INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Bioinformatics Research and Applications

ISSN online: 1744-5493 - ISSN print: 1744-5485
https://www.inderscience.com/ijbra

Codon usage in conserved sites is more biased compared to variable sites in the SARS-CoV-2 genome

Madhusmita Dash, Annushree Kurmi, Preetisudha Meher, Siddhartha Sankar Satapathy, Nima D. Namsa

DOI: 10.1504/IJBRA.2024.10060747

Article History:

Received:	17 May 2023
Last revised:	19 May 2023
Accepted:	19 July 2023
Published online:	14 March 2024

Codon usage in conserved sites is more biased compared to variable sites in the SARS-CoV-2 genome

Madhusmita Dash*

Department of Electronics and Communication Engineering,
NIT Arunachal Pradesh, Jote,
District: Papum Pare, Arunachal Pradesh – 791113, India
Email: madhusmita.dash81@gmail.com
*Corresponding author

Annushree Kurmi

Department of Computer Science and Engineering,
Tezpur University, Napaam,
Sonitpur – 784028, Assam, India
Email: annushreekurmi@gmail.com

Preetisudha Meher

Department of Electronics and Communication Engineering,
NIT Arunachal Pradesh, Jote,
District: Papum Pare, Arunachal Pradesh – 791113, India
Email: preetisudha1@gmail.com

Siddhartha Sankar Satapathy

Department of Computer Science and Engineering,
Tezpur University, Napaam,
Sonitpur – 784028, Assam, India
Email: ssankar@tezu.ernet.in

Nima D. Namsa

Department of Molecular Biology and Biotechnology,
Tezpur University,
Napaam-784028, Assam, India
Email: namsa@tezu.ernet.in

Abstract: High error rate in SARS-CoV-2 genome replication allows the virus to adapt to different environments and selective pressures. In this study, 35% of codons of the protein-coding sequences of the genome were observed to have undergone base substitution mutations. Machine learning based comparative analysis of usage between conserved codons and the remaining variable codons of the protein-coding genes revealed that the codon usage patterns between the

two groups are significantly different. Codon usage values in the variable region resemble genome composition, whereas the values in the conserved regions were highly variable. This differential codon usage suggests that the conserved regions are under influence of selection pressure in this virus genome. Further, the selection pressure on codon usage and the nucleotide substitution biases act towards increasing A and T base composition in SARS-CoV-2 genome. Our observations on the base substitution will help us in understanding evolution of this SARS-CoV-2 virus genome.

Keywords: SARS-CoV2 genome; base substitution mutation; selection; conserved region; codon usage bias; CUB.

Reference to this paper should be made as follows: Dash, M., Kurmi, A., Meher, P., Satapathy, S.S. and Namsa, N.D. (2024) 'Codon usage in conserved sites is more biased compared to variable sites in the SARS-CoV-2 genome', *Int. J. Bioinformatics Research and Applications*, Vol. 20, No. 1, pp.42–60.

Biographical notes: Madhusmita Dash is a PhD student in the Department of Electronics and Communication Engineering, National Institute of Technology, Arunachal Pradesh, India. She received her MTech and BTech degrees from Biju Patnaik University of Technology Rourkela Odisha. Her research interests include data science, IOT and genomics.

Annushree Kurmi is a PhD student in the Department of Computer Science and Engineering, Tezpur University, Assam, India. She received her MTech in Information Technology from Tezpur University. At present, she is also working as an Assistant Professor in the Department of Computer Science and Enginerring in The Assam Kaziranga University. Her research interests include bioinformatics, machine learning and software engineering.

Preetisudha Meher is an Assistant Professor in the Department of Electronics and Communication Engineering, NIT Arunachal Pradesh. She has received her PhD from National Institute of Technology (NIT), Rourkela, Odisha, India and MTech from National Institute of Science and Technology (NIST), Berhampur, Odisha, India. Her research interest includes low power VLSI, digital VLSI, embedded systems, IOT, and bioinformetics.

