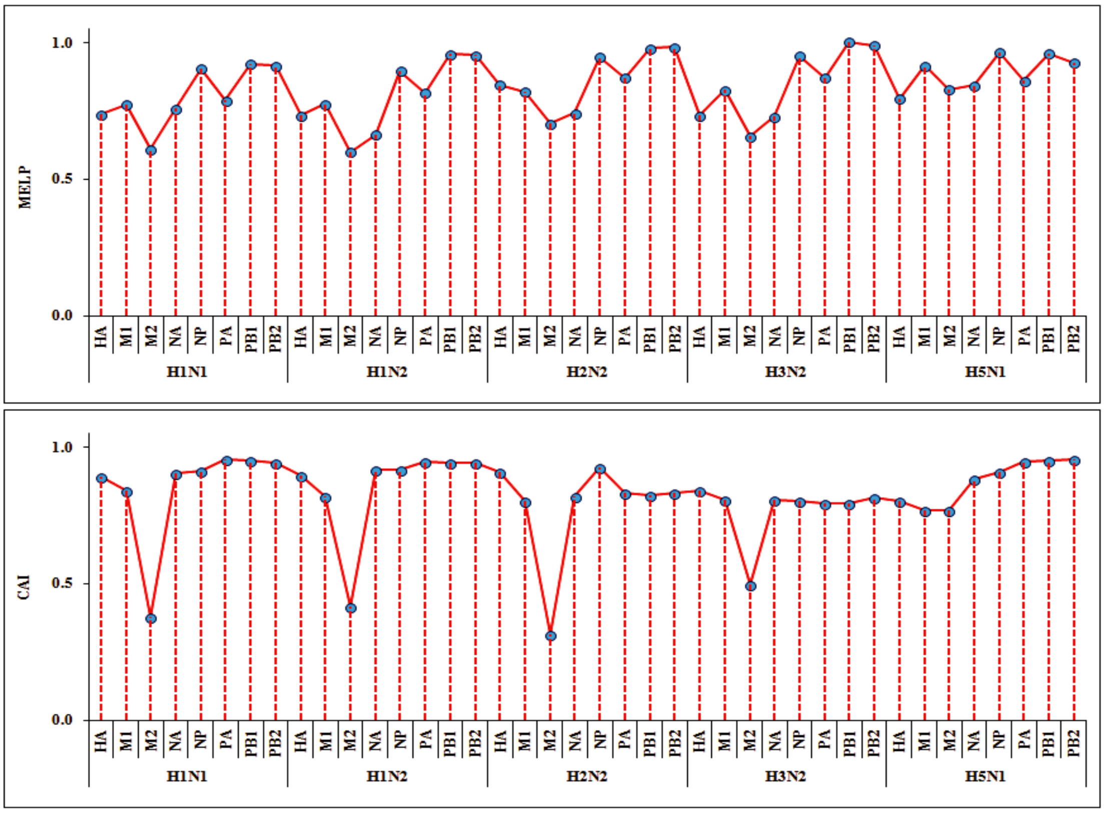


Fig. 4.4.1 Predictive expressivity of the genes (MELP and CAI) belonging to different IAV subtypes

4.4.2 MELP

MILC-based expression measure is another recently developed gene expression predictive measure given by Supek in 2005 which is independent of cds length and composition (Supek and Vlahovicek, 2005). We employed MELP along with CAI to predict expressivity of the IAV genes from all the subtypes covered in this study.

