A-6107
[1-7]

RESEARCH ARTICLE

Indian Journal of Agricultural Research

Analysis of Codon Usage Bias of Six Genes of Replicase/Coat Protein of Tobacco Mosaic Virus

Kevin Cheeran[1], Kuralayanapalya Puttahonnappa Suresh[2], Siju Susan Jacob[2],
Chirathahalli Shivamurthy Sathish Gowda[2], Narayanan Gejendiran[2],
Rajangam Sridevi[2], Sharanagouda S. Patil[2]

10.18805/IJARe.A-6107

ABSTRACT

Background: Tobacco mosaic virus (TMV) stands as a highly studied virus and consequently, its features and composition are extensively understood. It has been found to induce diverse infections in numerous plant species, with tobacco leaves being notably affected, showing mottled browning. Presently, the sole available method to control its spread is by removing infected plants. Understanding codon use bias is crucial as it could play a pivotal role in molecular interventions aimed at halting the virus's replication and multiplication, thereby helping to contain its propagation.

Methods: Currently, the research focuses on assessing codon bias within six genes related to the replicase/coat protein of TMV, namely TMVgp1, TMVgp2, TMVgp3, TMVgp4, TMVgp5 and TMVgp6. To conduct this analysis, various methods such as relative dinucleotide abundance, relative synonymous codon usage (RSCU), neutrality plot and parity rule 2 (PR2) plot were employed.

Result: All of the identified genes had a modest codon bias, according to the study on codon usage, as well as the function of mutation pressure in gene TMVgp3 and natural selection in genes TMVgp1, TMVgp2, TMVgp4, TMVgp5 and TMVgp6. The Research into codon use bias showed that the TMV virus's chosen genes are subjected to naturally occurring selection as well as mutational pressure.

Key words: Codon usage analysis, Plant virus, TMV, Tobacco mosaic virus.

INTRODUCTION

The Tobacco *mosaic* virus holds a special place in virology history, having been at the forefront of virus study since the late 1800s.

TMV, first virus to be discovered, that caused the mottled browning of tobacco leaves. It also infected other plants notably tomatoes. The transmission of the virus is by physical interaction between an infected plant and the damaged or scratched leaves of normal plants. The virus is highly stable and found in nature. They can be extracted from tobacco several years after their preparation. The only control measure present in today's world is to destroy the infected plants (Okada, 1998).

Tobacco mosaic virus, is only one stranded RNA virus, 6.5 kb in length, is a member of the genus *Tobamovirus* of the family *Virgaviridae* which is a rodlike in shape with a length of 300 nm in length and has a diameter of 18 nm. The capsid's of TMV consist of 2130 identical protein subunits, which are arranged around the RNA strand to form a helical structure, this leaves a hollow central cavity of 4 nm in diameter. Six overlapping genes of replicase/coat protein *viz.*, TMVgp1, TMVgp2, TMVgp3, TMVgp4, TMVgp5 and TMVgp6 were selected for codon usage bias analysis (Morozov *et al*., 1993).

In different organisms and species, there exists a nonhomogeneous utilization of synonymous codons, observed across various genes and genomes. The frequency of synonymous codon usage in a particular species is referred to as codon usage bias (CUB). The degree of bias varies significantly among different species and this codon use bias plays a role in the process of molecular evolution. The main factors contributing to codon usage bias are natural selection and mutational pressure. (Zhou, 2016).

MATERIALS AND METHODS

Collection of data

The six genes' entire nucleotide sequences (TMVgp1, TMVgp2, TMVgp3, TMVgp4, TMVgp5, TMVgp6) were obtained from the NCBI database of Tobacco Mosaic Virus in

[1]Sir M. Visvesvaraya Institute of Technology, Bengaluru-560 064, Karnataka, India.
[2]ICAR-National Institute of Veterinary Epidemiology and Disease Informatics, Bengaluru-560 064, Karnataka, India.

Corresponding Author: Sharanagouda S. Patil, ICAR-National Institute of Veterinary Epidemiology and Disease Informatics, Bengaluru-560 064, Karnataka, India.
Email: sharanspin13@gmail.com

How to cite this article: Cheeran, K., Suresh, K.P., Jacob, S.S., Gowda, C.S.S., Gejendiran, N., Sridevi, R. and Patil, S.S. (2023). Analysis of Codon Usage Bias of Six Genes of Replicase/Coat Protein of Tobacco Mosaic Virus. Indian Journal of Agricultural Research. DOI: 10.18805/IJARe.A-6107.

