INNSPUB

International Journal of Biosciences | IJB |
ISSN: 2220-6655 (Print), 2222-5234 (Online)
http://www.innspub.net
Vol. 19, No. 1, p. 31-45, 2021

RESEARCH PAPER OPEN ACCESS

# Analysis of Codon Usage and Nucleotide Bias in Severe Acute Respiratory Syndrome Coronavirus 2(SARS-CoV-2) Genes

**Satyabrata Sahoo**

*Department of Physics, Dhruba Chand Halder College, Dakshin Barasat, South 24 Parganas, W.B., India*

**Key words:** SARS-CoV-2; Relative synonymous codon usage (RSCU); Codon bias; Codon adaptaion index; Codon deoptimization; Statistical analysis.

http://dx.doi.org/10.12692/ijb/19.1.31-45 Article published on July 31, 2021

## Abstract

SARS-CoV-2 has recently emerged as a virus that poses a significant public health concern. The genetic features concerning the codon usage of SARS-CoV-2 genes were analyzed by the relative synonymous codon usage, the relative strength of codon bias, the effective number of codons (ENC), the codon adaptation index, and neutrality plot. Compositional analysis indicated that G and C at the first and second codon positions significantly affect synonymous codon choices. The mutational bias toward A/U may confer a selective advantage. The results suggest that mutation, together with selection dynamics, may play an essential role in shaping the pattern of codon usages in SARS-CoV-2 genomes. Turning to the codon usage preference and codon pair association in the viral genome, some of the most preferentially used codon observed across the genome did not occur at similar magnitudes in all genes. The possible co-evolution of the virus and its adaptation to the animal host has been discussed based on the codon adaptation index and codon de-optimization index.

* **Corresponding Author:** Satyabrata Sahoo ✉ dr_s_sahoo@yahoo.com

## Introduction

SARS-CoV-2, the newly emerged virus causing the outbreak of COVID-19, has posed a great threat to global public health[1]. It is a positive-stranded long RNA virus having a crown-like appearance and causes mostly respiratory and enteric diseases in different animals, including camels, cattle, cats, bats, and humans. The genome sequence of SARS-CoV-2 contains 29891 nucleotides, encoding for 9860 amino acids[2]. The genomic structure, typical of other beta coronaviruses, has 14 open reading frames (ORFs), encoding for 27 proteins. The organization of the genome is 5′ leader-UTR (untranslated region)-replicase-S (Spike)–E (Envelope)-M (Membrane)-N (Nucleocapsid)-3′UTR poly (A) tail with accessory genes interspersed within the structural genes at the 3′ end of the genome[1,3]. The replicase gene encodes non-structural proteins (nsps) which are essential for virus replication. The four major structural proteins: the spike (S) protein, nucleocapsid (N) protein, membrane (M) protein, and the envelope (E) protein are required to produce a structurally complete viral particle and which are essential for SARS-CoV-2 assembly and infection. Pathophysiology and virulence mechanisms of SARS-CoV-2 have links to the function of the nsps and structural proteins. Accessory genes are seen intermingled within the structural genes.

The pathogenesis of SARS-CoV-2 infection in humans remains unclear. The codon usage pattern of an organism provides useful insight to facilitate a better understanding of the structure and evolution of gene coding sequences of the species. It plays an important role in controlling the speed of translation elongation during mRNA translation[4]. Bioinformatic analysis revealed that codon usage bias could be useful in estimating the translational efficiency of a gene. By deliberately deoptimizing the codon usage viral gene expression can be downregulated which may be helpful for effective vaccine formulation against viral pathogens[5,6]. The purpose of the present study was to elucidate the codon usage pattern of SARS-CoV-2, which may help in understanding the viral pathogenesis.

## Materials and methods

The genome sequences for SARS-CoV-2(NC_045512.2) based on completeness and annotation were retrieved from the National Center for Biotechnology GenBank database (http:// www.ncbi.nlm.nih.gov).

*Analysis of nucleotide organization*

The overall frequencies of occurrence of the four nucleotides (A, G, C, and U), the occurrence of nucleotides (A, G, C and U at the first, second, and third position of all codons, the occurrence of GC at the first (GC1), second (GC2) or third position (GC3), the overall GC contents, AU1,AU2, and AU3 and the frequency of occurrence of each nucleotide at the third position of synonymous codons(A3, U3, G3, and C3) were calculated using an in-house Fortran program for the analysis of codon usage and nucleotide bias of genes in SARS-COV-2. When there is no external pressure, mutations should occur in a random rather than in a specific direction. This will result in uniform base composition at three positions of codons. However, in the presence of selective pressure, preference for a particular base would occur in three different positions.

*Analysis of dinucleotide composition*

The frequencies of 16 dinucleotides (GpA, GpC, GpG, GpU, CpA, CpC, CpG, CpU, UpA, UpC, UpG, UpU, ApA, ApC, ApG, and ApU) along with their expected frequencies were also calculated for the analysis of compositional bias in genes. The identification of favored dinucleotides and the patterns of dinucleotide usage may have the effect on selection of codons in genes. The ratio of observed and expected frequencies may be used for identification of over- or under-represented dinucleotides and is given by,

$$R_{XY} = \frac{f_{XY}}{f_X f_Y}$$

where,fx and $f_Y$ are the frequency of individual nucleotides (x and y, respectively), and fxy is the frequency of dinucleotides (xy) in the same sequence[7]. If the ratio of the observed to expected dinucleotide frequency is more than 1.2, the dinucleotide is

considered overrepresented, whereas values below 0.8 indicate an underrepresentation.

*Analysis of Codons usage*

a) The Relative Synonymous Codon Usage (RSCU):

The Relative Synonymous Codons Usage (RSCU) has been calculated to describe the synonymous codon usage pattern[8]. RSCU was calculated by determining the ratio of observed usage frequency of a codon to the expected frequency, given that all codons for a specific amino acid are used equally. Codons showing an RSCU value of 1 means no synonymous bias in the codon usage pattern of the gene, while codons with RSCU values >1 or <1 are showing positive or negative synonymous codon bias, are preferred or un-preferred codons for efficient translation, respectively.

RSCU has been calculated by using the following equation:

$$RSCU_i = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}}$$

Where *Xij* is the observed number of the i[th] codon for j[th] amino acid which has $n_i$ number of synonymous codons for the amino acid.

b) The Relative Strength of Codon Bias (RCBS):

The Relative Strength of Codons Bias ( RCBS) has been proposed to describe the codon usage pattern under the assumption of random codon usage in genes under study[9,10]. RCBS was calculated by determining the ratio of observed frequency of a codon to the expected frequency, given that base composition is biased at three sites of all codons in the gene under study. Codons showing an RCBS value of 1 means no codon bias or the codon usage frequency is similar to the expected value, while codons with RCBS values >1 or <1 are showing overrepresented or underrepresented codons (with respect to a randomized sequence) respectively in respect of compositional bias of nucleotides.