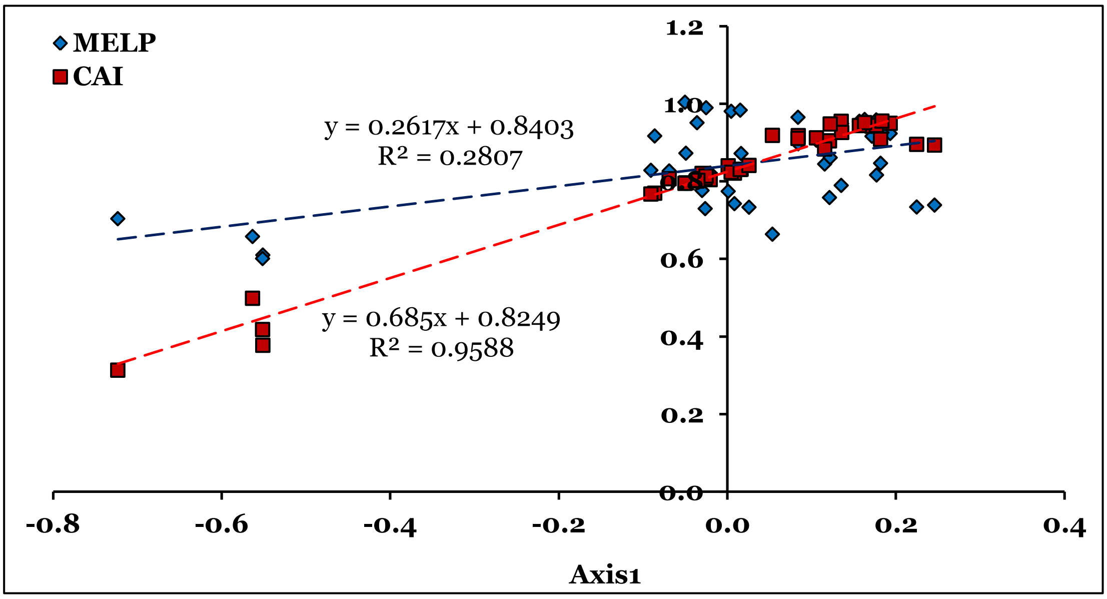


Fig. 4.4.2 Plot of the two gene expression measures MELP and CAI as a function of Axis1 of correspondence analysis

The results showed somewhat dissimilar trend of expressivity as indicated by CAI. The magnitude varied between the values of the two parameters (**Fig. 4.4.1**). The MELP values in all the genes leaving alone M2 (all subtypes) and NA (H1N2) were above 0.7 meaning higher expressivity. However, unlike CAI, in case of M2 genes the expressivity was over 0.60 (Mean MELP of 0.64 excluding M2 in H5N1). We used Pearson's correlation to see if there is any similarity between the two expression measures. The two gene expression measures *i.e.* MELP and CAI showed a significant positive correlation of 0.556 ($p < 0.01$). The axis1 of correspondence analysis showed strong significant negative correlation with CAI (r = -0.9792, p<0.01) and MELP (r = -0.5295, p<0.01) (**Fig. 4.4.2**).

4.5 Characterization the basic biochemical properties of proteins

It has been envisaged that the choice of codons has some liaison with the amino acid composition, especially, the basic biochemical properties are thought to exert some selective pressure on the codon usage profiles in some organisms (Lobry and Gautier, 1994). Most widely examined biochemical properties in this context include General/Grand average of hydropathy index (Gravy), Hydrophilicity (Hydro), Aromaticity (Aroma) and Isoelectric point (pI).

Table 4.5.1 Basic biochemical properties of the IAV proteins

	Gene	Prot_Length	MW(dalton)	Gravy	Hydro	Aroma	pI
H1N1	**HA**	564.5	61123.9	-0.3700	-0.0471	0.1007	7.0421
	M1	252.0	27099.6	-0.2560	0.0120	0.0507	9.4287
	M2	97.0	10907.8	-0.2987	0.1073	0.0867	4.8587
	NA	469.3	50671.4	-0.2582	-0.1745	0.1000	6.2036
	NP	499.8	55219.4	-0.5596	0.1972	0.0698	9.3500
	PA	725.3	81897.4	-0.4541	0.1791	0.0953	5.4788
	PB1	757.9	85159.2	-0.4925	0.0748	0.0900	9.3315
	PB2	756.0	84144.5	-0.3618	0.0953	0.0706	9.5306
H1N2	**HA**	565.1	61435.4	-0.3361	-0.0967	0.1000	6.4000
	M1	252.0	27150.8	-0.2344	0.0189	0.0500	9.3478
	M2	97.0	10934.2	-0.2733	0.1022	0.0900	5.0417
	NA	469.0	50526.4	-0.2900	-0.0600	0.0805	6.2305
	NP	498.0	55206.9	-0.5847	0.1905	0.0795	9.2832
	PA	716.0	81029.8	-0.4521	0.1732	0.0995	5.3526
	PB1	756.9	85040.5	-0.5070	0.0800	0.0900	9.2985
	PB2	759.0	84568.7	-0.3390	0.0990	0.0695	9.6167
H2N2	**HA**	562.0	61265.5	-0.3823	-0.0091	0.0905	6.3627
	M1	252.0	27185.6	-0.2075	0.0208	0.0500	9.3633
	M2	97.0	10773.1	-0.1842	0.0900	0.0892	5.0725
	NA	481.4	53228.4	-0.3601	0.0554	0.0817	7.5049
	NP	498.0	55413.0	-0.5952	0.1965	0.0800	9.2513
	PA	501.1	55537.5	-0.3734	0.0687	0.0807	7.7992
	PB1	484.0	53562.7	-0.3654	0.0683	0.0801	7.6970
	PB2	500.1	55421.3	-0.3729	0.0687	0.0807	7.7932