Submitted: 26-04-2023 **Accepted:** 25-09-2023 **Online:** 03-10-2023

FASTA Format. The sequences were aligned and modified using MEGA software after being screened for duplicates using DAMBE software. Using RDP software, recombinant areas in the sequences were eliminated (Tamura *et al*., 2007).

Overall nucleotide content analysis

The total number of nucleotides ((A),(T), (G), (C)) in each gene at each codon's third position and Using MEGA software, different compositions including GC, GC1, GC2, GC3 and GC12 were determined. In R programming, the GC content and mononucleotide frequencies were determined using the "seqinR" library (Tamura *et al*., 2007).

Examination of the relative dinucleotide abundance

The process of Relative Dinucleotide Abundance Analysis is employed to assess the dinucleotide occurrence in the pathogen's sequence. There are 16 possible combinations of dinucleotides. Analyzing the frequency of these dinucleotides offers valuable insights into the impact of mutation and selection pressures. To calculate the relative dinucleotide abundance of the virus's six genes, the approach introduced by Karlin and Burge in 1995 was utilized.

$$\text{Pxy} = \frac{\text{Fxy}}{\text{Fx*Fy}}$$

In the given equation,
Fx/Fy = Frequency of individual nucleotides.
Fxy = Frequency of dinucleotides.

To differentiate between high and low relative abundance, a careful criterion sets Pxy >1.23 as high and Pxy <0.73 as low. The software utilized for determining the frequency of dinucleotides was R Studio Programming (Beelagi *et al*., 2021a).

Examination of Relative Synonymous Codon Usage (RSCU)

The RSCU (Relative Synonymous Codon Usage) represents the ratio of the observed value of an amino acid to its predicted value. This analysis remains unaffected by factors such as sequence length or amino acid frequency. The RSCU values offer a concise summary of how each codon is distributed in the sequence. For instance, if a codon's RSCU value exceeds 1.6, it is considered overrepresented; if it falls below 0.6, it is deemed underrepresented; and if it ranges between 1.6 and 0.6, it is considered unbiased. The calculation of RSCU values was performed using the formula provided below (Sharp and Li, 1986; Bylaiah *et al*., 2021).

$$\text{RSCU} = \text{kij}/\Sigma \text{ i j kij}$$

The formula mentioned above utilizes the symbol kij to represent the observed number of the ith codon for the jth amino acid, which has ni synonymous codons. The R Studio Programming software was employed to compute and visualize the RSCU values for the genes.

Analysis of neutrality plot

The Neutrality Plot method is employed to investigate the impact of mutational pressure and natural selection on the codon usage pattern. To create the neutrality figure, GC3 data are plotted against the GC12 mean. If the GC3 levels are substantial and close to 1, the evolution of the codon usage pattern is notably influenced by mutational pressure. On the other hand, if the regression slope is equal to 0, it suggests that natural selection has a significant effect. The same method was applied to each TMV gene by mapping GC12 values against GC3 values. The mutational pressure is represented by the regression line on the neutrality plot.

Examination of parity rule 2 (PR2) plot

The study utilized a set of guidelines known as PR2 (Parity Rule 2) to conduct two investigations. The GC bias was graphed on the abscissa as G3/(G3+C3) and the AT bias on the ordinate as A3/(A3+T3). This analysis provides insights into the relative levels of natural selection and mutation pressure based on the genomic composition. The origin of both axes is set at 0.5 (X= 0.5, Y= 0.5). Points close to the origin indicate that natural selection and mutational pressure are not in conflict, demonstrating equality between A and T, as well as between G and C (Tao and Yao, 2020; Patil *et al*., 2021).

RESULTS AND DISCUSSION

Collection of data

The CDS the gene sequences for each, TMVgp1 (n=65, l =4850 bp), TMVgp2 (n=65, l=3350 bp), TMVgp3 (n=64, l = 1424 bp), TMVgp4 (n=47, l=806 bp), TMVgp5 (n=48, l = 122 bp) and TMVgp6 (n=51, l=479 bp) taken from the NCBI database of the TMV virus. All segments' nucleotide coding sequences were aligned using MEGA × software, which was also utilised to estimate nucleotide composition and identify stop codons from each segment's sequence (MUSCLE algorithm) for alignment.