RCBS has been calculated by using the following equation:

$$RCBS_{xyz} = \frac{f_{xyz}}{f(x)_1 f(y)_2 f(z)_3}$$

where $f_{xyz}$ is the normalized codon frequency of a codon *xyz* and $f_n(m)$ is the normalized frequency of base *m* at codon position *n* in a gene. The ratio of RSCU to RCBS indicates the influence of mutational bias over natural selection in the choice of codons in a gene. The optimal codons are identified as codons with RSCU>1 and RCBS>1, whereas for rare codons both RSCU<0.5 and RCBS<0.5.

c) The effective number of codons (ENC):

The concept of ENC is a simple and absolute measure of codon usage bias in the use of synonymous codons[11]. The value of ENC ranges from 20 to 61, with lower ENC values (<35) indicating strong codon usage bias of a gene. The effective number of codons has been calculated by

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

where $F_K$ values for k-fold degenerate amino acids can be estimated by

$$F_k = \sum_{i=1}^{k} \left(\frac{m_i}{m}\right)^2$$

where $m_i$ is the number of occurrences of ith codon for the k-fold degenerate amino acid having total m number of synonymous codons.

d) $GC_3$ measures the frequency of G or C at the third position of synonymous codons and can be used as an index of mutation bias on codon usage. It is measured by

$$GC_3 = \frac{\sum_{(NNS)\in C} f_{NNS}}{\sum_{(NNN)\in C} f_{NNN}}$$

where *N= any base, S=G or C,*and $f_{xyz}$ is the observed frequency of codon *xyz*.

e) Neutrality Plot:

The neutrality plot is an analytical method to analyze the influence of mutation bias and natural selection

on codon usage. In neutrality plot, a regression line was plotted between average GC contents at the first and second synonymous codon positions (GC12) and GC3 contents at the third synonymous codon position. A slope of the regression line is indicative of the mutational force. A regression plot with a slope of zero indicates no effect of directional mutation pressure, while a slope of 1 indicates complete neutrality[12].

f) The ENC-Plot (ENC vs GC3s) is commonly used to determine whether the codon usage of a gene is affected by mutation or selection. The ENC-plot is the comparison of the observed and expected number of genes based on GC3s on a single plot. Expected ENC values for all GC3s compositions were calculated using the equation

$$ENC_{exp} = 2 + S + \frac{29}{S^2 + (1-S)^2}$$

where S corresponds to the GC3s value and used to plot standard curve[13]. Data points located on or just below the standard curve ($ENC_{exp}$) indicate mutational pressure determines the codon usage bias, while data points located far away from the standard curve indicate that factors other than mutational pressure are affecting the codon usage bias.

g) The codon adaptation index (CAI):

The Codon Adaptation Index, CAI, a measure of codon bias based on RSCU values of the codons in reference to a set of highly expressed genes is given by,

$$CAI = \left( \prod_{1}^{N} w_i \right)^{\frac{1}{N}}$$

where N is the number of codons in the gene and relative adaptiveness of an i[th] codon, $w_i$ is defined as

$$w_i = \frac{(RSCU)_i}{(RSCU)_{i,max}}$$

$RSCU_i$ is the RSCU value of the $i^{th}$ codon for j[th] amino acid and $RSCU_{i,,\ max}$ is the RSCU value of the most frequent codon used for encoding j[th] amino acid in a reference to a set of highly expressed genes[14]. The score measured by CAI ranges from 0 to 1 indicating that the higher is the CAI values, the genes are more likely to be highly expressed. CAI was proposed as a measure of codon bias of a gene relative to a highly expressed reference set of genes. Although this method has been applied successfully for the prediction of highly expressed genes in a query genome sequence, it relies on the prior definition of a reference set of highly expressed genes.

However, we have introduced an alternative methodology to calculate the codon adaptation index of a gene from a whole-genome perspective to study the codon usage pattern and gene expression profile of an organism.

$$CAI_g = \prod_{i=1}^{N} (S_i)^{\frac{1}{N}}$$

where $S_i$ is the impact score of $i^{th}$ codon defined as

$$S_i = \frac{F_{ij}}{F_{max,j}}$$

where *Fij* is the number of occurrence of the i[th] codon for a j[th] amino acid which has $n_i$ number of synonymous codons for the whole set of coding sequences in a genome and $F_{max,j}$ is the observed number of the most frequent codon used for encoding j[th] amino acid. *N* is the codon length of a gene.

h) Relative Codon Deoptimization Index (RCDI):

The relative codon deoptimization index RCDI, which compares the similarities in codon usage of a gene with reference genome, may provide an estimate of the rate of viral gene translation in a host genome. It is given by

$$RCDI = \sum_{i-1}^{61} \frac{f_{ri}}{F_{ri}} \frac{N_i}{N}$$

where $f_{ri}$ is the relative frequency of codon i for a specific amino acid in the test sequence; $F_{ri}$ is the relative frequency of codon i for a specific amino acid

in the reference sequence; Ni is the number of occurrences of codon i in the test sequence; and N, the total number of codons in the test sequence[15]. The RCDI ranges from 1 (the codon usage of the test sequence is completely optimized to the codon usage of the reference genome) to N (increases with the deoptimization of the test sequence). If the RCDI value is close to one, and the higher translation rate can be predicted as well as being indicative of a greater adaptation to the host[16].

**Results and discussion**

The compositional analysis of SARS-COV-2 genome reveals that the genome is AU rich and A/U ending codons are more preferred than G/C ending codons. The U (32.3%) and A (29.9%) nucleotides occur more frequently than C (18.1%) and G (19.7%) nucleotides in SARS-CoV-2 genome. The GC content of the genome is 37.9%. The individual SARS-COV-2 gene has strong compositional bias with highest compositional value of G1 (48.78±4.03%). The average GC1(%),GC2(%),GC3(%), GC(%) and GC3s(%) in the genes are 45.07±4.91,34.73±9.38, 32.00±4.99, 37.47±4.85 and 29.57±4.67 respectively[Figure-1]. GC1 is highest in all genes except ORF10. Further analysis reveals significant abundance of T3s (48.73±5.73%) and A3s (38.36±5.42%), as compared to C3s (21.78±4.38) and G3s (16.02±2.55%) [Figure-2]. T3s is highest in all genes except ORF7a and ORF10.Further analysis in respect of structural genes revealed that, for E,S and M genes the nucleotide frequencies follow U>A>C>G and for N genes, the nucleotide frequencies are A > C>U > G . Although AU% is always greater than GC% in all structural genes, but the difference is less in case of M and N genes.

**Table 1.** CAI values of SARS-CoV-2 genes in respect of different animal hosts.

| GENE | Length | *SARS-COV-2* | *Homo sapiens* | *Gallus gallus* | *Naja atra* | *Bungarus multicinctus* | Rhinolophus ferrumequinum | *Marmota monax* |
|---|---|---|---|---|---|---|---|---|
| ORF1ab | 21291 | 0.723 | 0.654 | 0.743 | 0.704 | 0.707 | 0.742 | 0.816 |
| ORF1ab | 13218 | 0.723 | 0.652 | 0.742 | 0.702 | 0.706 | 0.747 | 0.819 |
| S | 3822 | 0.718 | 0.669 | 0.756 | 0.720 | 0.718 | 0.763 | 0.835 |
| ORF3a | 828 | 0.669 | 0.651 | 0.733 | 0.665 | 0.665 | 0.753 | 0.841 |
| E | 228 | 0.590 | 0.578 | 0.652 | 0.580 | 0.591 | 0.768 | 0.791 |
| M | 669 | 0.601 | 0.637 | 0.710 | 0.627 | 0.620 | 0.800 | 0.868 |
| ORF6 | 186 | 0.746 | 0.625 | 0.707 | 0.694 | 0.708 | 0.672 | 0.785 |
| ORF7a | 366 | 0.687 | 0.675 | 0.769 | 0.728 | 0.737 | 0.806 | 0.867 |
| ORF7b | 132 | 0.715 | 0.624 | 0.690 | 0.744 | 0.729 | 0.792 | 0.794 |
| ORF8 | 366 | 0.710 | 0.651 | 0.734 | 0.730 | 0.751 | 0.751 | 0.793 |
| N | 1260 | 0.584 | 0.699 | 0.767 | 0.662 | 0.627 | 0.819 | 0.909 |
| ORF10 | 117 | 0.606 | 0.586 | 0.666 | 0.595 | 0.576 | 0.611 | 0.728 |

The nucleotides at the first, second and third position of codons were observed to have variations which contribute to the codon biases and the codon pattern differences in structural genes. The overall nucleotide frequency at third codon position is highest for U. These findings may indicate directional mutation in the codon usage pattern in SARS-CoV-2 genes.

Dinucleotide composition may have the consequences on the intrinsic characteristics of the viral genome and appear to have a functional role in the infection and propagation of the viruses[17,18]. In SARS-CoV-2 genome, the calculated frequency ratios did not deviate much from 1 for most dinucleotides, but there are some exceptions [Figure-3]. The dinucleotide UpG and CpU showed over-representation, and, CpG and CpC dinucleotides showed under-representation in SARS-CoV-2. In the present study, we observed that UpG was the most abundant dinucleotide reflecting a high abundance of U in the SARS-CoV-2 genome, whereas, CpG was, the least abundant dinucleotide pair with the lowest odds ratio.