Table 4.5.1 *contd.*

	Gene	Prot_Length	MW(dalton)	Gravy	Hydro	Aroma	pI
H3N2	**HA**	511.4	56735.0	-0.3863	0.0806	0.0809	7.7553
	M1	464.6	51369.1	-0.3566	0.0739	0.0783	7.8216
	M2	435.2	48049.1	-0.3584	0.0714	0.0792	7.6421
	NA	482.8	53428.9	-0.3660	0.0694	0.0800	7.7082
	NP	469.1	51893.1	-0.3662	0.0681	0.0805	7.6171
	PA	464.3	51416.0	-0.3678	0.0763	0.0795	7.6780
	PB1	474.9	52629.2	-0.3744	0.0792	0.0810	7.5945
	PB2	493.8	54714.0	-0.3795	0.0781	0.0806	7.7221
H5N1	**HA**	494.9	54908.5	-0.3734	0.0794	0.0806	7.7185
	M1	444.0	49141.5	-0.3617	0.0778	0.0789	7.7337
	M2	456.1	50513.8	-0.3713	0.0814	0.0807	7.6318
	NA	466.3	51587.7	-0.3621	0.0747	0.0791	7.7542
	NP	456.6	50573.4	-0.3708	0.0810	0.0807	7.6386
	PA	478.5	52998.1	-0.3688	0.0752	0.0797	7.8163
	PB1	475.2	52616.4	-0.3691	0.0748	0.0801	7.7130
	PB2	458.5	50786.2	-0.3706	0.0804	0.0806	7.6506

The basic biochemical properties of the amino acids in IAV proteins are summarised in **table 4.5.1**. We performed correlation analysis between the basic biochemical properties of the amino acids with axis1 of correspondence analysis to examine if there is any relation between these parameters with that of codon usage.

Table 4.5.2 Correlation between the basic biochemical properties with axis1 of correspondence analysis

	Gravy	Hydro	Aroma	pI	Axis1
Gravy	1	-0.599**	-0.251	-0.333**	0.457**
Hydro	-0.599**	1	-0.123	0.18	0.128
Aroma	-0.251	-0.123	1	-0.668**	-0.049
pI	-0.333**	0.18	-0.668**	1	-0.427**
Axis1	0.457**	0.128	-0.049	-0.427**	1

** Correlation is significant at the 0.01 level (2-tailed).

Significant positive correlations were found between axis1 and Gravy ($r = 0.457$, $p < 0.01$), while axis1 and pI showed significant inverse correlation ($r = -0.427$, $p < 0.01$). Hydro and Aroma did not correlate significantly with codon usage. We also found significant inverse relationship between Gravy and Aroma ($r = -0.599$, $p < 0.01$) as well as Gravy and pI ($r = -0.333$, $p < 0.01$) along with a strong negative correlation between Aroma and pI ($r = -0.668$, $p < 0.01$). Thus, it can be concluded that the biochemical properties did have some relationship with codon usage profiles of the genes these IAV subtypes (**table 4.5.2**).