Analysis of the relative dinucleotide abundance frequency and the nucleotide makeup

To assess the extent of codon usage bias, we examined the nucleotide composition (A, T, G and C) and the nucleotide composition at position three (A3, T3, G3, C3) of the genes TMVgp1, TMVgp2, TMVgp3, TMVgp4, TMVgp5 and TMVgp6. Additionally, we calculated GC, GC1 (GC content at the 1st codon position), GC2 (GC content at the 2nd codon position) and GC3 (GC content at the 3rd codon position). Table 1 provides the frequency of nucleotide composition. This methodology allows us to estimate how each nucleotide influences the patterns of codon usage.

Upon considering the nucleotide composition of each investigated gene, it is apparent that A and T nucleotides are most frequently used across all six TMV genes. This prevalent use of A and T nucleotides in TMV might be attributed to a hereditary trait.

Dinucleotide bias can also affect codon usage bias. The R Studio program was utilized to calculate the relative abundance of all 16 dinucleotides for each TMV gene. Upon comparison to a theoretical value, the abundance frequency of each segment was found to be less consistent (equal to 1.0). Based on the abundance frequency, values exceeding 1.23 are considered overrepresented, while values below 0.78 are categorized as underrepresented.

The relative dinucleotide abundance frequencies of the all the 6 genes of TMV are depicted in Fig 1.

TMVgp1: This gene has 6 overrepresented dinucleotide bases, they are AG (1.389), GA (1.528), GG (1.326), GT (1.284), TG (1.670) and TT (1.583). It also had 5 underrepresented dinucleotide bases which are AC (0.695), CC (0.426), CG (0.706), CT (0.722), TA (0.752).

TMVgp2: This gene had a single overrepresented and underrepresented dinucleotide bases, they were CA (1.274) and TA (0.656) respectively.

TMVgp3: The dinucleotide bases, AA (1.885), CA (1.252), TG (1.265) were overrepresented and CG (0.763), TA (0.616) were underrepresented.

TMVgp4: The gene has 2 overrepresented and 3 underrepresented dinucleotide bases, they are CC (1.379), TC (1.323) and AC (0.744), GC (0.752), TA (0.667) respectively.

TMVgp5: The gene has 3 overrepresented and 2 underrepresented dinucleotide bases which are AT (1.25), CG (1.424), TC (1.505) and AC (0.771), GC (0.712) respectively.

TMVgp6: The gene has a single underrepresented dinucleotide base AT (0.762).

The dinucleotides AG, GA , GG , GT , TG and TT of TMVgp1 were overrepresented along with CA of TMVgp2, AA , CA , TG of TMVgp3 , CC , TC of TMVgp4 and AT, CG , TC of TMVgp5. It was demonstrated that each gene has its own set of abundant dinucleotides. In the same manner dinucleotides AC, CC, CG, CT, TA of TMVgp1 along with CA, TA of TMVgp2, CG, TA of TMVgp3, AC, GC, TA of TMVgp4, AC, GC of TMvgp5 and AT of TMVgp6 were underrepresented.

Table 1: Nucleotide compositions of six genes of TMV.

	Genes of TMV											
Nucleotides	TMVgp1		TMVgp2		TMVgp3		TMVgp4		TMVgp5		TMVgp6	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
T	28.04	0.33	27.43	0.13	29.1	0.26	27.6	0.18	32.83	0.85	26.9	0.25
C	19.25	0.4	20.12	0.12	17.27	0.17	15	0.17	21.62	0.77	20.95	0.24
A	28.29	0.59	27.87	0.2	30.26	0.15	31.38	0.25	25.09	1.02	29.63	0.26
G	24.42	0.56	24.59	0.2	23.37	0.24	26.03	0.23	20.47	0.82	22.52	0.2
GC	43.67	0.58	44.7	0.23	40.64	0.34	41.02	0.31	42.09	1.11	43.47	0.33
GC1	47.8	1.71	49.56	0.15	43.9	1.98	45.15	0.28	44.92	3.99	44.72	0.47
GC2	36.95	1.63	37.64	0.13	34.16	2.22	34.08	0.33	34.78	3.22	47.51	0.21
GC3	46.27	1.71	46.91	0.59	43.87	1.44	43.85	0.62	46.58	1.43	38.18	0.89
T3	30.68	1.2	30.15	0.35	31.9	1.37	31.94	0.44	33.28	3.42	30.04	0.76
C3	19.76	0.64	20.74	0.31	17.79	0.57	17.8	0.44	24.63	1.43	19.17	0.76
G3	23.05	2.1	22.94	0.49	24.23	1.86	24.21	0.48	20.14	3.77	31.77	0.5
A3	26.51	1.55	26.17	0.5	26.08	1.26	26.05	0.38	21.95	1.86	19.01	0.58