**Table 2.** RCDI values of SARS-CoV-2 genes in respect of different animal hosts.

| GENE | Length | *Homo sapiens* | *Bungarus multicinctus* | *Naja atra* | *Gallus gallus* | *Rhinolophus ferrumequinum* | *Marmota monax* |
|---|---|---|---|---|---|---|---|
| ORF1ab | 21291 | 1.961 | 1.410 | 2.037 | 1.545 | 2.178 | 1.878 |
| ORF1ab | 13218 | 1.910 | 1.397 | 1.994 | 1.529 | 2.135 | 1.857 |
| S | 3822 | 2.226 | 1.389 | 1.848 | 1.536 | 2.098 | 1.839 |
| ORF3a | 828 | 1.761 | 1.379 | 2.683 | 1.469 | 2.228 | 1.757 |
| E | 228 | 2.011 | 1.906 | 4.037 | 1.931 | 2.446 | 2.383 |
| M | 669 | 1.622 | 1.494 | 3.380 | 1.473 | 2.011 | 1.728 |
| ORF6 | 186 | 2.713 | 2.170 | 2.246 | 2.143 | 3.405 | 2.622 |
| ORF7a | 366 | 2.298 | 1.460 | 2.247 | 1.601 | 2.270 | 1.928 |
| ORF7b | 132 | 2.628 | 1.991 | 2.083 | 2.083 | 2.913 | 2.639 |
| ORF8 | 366 | 2.314 | 1.640 | 2.895 | 1.971 | 2.547 | 2.564 |
| N | 1260 | 2.626 | 1.375 | 1.980 | 1.371 | 1.860 | 1.501 |

The low relative abundance of CpG may be linked to DNA methylation mechanism of vertebrates. DNA methylation was found to occur predominantly on cytosines followed by guanine residues and the major function of DNA methylation is the suppression of gene expression[19]. Thus, similar pattern of dinucleotide compositions appear to have a functional role in the infection and propagation of the virus. To evade host immunity, the virus may maintain low CpG content[20]. Among other dinucleotides mild suppression of GpU and UpC, and slight overrepresentation of ApG and GpA have been observed. Similar dinucleotide biases are also observed in individual genes of SARS-CoV-2.

The CpG dinucleotide is underrepresented in all SARS-COV-2 genes except E and ORF10, and UpG is overrepresented except ORF10.

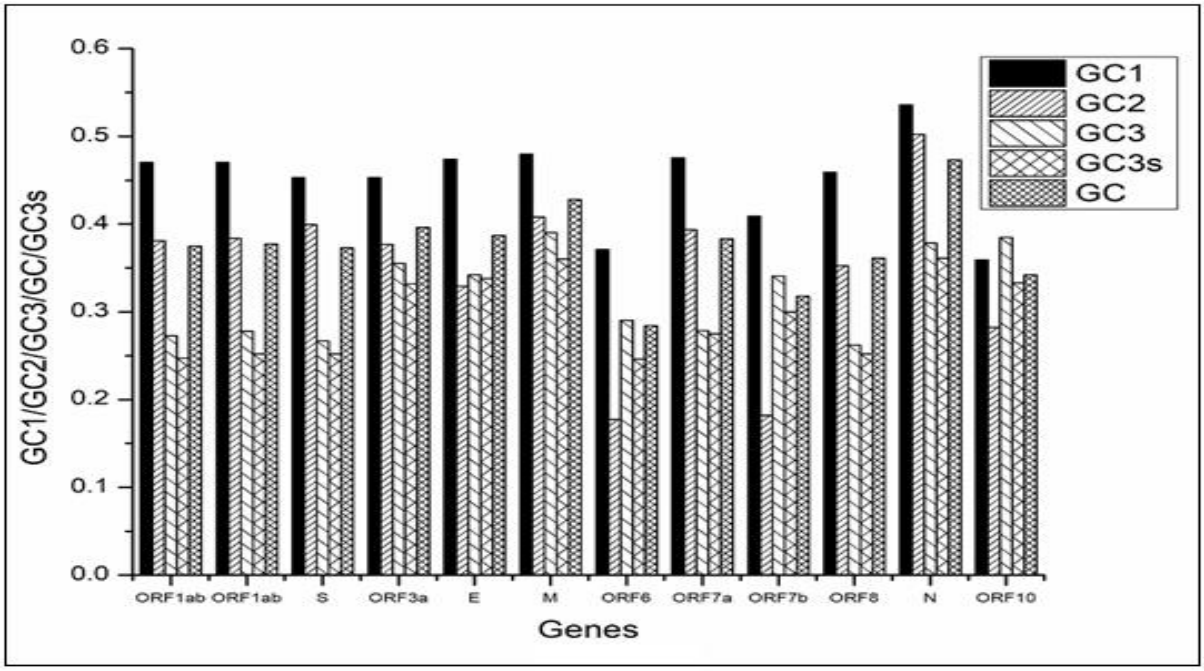


**Fig. 1.** Overall GC content, GC1 (at first site of codons), GC2 (at second site of codons) and GC3 (at third site of codons), and GC3s (at synonymous third codon sites) in all coding sequences of SARS-CoV-2 genome.

The abundance of GpC dinucleotide is high in all SARS-COV-2 genes except ORF6 and ORF8. The ratio CpG/GpC may be important in estimating the role of evolutionary process and mutational pressure acting upon constituent nucleotides. The significant lower value for S gene demonstrates that CpG dinucleotides are susceptible to evolutionary pressures.

ENC is a measure of bias from equal codon usage in a gene ENC values are calculated to determine the

inter-genic codon bias. For SARS-CoV-2 genes (ORFs having length greater than 200), ENC values ranges from 33.57 to 47.05 with a mean value 40.89±7.02. The highest ENC value is 47.05 for M gene, whereas the lowest value is 33.57 for E gene. The more biased a gene is, the smaller is the ENC value.

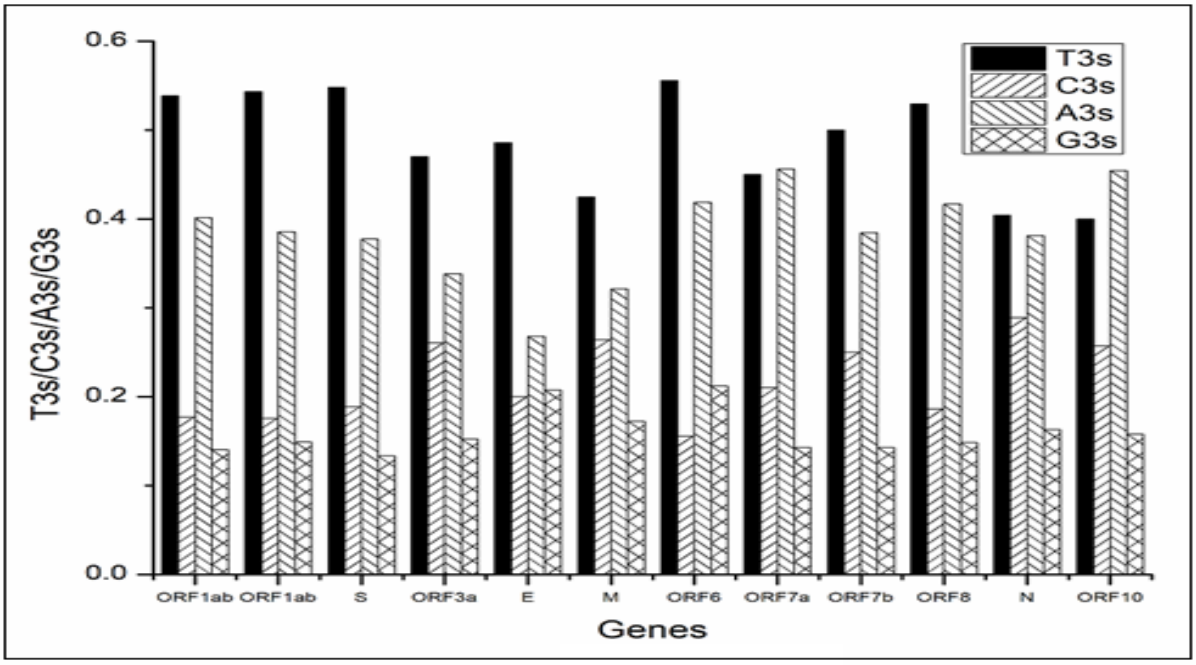


**Fig. 2.** A3, C3, G3, and T3 (A, C, G, T at third site of codons), and GC3s (at synonymous third codon) in all coding sequences of SARS-CoV-2 genome.