Siddhartha Sankar Satapathy is an Associate Professor in the Department of Computer Science and Engineering, Tezpur University, Assam, India. He received his MTech and PhD degrees from Tezpur University. His research interests include bioinformatics and ad hoc networks.

Nima D. Namsa is an Assistant Professor in the department of Molecular Biology and Biotechnology, Tezpur University, Assam. He was a JRF and SRF in the Department of Microbiology and Cell Biology, Indian Institute of Science, Bangalore. He received his Post-PhD training from the Department of Microbiology and Immunology, School of Medicine, Stanford University, California, USA and PhD degree from the of Department of Microbiology and Cell Biology, Indian Institute of Science, Bangalore. His research interests include rotavirus, scrub typhus, and molecular evolution of RNA viruses.

1 Introduction

The unequal usage of synonymous codons, known as codon usage bias (CUB), is determined by the combined influences of selection and mutation forces in the genome of any organism (Grantham et al., 1980; Ikemura, 1981, 1985; Sharp and Li, 1986; Sharp et al., 2010). The preferential usage of specific codons among the synonymous ones in high-expression genes indicates influences of the translational selection on codon choices (Sharp et al., 2005). When selection is weak, codon usage in an organism is primarily influenced by complex mutational forces and drift (Bulmer, 1991). Nucleotide substitution, context-dependent mutation, and repair-associated bias are the prominent genome-wide mutational factors acting on codon usage (Knight et al., 2001; Plotkin and Kudla, 2011). Organisms prefer specific codons among synonymous codons recognised faster by cognate tRNA (Ikemura, 1981, 1985). Alternatively, a codon interacting more effectively with the anticodon over another synonymous codon can be selected for optimum translation. Codons preferred in high-expression genes compared to low-expression genes for fast or accurate translation are considered optimal codons (Ikemura, 1981). This translational selection has been shown across bacteria (Sharp et al., 1988; Akashi, 2003; dos Reis et al., 2003) and more prominently among bacteria with rapid growth rates (Sharp et al., 2005) because fast translation is a vital necessity for them.

In general, a virus depends on its host's cellular mechanism for replication, survival, and evasion from the host's immune system. The CUB has been reported to adapt to the host for replicative fitness and virulence in the different viral genomes (Burns et al., 2006; Mueller et al., 2006; Costafreda et al., 2014). CUB in virus genomes is being studied from different points of view, such as adaptation to their hosts (Tian et al., 2018), the extent of respiratory virulence (Chen and Yang, 2022), and the compositional difference between conserved and variable amino acid residues (Klitting et al., 2016). Genome composition in RNA viruses reflects codon usage and therefore CUB is considered to be mainly under mutational pressure (Jenkins and Holmes, 2003; Belalov and Lukashev, 2013; Yao et al., 2020). Apart from the genome composition, limited evidence of host-specific translational selection pressure also has been reported in human RNA viruses (Jitobaom et al., 2020), in the MERS-CoV (Hussain et al., 2020) and influenza virus H1N1 (Wong et al., 2010). A recent study reported the coevolution of virus and host codon usage (Chen et al., 2020). The coronavirus SARS-CoV-2 causal agent of severe acute respiratory syndrome (SARS) disease originates in Wuhan, China; the genome evolves due to rapid mutation inside the host cell (Denison et al., 2011). World Health Organisation (WHO) classified the novel coronavirus pneumonia epidemic caused by SARS-CoV-2 as a public health emergency of international concern in 2020 (Yang and Wang, 2020; Zheng, 2020). Similar to the other viral genome, CUB in SARS-CoV-2 is mainly influenced by mutational pressure (Daron and Bravo, 2021). Gene-specific mutation pattern in SARS-CoV-2 is believed to positively impact viral evolution by increasing its adaptation to human codon usage (Chen and Yang, 2022; Ramazzotti et al., 2022).