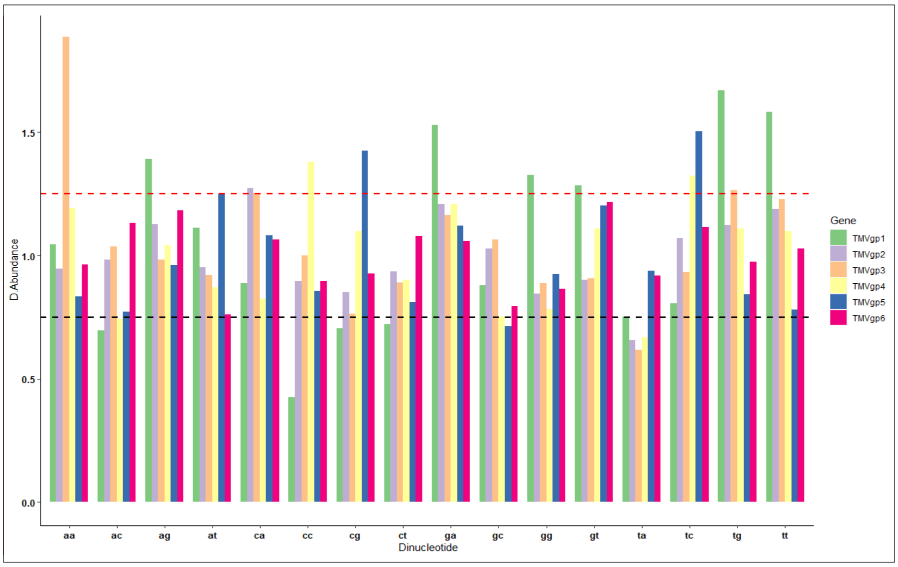

Fig 1: Dinucleotide composition of six genes of TMV.

Examination of relative synonymous codon usage (RSCU)

The proportion synonymous codon usage of six genes was determined and plotted using the R studio programme. The RSCU range of 0.6 to 1.6 is used to differentiate the frequency values of each synonymous codon. Overrepresented synonymous codons have a value of >1.6, while underrepresented synonymous codons have a value of <0.6. Yellow and red highlights, respectively, are present for the codons that are over- and under-represented (Table 2). Codons with a frequency value much more than 1.0 are referred to as high frequency or positively biassed codons. Codons with a lower frequency or those that are negatively biassed have a frequency below 1.0.

TMVgp1: This gene has 3 overrepresented and 7 underrepresented codons, they are AGA, AGG, TTG and ATA, CGC, CGG, CGT, CTA, CTC, GTA respectively. The gene has 24 high frequency and 31 low frequency codons. Among the 24 high frequency codons, 11 codons dominantly ended with the nucleotide T and 15 out of 31 low frequency codons dominantly ended with the nucleotide C.

TMVgp2: This gene has 2 overrepresented and 5 underrepresented codons, they are AGA, AGG and ATA, CGC, CGG, CTC, GTA respectively. The gene has 25 high frequency and 29 low frequency codons. Among the 25 high frequency codons, 10 codons dominantly ended with the nucleotide T and 13 out of the 29 low frequency codons dominantly ended with the nucleotide C.

TMVgp3: This gene has 7 overrepresented and 12 underrepresented codons, they are AGA, AGT, CCA, GCA, GTT, TCT, TTG and AAC, AGC, CGC, CGG, CGT, CTA, GCC, GGG, GTC, TCC, TTA, TTC respectively. The gene has 23 high frequency and 32 low frequency codons. Among the 23 high frequency codons, 11 codons dominantly ended with the nucleotide T and 12 out of the 32 low frequency codons dominantly ended with the nucleotide C.