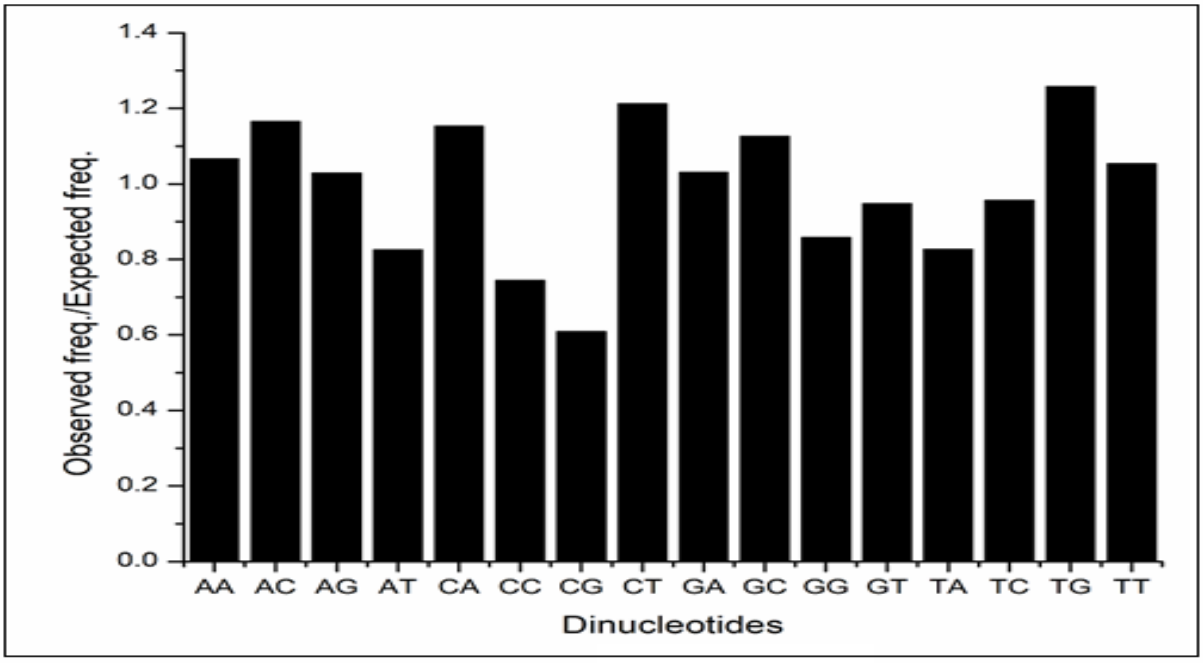


**Fig. 3.** Ratio of observed frequency and expected frequency of dinucleotides used in SARS-CoV-2 genomes under study.

The ENC-GC3 plot, which indicates that mutation plays a role in defining the codon usage variation among those genes, we observe that all the data points of the ENC-GC3s plot [Figure-4] fall below the expected curve**.** It indicates that not only mutation but also other factors, such as natural selection, are

likely to be involved in determining the selective constraints on codon bias influencing the codon usage in SARS-COV-2.The mutations are independent single-site events; however, the base frequencies of the third codon position in SARS-CoV-2 genes are influenced by bases at the first and second codon positions. The correlation (r=0.268) between GC3 and GC12 is statistically significant, suggesting that codon usage is influenced by the mutational pressure. In the neutrality plot [Figure-5], except for ORF10, all genes are far from the regression line. The slope of the regression line indicates a significant impact (39.8%) of neutral evolution i.e.mutational pressure on codon preference in SARS-COV-2 genes. But, the mutational force is not the major factor affecting the codon usage.

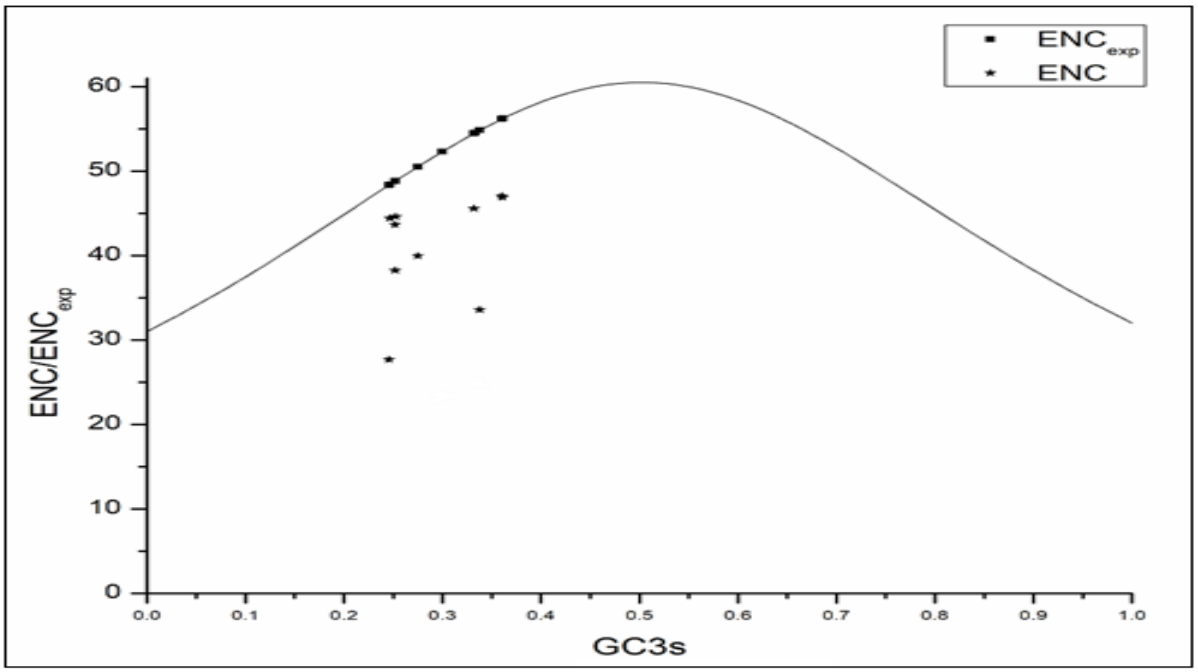


**Fig. 4.** ENC of protein coding genes of SARS-CoV-2 were plotted against GC content at the third codon position (Nc-plot). The expected ENC from GC3 are shown as a continuous curve.