The underlying mechanism of the rapid evolution of SARS-CoV-2 is yet to be thoroughly investigated. In this study, we have done a detailed computational analysis on CUB and base substitution dynamics in this viral genome considering a global dataset of

25,135 filtered sequences of the SARS-CoV-2 genome from the early phase of the pandemic. We observed that the codon usage in the conserved sites is highly biased, whereas the variable site codon usage resembles genome composition indicating conserved sites are under more substantial selection. Further, organisms differ with regard to the identities of the selected codons (Sharp et al., 1988). The base composition of the selected codons may or may not match the genome base composition. In the case of the former, (G+C)-rich genomes tend to have (G+C)-rich selected codons, while (A+T)-rich microorganisms tend to have (A+T)-rich selected codons. If the G+C composition of selected codons matched with the genome G+C composition, it would indicate that the selection on codon usage and the nucleotide substitution biases act in the same direction to determine the nucleotide content of a genome (Hershberg and Petrov, 2009). Our study results demonstrate that the selection on codon usage and nucleotide substitution biases are acting in the same direction towards increasing A and T bases in this SARS-CoV-2 virus genome.

2 Materials and methods

2.1 SARS-CoV-2 CDS sequences considered for codon usage analysis

We have considered the SARS-CoV-2 virus reference genome (NC_045512.2) consisting of 29,903 bases reported in the NCBI database for our codon usage study. We also took these genes' protein-coding sequence (CDS) annotations from the NCBI database. The genome of SARS-CoV-2 consists of twenty-six protein coding genes out of which sixteen non-structural protein genes are nsp1 through nsp16 which are distributed over one open reading frame (ORF1), four genes E, M, N, and S code for structural proteins and the remaining six genes code for accessory proteins ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10 (Holmes, 2003; Angeletti et al., 2020; Mousavizadeh and Ghasemi, 2021). Codon usage of all these twenty-six genes we considered in our analysis. Unlike the other protein coding genes, the nsp12 gene that encodes for RNA-dependent RNA polymerase (RdRp) includes a ribosomal –1 frameshift site at base position 13468. Considering the repetition of the base position at 13468, we reconstructed the ORF of the nsp12 gene for codon usage analysis.

2.2 Segregating conserved codons from the variable ones in SARS-CoV-2 CDS sequences

Considering a simple computational approach, we have estimated conserved and variable codon positions in the virus genome. The methodology for finding conserved and variable codons is illustrated with an alignment of the sample set of ten strains of the nsp11 gene in Table 1. In the CDS of a protein-coding gene, if a codon position has not undergone any base substitution in the alignment of strains, the codon is considered conserved. Possibly any change in that codon position is deleterious for the protein, and therefore no base substitution leading to either synonymous or non-synonymous codon change was observed. Whereas, if a codon has undergone base substitution at least in one of the strains, it is considered variable. We assumed that the variable positions could

accommodate mutations without impacting the protein function. A similar method of aligning a large number of sequences to find base substitutions has been used in recent studies (Aziz et al., 2022). A base substitution in a codon might result in to synonymous change that do not alter encoded amino acid or might change the encoded amino acid resulting non-synonymous change. We considered codons with both synonymous and non-synonymous substitutions as variable ones. We wrote computer programs for finding base substitutions and conserved/un-conserved codons in Python.

For the base substitution analysis, we have considered 25,135 SARS-CoV-2 strains downloaded from the GISAID database (https://www.gisaid.org/). We filtered out these strains from the 46,076 high-coverage SARS-CoV-2 strains sequences reported in the GISAID database until 24th July 2020, sampled from patients in 95 countries. All these fully sequenced SARS-CoV-2 virus strains were of size 29,903 bases each and had no internal stop codons, deletions, and ambiguous nucleotides other than A/T/G/C. We created a local BLAST database of all these strains and extracted the alignment of the individual genes from the database for finding base substitutions.