TMVgp4: This gene has 7 overrepresented and 15 underrepresented codons, they are AGA, AGT, CAT, CTT, GGA, TCG, TGT and ACG, AGC, CAC, CCA, CGC, CGG, CGT, CTA, CTC, GGC, GGG, GTA, TCC, TGC, TTC respectively. The gene has 26 high frequency and 28 low frequency codons. Among the 26 high frequency codons, 10 codons dominantly ended with the nucleotide T and 10 out of the 28 low frequency codons dominantly ended with the nucleotide C.

TMVgp5: This gene has 11 overrepresented and 32 underrepresented codons, they are AAA, AAT, CAC, CAG, CCC, CCG, CGG, CGT, GGC, GTT, TTT and AAG, AAC, ACA, ACC, ACG, ACT, AGT, ATA, CAT, CAA, CCA, CCT, CGC, CTC, GAA, GAC, GAG, GAT, GCA, GCC, GCG, GCT, GGA, GGG, GGT, GTA, TCT, TGC, TGG, TGT, TTC, TTG respectively. The gene has 23 high frequency and 37 low frequency codons. Among the 23 high frequency codons, 7 codons dominantly ended with the nucleotide T and nucleotide C. 10 out of the 37 low frequency codons dominantly ended with the nucleotide A.

TMVgp6: This gene has 11 overrepresented and 26 underrepresented codons, they are AAT, ACT, AGA, AGG, ATA, GAC, GGA, GGT, TCT, TGT, TTA and AAC, ACA, ACC, ATT, CAC, CAT, CCC, CCG, CGA, CGC, CGG, CGT, CTA, CTC, CTG, CTT, GAT, GCT, GGC, GGG, GTC, TAT, TCC, TCG, TGC, TTT respectively. The gene has 27 high frequency and 30 low frequency codons. Among the 27 high frequency codons, 9 codons dominantly ended with the nucleotide T and nucleotide A. 11 out of the 30 low frequency codons dominantly ended with the nucleotide C.

The analysis of Relative Synonymous Codons (RSCU) revealed the following distribution among the TMV genes: TMVgp1 had 24 high frequency and 31 low frequency codons, TMVgp2 had 25 high frequency and 29 low frequency codons, TMVgp3 had 23 high frequency and 32 low frequency codons, TMVgp4 had 26 high frequency and 28 low frequency codons, TMVgp5 had 23 high frequency and 37 low frequency codons and TMVgp6 had 27 high frequency and 30 low frequency codons. This analysis underscored the significance of dinucleotide and mononucleotide compositions in influencing the codon usage pattern within TMV.

Examination of parity rule 2 (PR2) plot

The PR2 origin indicates the direction and extent of bias. The PR2 bias plot provides valuable information when evaluating the biases at the third position of AT and GC content. According to Chargaff's second parity rule (PR2), the nucleotide composition of DNA follows A=T and G=C. Therefore, the origin represents the point where bias has not developed. The X-axis represents the values of [G3/(G3+C3)] and the Y-axis represents the values of [A3/(A3+T3)]. For each TMV viral gene selected in this study, the mean values of [G3/(G3+C3)] and [A3/(A3+T3)] were as follows:

TMVgp1: GC and AT bias were calculated to be 0.43 and 0.56, respectively. The AT's dominant over the GC indica.

TMVgp2: GC and AT bias were calculated to be 0.44 and 0.55, respectively. The AT's dominant over the GC indica.

TMVgp3: GC and AT bias were calculated to be 0.40 and 0.59, respectively. The AT's dominant over the GC indica.

TMVgp4: GC and AT bias were calculated to be 0.41 and 0.58, respectively. The AT's dominant over the GC indica.

TMVgp5: GC and AT bias were calculated to be 0.42 and 0.57, respectively. The AT's dominant over the GC indica.

TMVgp6: GC and AT bias were calculated to be 0.43 and 0.56, respectively. The AT's dominant over the GC indica.

There is a bias in the genes analysed in this study because none of the genes had an AT=GC composition. The genes TMVgp2, TMVgp4 and TMVgp6 exhibited a little less bias than the genes TMVgp1, TMVgp3 and TMVgp5 because their sites were situated further from the origin (Fig 2).

Table 2: Relative synonymous codons usage of each amino acid in six genes of TMV.