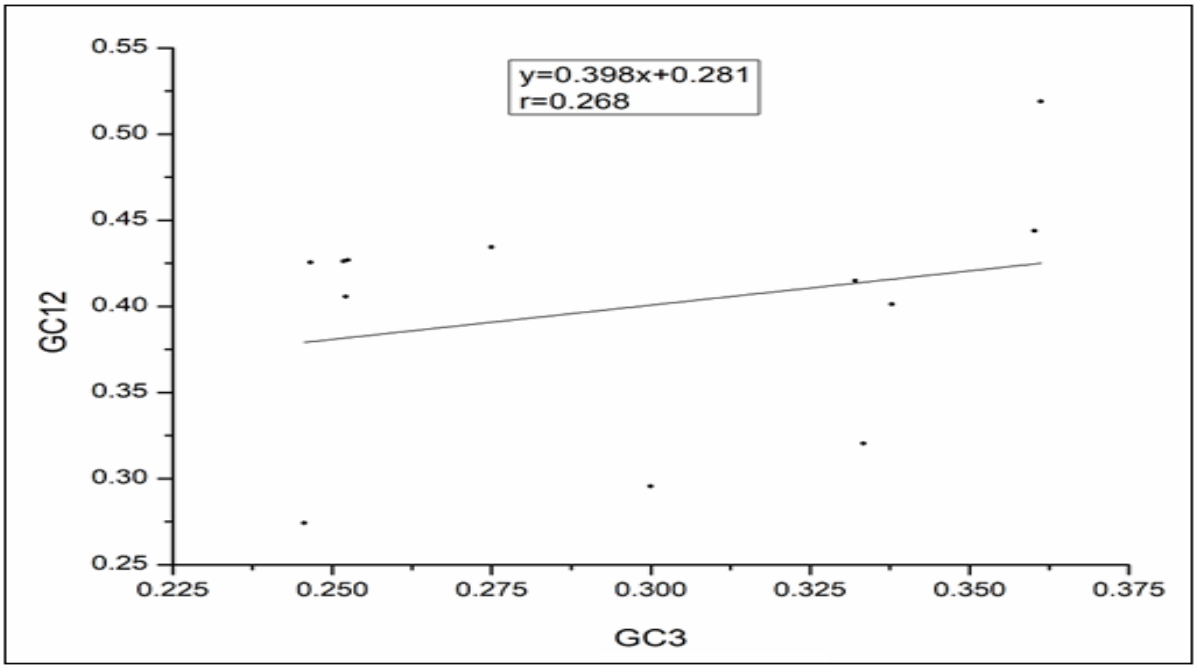


**Fig. 5.** Neutrality plot (GC12 against GC3). The regression line is y=0.398x+0.281, r =0.268.

The codon usage bias effectively regulates the codon usage pattern of a genome. Irrespective of synonymous codon usage, the codon usage in respect of whole genome perspective may be described by RCBS [Figure-6]. In the present study, we observed that GCU for Ala, AGA for Arg, AAC for Asn, GAC for Asp, UGC and UGU for Cys, CAA and CAG for Gln, GAA and GAG for Glu, GGA,GGC,GGU for Gly, CUC,CUU, UUA, and UUG for Leu, AAA and AAG for Lys, AUG for Met, UUC and UUU for Phe, CCA and CCU for Pro, UCA for Ser, ACA and ACU for Thr, UGG for Trp, and UAC for Tyr are the overrepresented codons (RCBS>1) in SARS-CoV-2 genes.

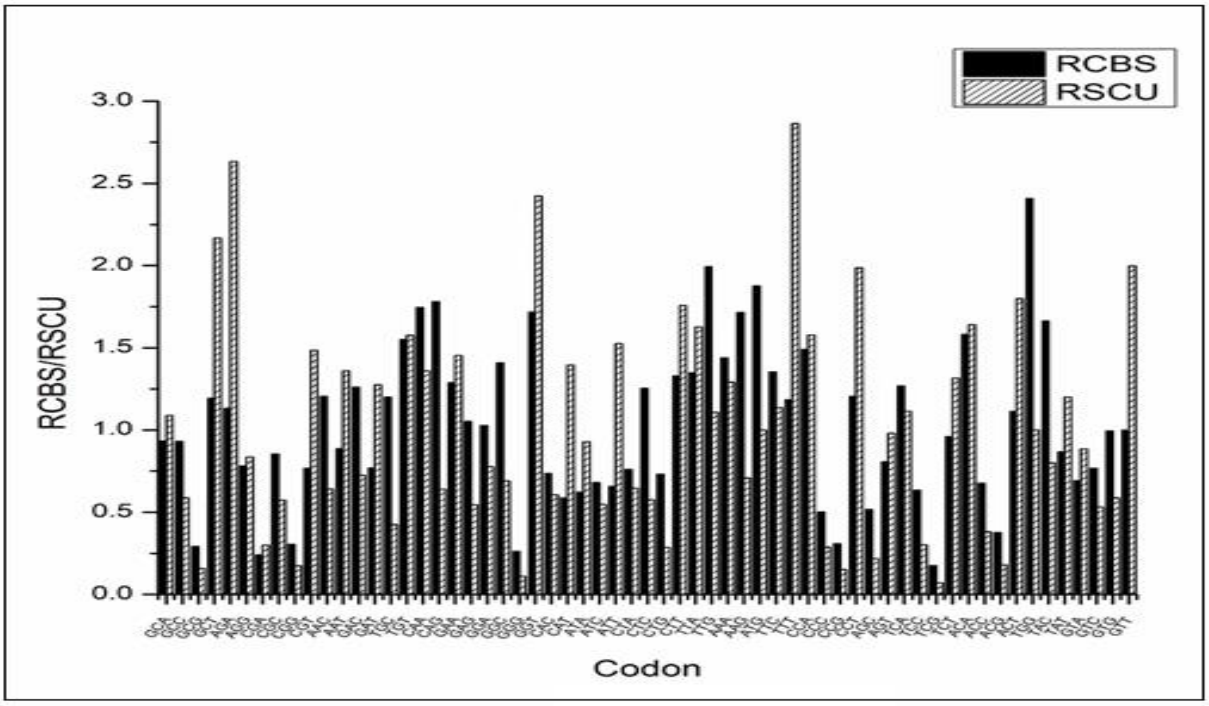


**Fig. 6.** RSCU and RCBS values of codons used in SARS-CoV-2 genome under study.

The relative abundance of dinucleotides ApA,ApC, ApG, CpA, CpU, GpA,GpC,UpG,UpU are reflected in the codon usage bias. RSCU values of 59 synonymous codons, excluding AUG and UGG, which just encode one amino acid, were calculated to explore the influence of synonymous codon usage bias in SARS-CoV-2 genes. As for the synonymous codon usage, the virus has evolved to a set of preferred 26 synonymous codons (RSCU>1) [Figure-6].

The results showed that, among the 26 preferred synonymous codons, fifteen codons terminated with U, which were GCU for Ala,GGU for Arg,AAU for Asn,GAU for Asp,UGU for Cys,GGU for Gly,CAU for His, AUU for Ile,CUU for Leu,UUU for Phe,CCU for Pro,UCU for Ser,ACU for Thr,UAU for Tyr and GGU for Val, and nine codons terminated with A , which were GCA for Ala, AGA for Arg, CAA for Gln, GAA for Glu, UUA for Leu, AAA for Lys, CCA for Pro, UCA for Ser, ACA for Theo, one ended with G, which was UUG for Leu and one ended with C which was UUC for Phe. It is interesting that codons ending with U were the most frequently employed among the twenty six synonymous codons, which was in according with the result of U being the most abundant among the third position of the four kinds of nucleotides, indicating that nucleotide bias was displayed in the SARS-CoV-2 genes. Thus, the preferred codons were influenced by compositional constraints. Although, different synonymous codons favoured by an organism for translational efficiency in different genes are identified by RSCU, the set of optimal codons used in a gene effectively measures its expressivity.The optimal codons enhance the rate of elongation while non-optimal codons slows it down[20].In the present study, we observed that GCU(Ala), AGA(Arg),

UGU(Cys), CAA(Gln) ,GAA(Glu), GGU(Gly), CUU, UUA ,UUG(Leu), AAA(Lys), UUC, UUU(Phe), CCA, CCU(Pro), UCA(Ser), ACU, ACA (Thr), are optimal codons in SARS-CoV-2 genes whereas, rare codons are GCG (Ala), CGA, CGG (Arg), GGG (Gly), CCC, CCG (Pro), AGC, UCG (Ser), ACG (Thr). The low relative abundances of CpG, CpC are reflected in the set of rare codons which are associated with generally slower rate of protein synthesis. This indicates that the selection pressure leading to low CpG and CpC is actively involved in the codon usage pattern of SARS-CoV-2.