2.3 Estimating selection on codon usage bias in SARS-CoV-2

Comparative analysis based on gene expression helps estimate translational selection in organisms, where the expression level of the genes is known either experimentally or from the gene information. However, the virus genome is highly compact, consisting of very few genes that are considered essential for the survival of the virus. Virus genomes do not encode any tRNA, and the protein gene translation relies on the tRNA of their hosts (Albers and Czech, 2016; Tian et al., 2018). Previous studies have reported substantial similarities between the codon usage of virus and their hosts (Lucks et al., 2008; Bahir et al., 2009). Therefore, gene expression-based comparative analysis of codon usage is not feasible to study selection in the viral genome. However, in the viral genome, selection on codon usage can be feasible to study using an alternate approach. Vital sites of a protein that are important for its function might have conserved codons. Translational accuracy might be most important at these vital sites; hence, the preferred codons are expected to be used at evolutionarily conserved sites. A comparative codon usage analysis between conserved sites and the remaining variable sites in a gene sequence can be a good strategy for understanding selection mechanisms (Jia and Higgs, 2008). As discussed earlier, we segregated conserved sites from the variable sites considering a large set of viral strains from the public database sampled from human patients across the globe for understanding the selection pressure on the SARS-CoV-2 genome.

2.4 Codon usage bias measures

Different measures have been proposed to measure biased codon usage in the coding sequence of genes (Roth et al., 2012). One of the popular measures is the effective number of codons (N_c) (Wright, 1990) that estimates overall CUB in a gene. For a gene, N_c values can vary between 20.0 and 61.0 for uniform and extremely biased codon usage scenarios. Mathematically N_c is defined as follows:

Table 1 Conserved and un-conserved amino acid positions in the nsp11 gene

Sl. no.	*Info*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
1	>hCoV-19/England/LIVE-99A77/2020	TCA	GCT	GAT	GCA	CAA	TCG	TTT	TTA	AAC	GGG	TTT	GCG	GTG
2	>hCoV-19/India/ILSCV20166/2020	TCA	GCT	**GGT**	GCA	CAA	TCG	TTT	TTA	AAC	GGG	TTT	GCG	GTG
3	>hCoV-19/Scotland/CVR3377/2020	TCA	GCT	GAT	GCA	CAA	TCG	TTT	TTA	AAC	GGG	TTT	**TCG**	GTG
4	>hCoV-19/Scotland/EDB153/2020	TCA	**TCT**	GAT	GCA	CAA	TCG	TTT	TTA	AAC	GGG	TTT	GCG	GTG
5	>hCoV-19/Senegal/62365/2020	TCA	GCT	GAT	GCA	CAA	TCG	TTT	TTA	AAC	**GGA**	TTT	GCG	GTG
6	>hCoV-19/SouthKorea/CoV55/2020	TCA	GCT	GAT	GCA	CAA	TCG	TTT	TTA	AAC	GGG	TTT	GCG	GTG
7	>hCoV-19/Spain/Valencia597/2020	TCA	GCT	GAT	GCA	CAA	TCG	TTT	TTA	AAC	GGG	TTT	GCG	GTG
8	>hCoV-19/Sweden/20-04631/2020	TCA	GCT	GAT	GCA	CAA	TCG	TTT	TTA	AAC	GGG	TTT	GCG	GTG
9	>hCoV-19/USA/NY-PV09362/2020	TCA	GCT	GAT	GCA	CAA	TCG	TTT	TTA	AAC	GGG	TTT	GCG	GTG
10	>hCoV-19/Wuhan/WIV02/2019	TCA	GCT	GAT	GCA	CAA	TCG	TTT	TTA	AAC	GGG	TTT	GCG	GTG

Notes: The table presents a base substitution scenario in an alignment of a sample set of ten nsp11 genes from the GISAID database. In this gene alignment, there is no base substitution in amino acid positions 1, 4–9, 11, and 13, categorised as conserved. In contrast, positions 2, 3, 10, and 12 have base substitutions and are therefore considered variable. Base substitutions in amino acid positions 2, 3, and 12 are non-synonymous, whereas, in position 10, base substitution is synonymous. Codons with base substitutions are shown in boldface and shaded dark.