Genes	TMVgp1	TMVgp2	TMVgp3	TMVgp4	TMVgp5	TMVgp6
aaa	1	0.8923077	1.1794872	1.0769231	2	1
aac	0.8	0.9268293	0.5555556	0.6	0	0.2
aag	1	1.1076923	0.8205128	0.9230769	0	1
aat	1.2	1.0731707	1.4444444	1.4	2	1.8
aca	1.0989011	1.2352941	0.7272727	1.4545455	0	0.5
acc	0.8791209	0.8235294	1.0909091	1.4545455	0	0.5
acg	0.7472527	0.7058824	0.9090909	0.3636364	0	0.75
act	1.2747253	1.2352941	1.2727273	0.7272727	0	2.25
aga	2.4705882	2.0307692	3.9	3.4285714	0.8571429	2.7272727
agc	0.761194	0.8484848	0.5625	0.2857143	1.5	1.125
agg	1.6235294	1.7538462	1.2	0.8571429	0.8571429	1.6363636
agt	0.8507463	0.6060606	1.6875	1.7142857	0	1.125
ata	0.5853659	0.4117647	0.9310345	0.75	0	2
atc	0.9878049	1	1.0344828	1.25	1.5	0.6666667
att	1.4268293	1.5882353	1.0344828	1	1.5	0.3333333
caa	0.8727273	0.8823529	0.8421053	0.6666667	0	1.3333333
cac	0.7804878	0.6666667	1.0909091	0	2	0
cag	1.1272727	1.1176471	1.1578947	1.3333333	2	0.6666667
cat	1.2195122	1.3333333	0.9090909	2	0	0
cca	1.3584906	1.0909091	1.6842105	0	0	1.5
ccc	0.8301887	0.969697	0.6315789	1	2	0.5
ccg	1.0566038	1.0909091	1.0526316	1.5	2	0.5
cct	0.754717	0.8484848	0.6315789	1.5	0	1.5
cga	0.7764706	0.8307692	0.6	0.8571429	0.8571429	0.5454545
cgc	0.4941176	0.5538462	0.3	0.4285714	0	0
cgg	0.1411765	0.1846154	0	0.4285714	1.7142857	0.5454545
cgt	0.4941176	0.6461538	0	0	1.7142857	0.5454545
cta	0.5316456	0.6055046	0.3829787	0.3	1.5	0.5
ctc	0.5696203	0.5504587	0.6382979	0.3	0	0
ctg	0.835443	0.8807339	0.7659574	0.9	1.5	0.5
ctt	1.2911392	1.4311927	0.893617	1.8	1.5	0
gaa	1	1	1	0.7058824	0	1.1428571
gac	0.7884615	0.8923077	0.6111111	0.7058824	0	1.75
gag	1	1	1	1.2941176	0	0.8571429
gat	1.2115385	1.1076923	1.3888889	1.2941176	0	0.25
gca	1.4414414	1.3658537	1.6296296	0.9230769	0	1.1428571
gcc	0.7207207	0.8292683	0.4444444	1.2307692	0	1.1428571
gcg	0.972973	0.9756098	0.8888889	0.6153846	0	1.1428571
gct	0.8648649	0.8292683	1.037037	1.2307692	0	0.5714286
gga	1.3142857	1.2765957	1.5238095	2.25	0	2
ggc	0.8	0.7659574	0.952381	0.25	4	0
ggg	0.7428571	0.9361702	0.3809524	0.5	0	0
ggt	1.1428571	1.0212766	1.1428571	1	0	2
gta	0.481203	0.3529412	0.9655172	0.125	0	1.1428571
gtc	0.7819549	0.9019608	0.4137931	1.125	1.3333333	0.5714286
gtg	1.2030075	1.254902	0.9655172	1.25	0.6666667	1.1428571
gtt	1.5338346	1.4901961	1.6551724	1.5	2	1.1428571
tac	1.030303	1.0434783	1	1	0.6666667	1.5
tat	0.969697	0.9565217	1	1	1.3333333	0.5
tca	1.119403	1.2727273	0.75	1.1428571	1.5	1.125

Table 2: Continue.....