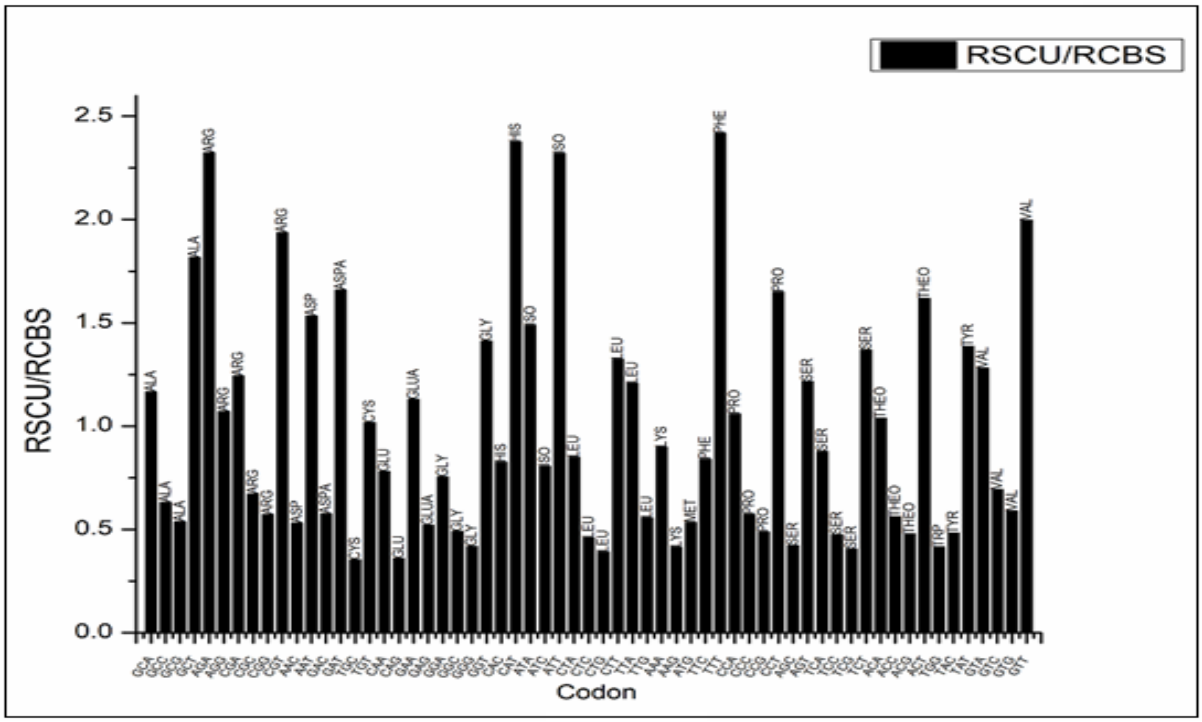


**Fig. 7.** Ratio of RSCU and RCBS of codons used in SARS-CoV-2 genomes under study.

The optimal and rare codons appeared to be highly structured. We found that optimal codons (RSCU value>1, RCBS>1) were A-ended or U-ended, and rare codons (RSCU<0.5 and RCBS<0.5) are G-ended or C-ended. The codon optimality has been shown to affect mRNA stability due to its role in affecting translation elongation[21]. To explore the amino acid usage trend in SARS-CoV-2 genes, we calculated the number of each amino acid for all ORFs across the genome. Wide variation of amino acid usage was observed among genes. The mean amino acid usages of leucine, valine, serine, isoleucine and serine were high for the novel virus, while amino acids such as cysteine, histidine, methionine and tryptophan were low [Figure-8]. Cys(UGU,UGC) is absent in nucleocapsid (N) gene; His(CAU,CAC), Gln(CAA,CAG) and Trp (UGG) is absent in envelope (E)gene. To find out the putative relationship of base composition at different synonymous codon position, when the values of A, U, G, C and GC content were compared with the A3, U3, G3, C3 and GC3 content i.e. at third synonymous codon positions, the significant positive correlations were observed only between A and A3 (r = 0.634), C and C3 (r = 0749), GC and GC3(r=0.568). Moreover, significant positive correlation between C and GC3 (r = 0.711) and a negative correlation between C and T3(r=-0.722) suggested that selection force along with mutational pressure both have significantly influence the codon usage pattern in SARS-CoV-2 genes. Based on the statistical analyses on codon usage biases among the nine CDSs, we observed that selection forces, along with mutational pressure were found to play a major role in shaping SARS-CoV-2 codon usage. CAI is a statistical method used to predict gene expression levels and to analyze codon usage bias, which refers to differences in the frequency of occurrence of synonymous codons in coding DNA.

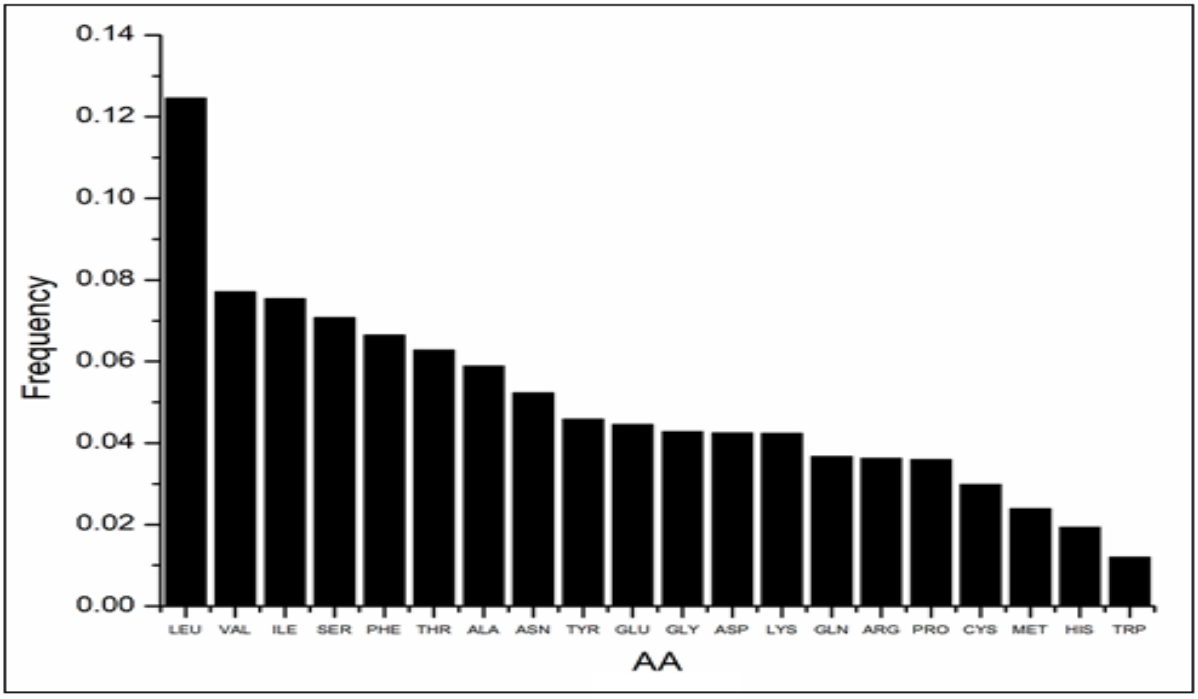


**Fig. 8.** Frequency of amino acids (AA) used in protein coding genes of SARS-CoV-2 genomes under study.

This variation in codon bias can also be interpreted as evidence of selective pressure on the usage of synonymous codons. It is hypothesized that CAI values are associated with selection pressure, with highly expressed codons selected. The strong correlation between CAI and GC3s (r=-0.921) [Figure-9] indicates that codon bias in SARS-CoV-2 genes are strongly influenced by mutational force.