For an amino acid A with degeneracy k, i.e., with k number of synonymous codons, each with counts $n_1, n_2, \ldots, n_k$, $n = \sum_{i=1}^{k} n_i$ and $p_i = n_i/n$, effective number of codons N_c is calculated as follows:

$$N_c = \frac{1}{F_A} \tag{1}$$

where

$$F_A = \frac{n\sum_{i=1}^{k} p_i^2 - 1}{(n-1)} \quad n > 1 \tag{2}$$

Finally, for the universal genetic code table, the formula of N_c for a gene can be given as:

$$N_c = 2 + \frac{9}{\overline{F}_2} + \frac{1}{F_3} + \frac{5}{\overline{F}_4} + \frac{3}{\overline{F}_6} \tag{3}$$

Here $\overline{F}_i (i = 2,\ 3,\ 4\ and\ 6)$ represents average values of F_A for all the amino acids with degeneracy *i*.

Nucleotide base composition is the primary factor for CUB in a gene sequence. Codon selection for optimum translation is another factor among several other selection factors that also influence codon usage in a gene. The effective number of codons prime (N_c') (Novembre, 2002) is a variant of N_c that measures CUB in a gene due to factors other than nucleotide composition. One advantage of using N_c and N_c' is that, unlike the other CUB measures such as CAI (Sharp and Li, 1987), FoP (Ikemura, 1981), no additional reference information is required for calculating N_c and N_c'values of a gene and these values can be calculated from the gene's nucleotide sequence. In this article, we have used improved implementation of N_c and N_c' (Satapathy et al., 2017) measures available in the web portal (http://agnigarh.tezu.ernet.in/~ssankar/cub.php).

2.5 Relative synonymous codon usage

Relative synonymous codon usage (RSCU) of individual codons is calculated as the ratio of actual codon usage to expected usage when all the synonymous codons are used uniformly in a given gene sequence. Mathematically RSCU is defined as follows

$$RSCU_i = \frac{x_i}{\frac{i}{n}\sum_{i=1}^{n} x_i} \tag{4}$$

Here x_i is the count of a codon i in the given gene sequence encoding an amino acid, and n is the codon degeneracy or the number of synonymous codons coding for that amino acid. RSCU value 1.00 for a codon suggests that the codon is used as expected, and any deviation from 1.00 denotes biased usage. RSCU values of the codons are independent of their degeneracy. We have used these RSCU values in machine learning analysis.