Table 2: Continue.....

tcc	0.7164179	0.8484848	0.1875	0.2857143	1.5	0
tcg	1.0746269	1.0909091	0.9375	1.7142857	1.5	0
tct	1.4776119	1.3333333	1.875	0.8571429	0	2.625
tgc	0.9333333	1	0.8333333	0.3333333	0	0
tgg	1	1	1	1	0	1
tgt	1.0666667	1	1.1666667	1.6666667	0	2
tta	0.9873418	1.1559633	0.5106383	1.5	1.5	3.5
ttc	0.6410256	0.6808511	0.5517241	0.5454545	0	1.5
ttg	1.7848101	1.3761468	2.8085106	1.2	0	1.5
ttt	1.3589744	1.3191489	1.4482759	1.4545455	2	0.5

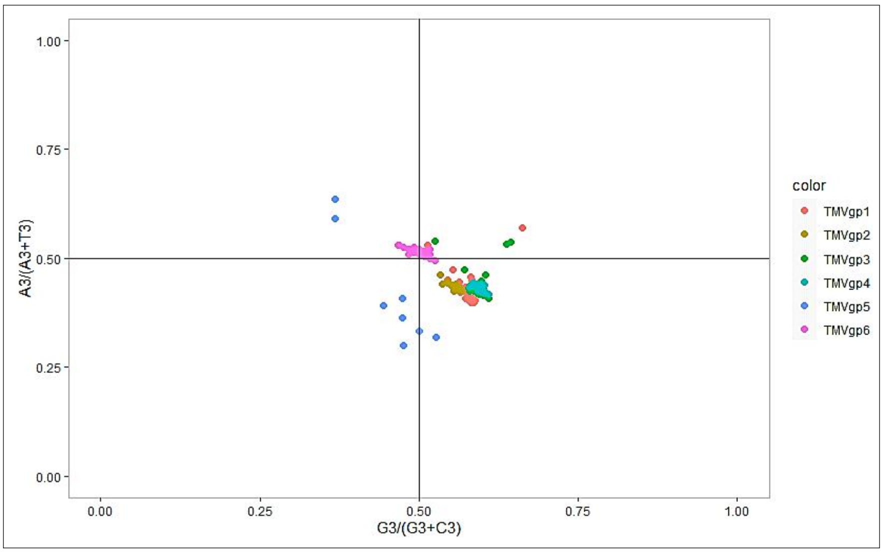

Fig 2: Parity rule 2 plots AT-bias against GC-bias. Each point represents six gene sequences of TMV.

The parity rule 2 plot reveals bias at the third position of AT and GC in all the chosen genes, suggesting that natural selection has a major impact on the pressure of mutation.

The parity rule 2 showed that none of the genes evaluated in this study have an AT=GC composition, showing a bias among the genes examined. The genes TMVgp2, TMVgp4 and TMVgp6 exhibited a smaller bias than the genes TMVgp1, TMVgp3 and TMVgp5 because the sites of the TMVgp2, TMVgp4 and TMVgp6 genes were positioned further from the origin. Due to this, the parity rule 2 plot demonstrates bias in all of the chosen genes at the third position of AT and GC, suggesting that natural selection has a major influence on the pressure of mutation.

Analysis of neutrality plot

The neutrality was assessed and graphed by comparing the nucleotide composition of GC12 (mean value of GC1 and GC2) with GC3, aiming to identify the influences of natural selection and mutational pressure. The slope of the regression line in the graph indicates the evolutionary rate of natural selection and mutational pressure. Additionally, the regression coefficient against GC12 and GC3, acting as a natural-mutational equilibrium coefficient, is also considered. The interpretation of the neutrality plot for this virus is as follows:

TMVgp1: In the case of this gene, the neutrality plot displayed a negative regression line and a significant negative R-value with y = 0.473 - 0.103x, where R2 = 0.09. The neutrality at 10.3% suggests that natural selection has a more substantial influence than mutational pressure in shaping the codon usage bias.

TMVgp2: In the case of this gene, the neutrality plot showed a positive regression line and a significant positive R-value with y = 0.412 + 0.0521x, where R2 = 0.09. The neutrality at 5.21% indicates that natural selection has a dominant role in shaping the codon usage bias, exerting more influence than mutational pressure.

TMVgp3: In the case of this gene, the neutrality plot exhibited a negative regression line and a significant negative R-value with y = 0.673 - 0.644x, where R2 = 0.80. The neutrality at 64.4% indicates that mutational pressure plays a dominant role in shaping the codon usage bias of this gene, exerting more influence than natural selection.