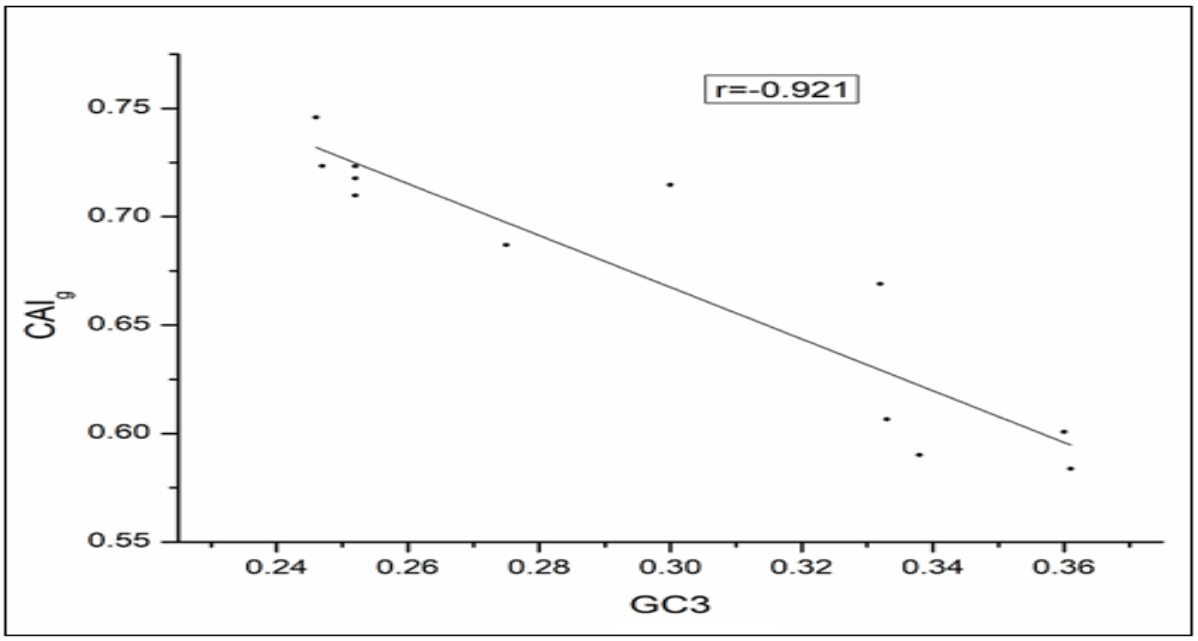


**Fig. 9.** CAI plotted against GC3 for each protein coding genes of SARS-CoV-2 genomes under study.

A comparative analysis based on CAI values of SARS-CoV-2 genes among different hosts considered for the present study are used to determine the adaptation of virus to a host. The higher are the CAI values, the viral genes might be more adapted to a particular host. The CAI values (Table-1) for the potential hosts indicated that host adaptation of SARs-CoV-2 was greatest for *Marmota monax*, followed by

*Rhinolophus ferrumequinum, Gallus gallus, Naja atra, Bungarus multicinctus*; while it was least for human. Compared to CAI value of SARS-CoV-2 in respect of its whole genome perspective, codon usage of human is less supportive compared to other hosts [Fig.-10]. Interestingly, we observe that CAI values of N gene and M gene of SRARS-CoV-2 calculated in reference to RSCU of human genes are greater than that calculated in respect of the virus genome and the difference is significantly large for N gene indicating its outstanding feature of codon usage to allow higher exposition to human(host) immune system.

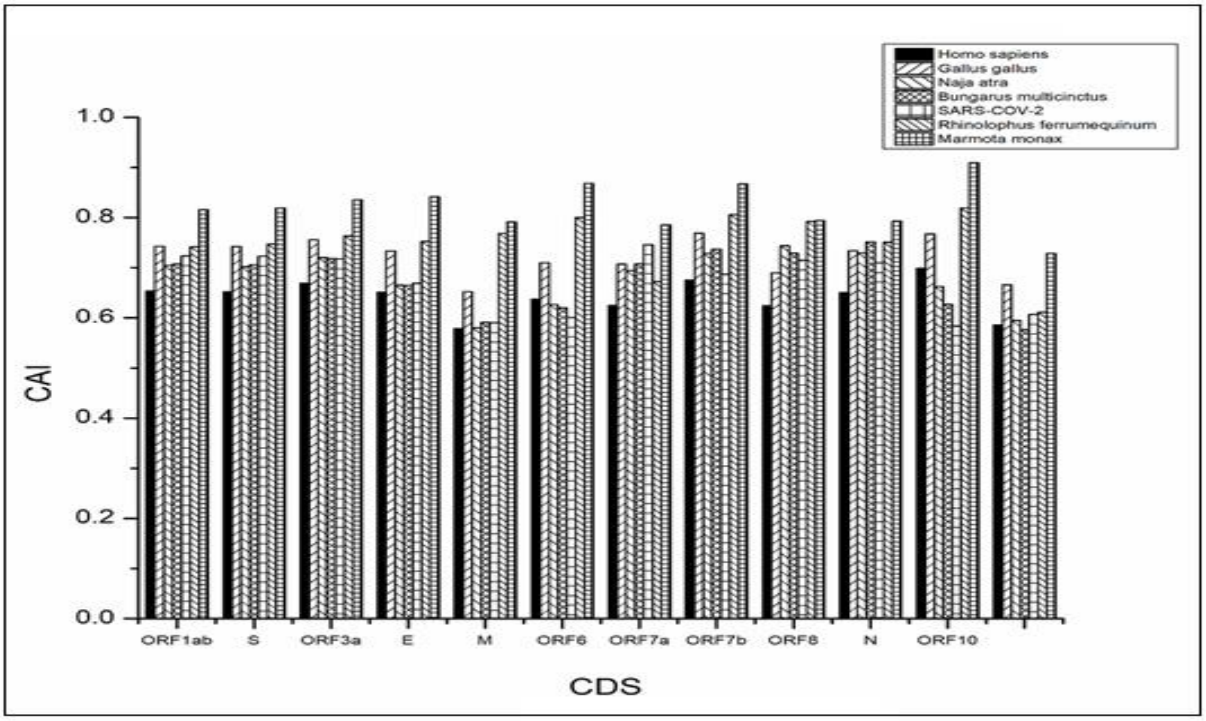


**Fig. 10.** CAI with reference to different animal hosts plotted against each protein coding genes of SARS-CoV-2 genomes under study.

For understanding the possible co-evolution of virus and host genomes, deoptimization is measured by RCDI values by comparing the codon usage of a virus with that of its host [Figure-11]. A lower RCDI value indicates higher adaptation of a virus to the translational machinery of its host. The lowest mean RCDI value (1.703±0.448) of *Bungarus multicinctus* suggested that the virus replicated best in *Bungarus multicinctus*. The significant negative correlation between CAI and RCDI for *Rhinolophus ferrumequinum* (r=-0.864), *Marmota*(r=-0.770), *Gallus gallus*(r=-0.737), *Naja atra*(r=-0.646) indicated that the codon usage pattern of the viral genes effectively used the translational machinery of the host [Table-2]. The weak negative correlation in case *Bungarus multicinctus* with lowest RCDI value and relatively low CAI values with respect to other hosts, indicating the potential of the virus to be replicated successfully in a host with different codon usage patterns.The average RCDI value of 2.281 ± 0.476 for human indicates that SARS-CoV-2 is less adapted to human compared to other hosts considered in the present study. The higher value of RCDI>2 may reflect the expression of genes at low translational rate to achieve error-proof translation of viral proteins[22]. In respect of human, the strong codon deoptimization was observed in all accessory genes and structural genes except M.

The structural gene M is comparatively more adapted to host and showed low RCDI value(1.622) indicating its significant role in the pathogenesis of the virus in human. CAI values of SARS-CoV-2 genes using codon usage of *Bungarus multicinctus* is very much comparable in respect of whole genome of the virus itself.