2.6 Machine learning based analysis of codon usage bias in SARS-CoV-2

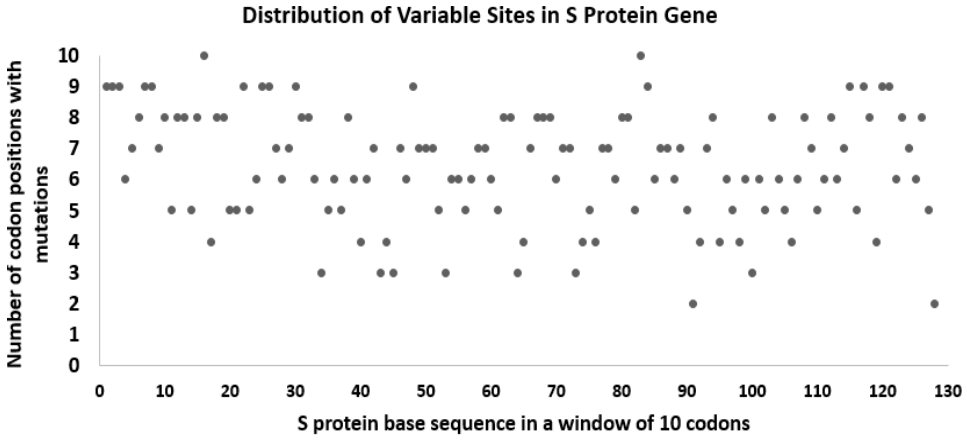
For presenting difference in codon usage between conserved and variable sites, we have used seven machine learning based classifiers, decision tree (Quinlan, 1986), Gaussian Naïve Bayes (GNB) (Pérez et al., 2006), k-nearest neighbours (kNN) (Baek and Sung, 2000), logistic regression (LR) (Hastie et al., 2009), random forest (RF) (Breiman, 2001), support vector machine (SVM) (Noble, 2006) and eXtreme Gradient Boosting (XGB) (Sheridan et al., 2016). Implementation of these classifiers are available in the sklearn library of Python. These methods have been shown to be effective in genome composition analysis in a recent study by Kurmi et al. (2023). In a scenario of clear difference, classifiers would result higher classification scores. For the classification task, in each dataset, we had features in terms of usage of the 61 codons and a target variable with two classes- conserved and variable sites which can be represented as $D = \{X, y\}$, where the feature dataset D is with X columns, and y is the target variable: $X = \{x_1, x_2, x_3, \ldots, x_n\}$, $n = 61$, $y = \{1 \text{ or } 0\}$, 1 represents conserved site and 0 represent variable site.

We used receiver operating characteristics (ROC) analysis to estimate performance of the classifiers. ROC is a 2-D probability plot between true positive rates (TPR) and false positive rates (FPR) and the area under ROC curve (AUC) value gives us an idea about the quality of the ROC curve. Here TPR is defined as the values that are positive in the actual class and have been correctly predicted as positive by the machine learning model. For example, codon usage values are from conserved region of a gene and it has been predicted as conserved. FPR are the values that are negative in the actual class but they have been wrongly predicted as positive by the machine learning model. For example, codon usage values are from variable region of a gene and it has been predicted as conserved.

3 Results and discussion

Before segregating variable and conserved codon positions, we estimated base substitution in all the CDS regions of the SARS-CoV2 genome in order to understand overall pattern of the substitutions. Each of the four bases A, T, G, and C can be substituted by the other three bases, resulting in twelve directional base substitution mutations. Out of the twelve mutations, four mutations (C→T, C→T, A→G, and G→A) are called transitions, and the remaining eight mutations are called transversions. Similar to the reports in the research literature (Lewis et al., 2016), we observed that the amount of transitions was more than the transversions in all the CDS sequences. Among the mutations, C→T was observed to be most frequent, possibly due to the rapid deamination of cytosine to uracil (Lewis Jr et al., 2016). The G→T transversion was the second most frequent base substitution. The other three transitions were more frequent compared to the reaming transversions. These observations resembling previously reported base substitutions in research literature suggested that the computational methodology employed here is correct. Considering a high frequency of transitions in a few RNA viruses reported earlier (Holland, 2006; Simmonds and Ansari, 2021) a surprisingly high frequency of G→T transversion observed in the genome as compared to the G→A, T→C, and A→G transitions suggest that the unusual high frequent G→T transversion might be unique to the genome of SARS-CoV-2.

Figure 1 Distribution of codons with mutations along the gene sequence of s protein



Notes: The figure presents the distribution of variable codons in the S protein gene. Each dot in the Y-axis represents the number of codon positions with mutations (details as described in the MM section) in a window of a sequence of 10 codons represented in the X-axis. It is evident from the figure that the mutations are spread across the length of the S protein gene.