TMVgp4: In the case of this gene, the neutrality plot displayed a positive regression line and a significant positive R-value with y = 0.309 + 0.198x, where R2 = 0.29. The neutrality at 19.8% indicates that natural selection plays a dominant role in shaping the codon usage bias, exerting more influence than mutational pressure.

TMVgp5: In the case of this gene, the neutrality plot exhibited a negative regression line and a significant negative R-value with $y = 0.398\text{-}3.76 \times 10 \quad 5x$, where R2 <0.01. The neutrality at 0.37% indicates that natural selection plays a dominant role in shaping the codon usage bias, exerting more influence than mutational pressure.

TMVgp6: In the case of this gene, the neutrality plot displayed a negative regression line and a significant negative R-value with y = 0.468 - 0.019x, where R2 <0.01. The neutrality at 1.9% indicates that natural selection plays a dominant role in shaping the codon usage bias, exerting more influence than mutational pressure.

From the neutrality plots of the 6 genes, 5 genes have indicated natural selection will shape its codon usage bias over mutational pressure.

To assess the driving forces behind bias and understand the evolutionary factors involved, a neutrality analysis and plot were conducted. If the regression coefficient is less than 0.5, natural selection is the primary cause of bias; if it is more than 0.5, mutational pressure is the main cause of bias. Among the 6 genes analyzed, 5 genes (TMVgp1, TMVgp2, TMVgp4, TMVgp5 and TMVgp6) indicated that natural selection shapes their codon usage bias more than mutational pressure, whereas TMVgp3 showed that mutational pressure plays a more dominant role in shaping its codon usage bias over natural selection.

Viral mutations can arise from various factors, including polymerase fidelity, sequence context, template secondary structure, cellular environment, replication procedures, proofreading and accessibility to post-replicative repair.

CONCLUSION

The findings from the codon usage bias research study revealed that all six genes of TMV exhibit bias, with both natural selection and mutational pressure significantly influencing the pattern of codon usage bias. Natural selection affects GC3 and GC12 in all genes except TMVgp3, where mutational pressure is identified as a contributing factor based on the neutrality plot. These results indicate that the study can be valuable in developing control measures and gaining insights into the evolutionary characteristics of the Tobacco Mosaic virus.

Conflict of interest: None.

REFERENCES

Beelagi, M.S. (2021a). Synonymous codon usage pattern among the S, M and L segments in crimean-congo hemorrhagic fever causing virus. Bioinformation. 17(4): 479-491. https://doi.org/10.6026/97320630017479.

Bylaiah, S., Shedole, S., Suresh, K.P., Gowda, L., Patil, S.S., Indrabalan, U.B. and Shivamallu, C. (2021). Relative analysis of codon usage and nucleotide bias between anthrax toxin genes subsist inpxo1 plasmid of bacillus anthracis. Annals of the Romanian Society for Cell Biology. 25(4): 5758-5774.

Karlin, S. and Burge, C. (1995) Dinucleotide relative abundance extremes: Agenomic signature. Trends Genet. 11(7): 283-290.

Morozov, S.Y., Denisenko, O.N., Zelenina, D.A., Fedorki, O.N., Solovyev, A.G., Maiss, E, Casper, R and Atabekov, J.G (1993). A novel open reading frame in tobacco mosaic virus genome coding for a putative, small positively charged protein. Biochimie. 75: 659-665.

Okada, Y. (1998). Tobacco mosaic virus. Uirusu. Journal of Virology. 48(1): 97-102. https://doi.org/10.2222/jsv.48.97.

Patil, S.S., Indrabalan, U.B., Suresh, K.P. and Shome, B.R. (2021) Analysis of codon usage bias of classical swine fever virus. Veterinary World. 14(6): 1450-1458.

Sharp, P.M. and Li, W.H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24(1-2): 28-38.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007). MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24(8): 1596-1599.

Tao, J. and Yao, H. (2020). Comprehensive analysis of the codon usage patterns of polyprotein of Zika virus. Prog. Biophys. Mol. Biol. 150(1): 43-49.

Zhou, J. and Teo, Y.Y. (2016). Estimating time to the most recent common ancestor (TMRCA): Comparison and application of eight methods. European Journal of Human Genetics. 24(8): 1195-1201. https://doi.org/10.1038/ejhg.2015.258.