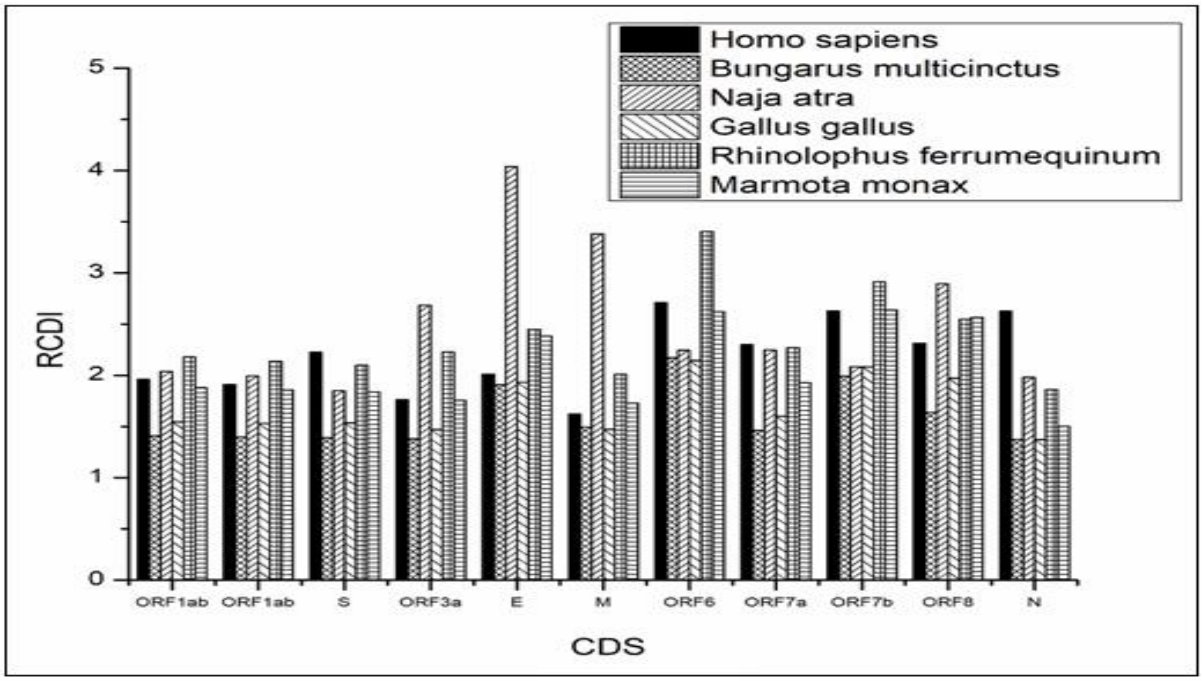


**Fig. 11.** RCDI with reference to different animal hosts plotted against each protein coding genes of SARS-CoV-2 genomes under study.

The observed data suggest that the SARS-CoV-2 might more effectively use snake's (*B. multicinctus*) translation machinery than that of other animals. Thus, *B. multicinctus* might serve as intermediate hosts that transmit viruses to humans.

## Conclusions

SARS-CoV-2, a highly contagious virus, spreading rapidly across the world, has now become a severe threat to public health. The future evolution, adaptation, and spread of this virus warrant urgent investigation**.** In the present study, we comprehensively analyzed the codon usage pattern of SARS-CoV-2 to gain an insight into the mechanism of viral pathogenesis. Based on the codon usage, we also attempted to discuss the possible coevolution and its adaptation to animal hosts. Taking all the results together, our studies reveal that SARS-CoV-2 has a relatively low codon usage bias, which is shaped by both mutation pressure and natural selection. Our results revealed that the novel virus is less adapted to human compared to other hosts as evidenced by CAI values. RCDI analysis indicated that *B. multicinctus* might serve as intermediate hosts that transmit viruses to humans. The result of this study might also be useful to optimizing protein expression. The codon usage information can also be used to reduce the viral protein synthesis by choosing underrepresented codons or by increasing rare dinucleotides like CpG in a tunable manner during replication of the pathogen.

The information on codon usage of SARS-CoV-2 may pave the way to design strategies such as the use of the least preferred codons to modify the SARS-CoV-2 genome to reduce virulence for the development of a safe and effective vaccine.

## Conflicts of Interest

The authors declare that they have no conflicts of interest related to this study.

## Funding

The authors acknowledge the Science and Engineering Research Board, DST, Govt. of India for the financial support under the fixed grant scheme MATRICS [File No: MTR/2019/000274].

## References

**Wu F, Zhao S, Yu B.** 2020.A new coronavirus associated with human respiratory disease in China. Nature **579,** 265–269.
https://doi.org/10.1038/s41586-020-2008-3

**Chan JFW, Kok KH, Zhu Z,Chu H, To KKW, Yuan S, Yuen KY.** 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerging Microbes & Infections **9(1),** 221–236. https://doi.org/10.1080/22221751.2020.1719902

**Wu A, Peng Y, Huang B.** 2020a. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China, Cell Host & Microbe **27,** 325-328. https://doi.org/10.1016/j.chom.2020.02.001.

**Liu Y.** 2020. A code within the genetic code: codon usage regulates co-translational protein folding, Cell Communication and Signaling **18,** 145. https://doi.org/10.1186/s12964-020-00642-66.

**Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S**. 2008. Virus attenuation by genome-scale changes in codon pair bias. Science **320(5884),** 1784-1787. https://doi.org/10.1126/science.1155761.

**Baker SF, Nogales A, Martínez-Sobrido L**. 2015. Downregulating viral gene expression: codon usage bias manipulation for the generation of novel influenza A virus vaccines: Future Virology **10(6),** 715–730. https://doi.org/10.2217/fvl.1531.

**Lytras S, Hughes J.** 2020. Synonymous Dinucleotide Usage: A Codon-Aware Metric for Quantifying Dinucleotide Representation in Viruses. Viruses **12,** 462. https://doi.org/10.3390/ v12040462.

**Sharp PM, Wen-Hsiung L.** 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. Journal of Molecular Evolution **24,** 28–38. https://doi.org/10.1007/BF02099948.

**Sahoo S, Das S.** 2014. Analyzing gene expression and codon usage bias in diverse genomes using a variety of models. Current Bioinformatics **9**, 102-112. https://doi.org/10.2174/157489360899914010911424 7.

**Rakshit R, Sahoo S.** 2017. In Silico Prediction of Gene Expression Based on Codon Usage: A Mini Review. Journal of Investigative Genomics **4(2),** 42-45. https://doi.org/10.15406/jig.2017.04.00063.

**Wright F.** 1990.The 'effective number of codons' used in a gene. Gene 87, 23–29. https://doi.org/10.1016/0378-1119(90)90491-9.

**Khandia R, Singhal S,Kumar U, Ansari A, Tiwari R, Dhama K, Das J, Munjal A, Singh RK.** 2019. Analysis of Nipah Virus Codon Usage and Adaptation to Hosts. Frontiers in Microbiology **10,** 886. https://doi.org/10.3389 /fmicb.2019.00.886

**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 Research International **2013,** Article Id 406342. https://doi.org/ 10.1155/2013/406342.

**Sharp PM, Li WH.** 1987. The codon adaptation index– a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research **15,** 1281– 1295. https://doi.org/10.1093/nar/15.3.1281.

**Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E.** 2006. Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. J. Virology **80,** 9687–9696. https://doi.org/10.1128/JVI.0073.8-06.

**Butt AM, Nasrullah I, Qamar R, Tong Y.** 2016. Evolution of codon usage in Zika virus genomes

is host and vector specific. Emerging Microbes Infection **5,** e107. https://doi.org/10.1038/emi.2016.106

**Belalov IS, Lukashev AN.** 2013.Causes and implications of codon usage bias in RNA viruses. PLoS One **8**, e56642. https://doi.org/10.1371/journal.pone.0056642.

**Simmonds P, Xia W, Baillie JK, McKinnon K.** 2013. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla –selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genomics **14,** 610. https://doi.org/10.1186/1471-2164-14-610.

**Jang HS, Shin WJ, Lee JE, Do JT.** 2017. CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. Genes (Basel) **8(6),** 148. https://doi.org/10.3390/genes 8060148.

**Vetsigian K, Goldenfeld N.** 2009. Genome rhetoric and the emergence of compositional bias. PNAS **106,** 215–220. https://doi.org/10.1073/pnas.0810122106.

**Zhao F, Yu CH, Liu Y.** 2017. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. Nucleic Acids Research **45(14),** 8484–8492. https://doi.org/10.1093/nar/gkx501.

**Puigbo P, Aragones L, Garcia-Vallve S.** 2010. RCDI/eRCDI: a web-server to estimate codon usage deoptimization. BMC Research Notes **3,** 87. https://doi.org/10.1186/1756-0500-3-87