3.1 Variable sites in the SARS-CoV-2 gene are distributed across the gene sequence

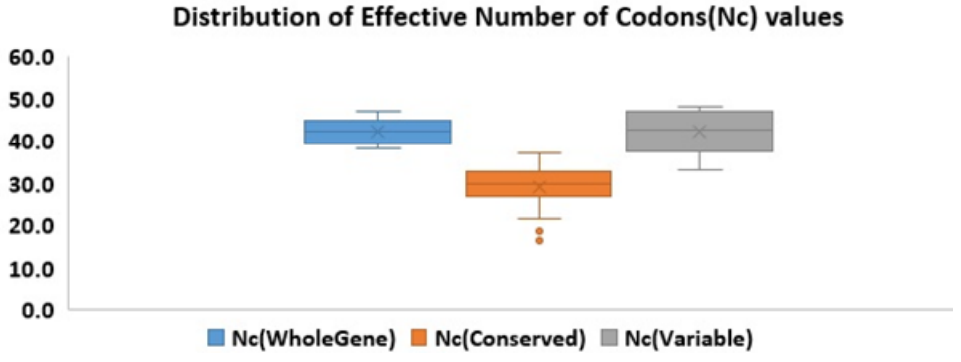
For estimating variable and conserved sites in the coding sequences, we found base substitutions in the alignment of all the genes individually. Base substitutions were abundantly observed along the length of the codon sequence of all the genes. Among the larger genes with a size of more than 100 amino acids, the accessory protein gene ORF3a was found with the least percentage (13.77%) of variable sites. In contrast, the non-structured protein gene nsp10 had the highest rate (48.20%). In general, non-structured protein genes had a high percentage of variable sites, whereas accessory proteins had the lowest proportion. Among the structural protein genes, the portion of variable sites for S, M, N, and E was 35.48%, 31.84%, 20.95%, and 38.16%, respectively. These variable sites were not locally clustered but spread across the length of the ORF of the genes. For example, Figure 1 presents the distribution of codons with mutations along the gene sequence of the S protein.

3.2 Conserved codons are comparatively under stronger selection compared to variable ones in the SARS-CoV-2 genome

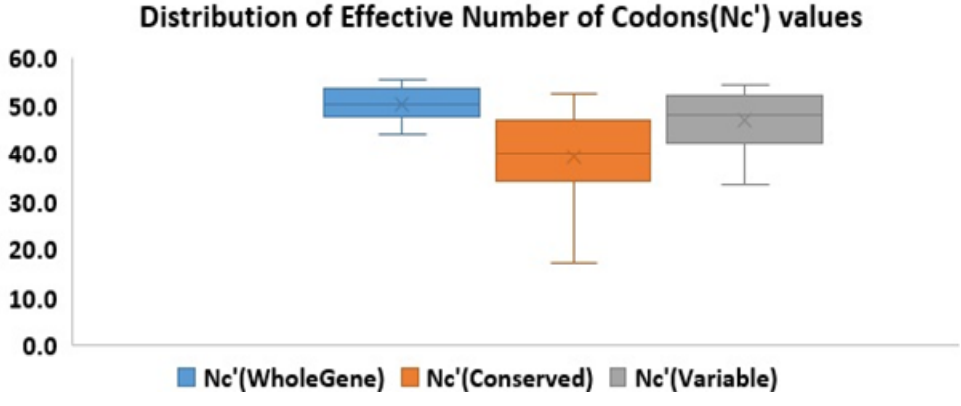
SARS-CoV-2 genome replication typically has a high error rate and is expected to be under weak selection. Therefore, uniform codon usage was expected in the genes. Accordingly, N_c values were expected to be on the higher side in the range of 20.00 to 61.00. We calculated N_c values in the genes as a whole and separately considering only the unserved and un-conserved codons (Table 2). It can be seen from Figure 2(a) that the distribution of the N_c values of the gene sequence considered as a whole and of the un-conserved codons are similar and on the higher side of the range. However, the N_c'values of the conserved codons are significantly (p-value <0.0001) lower than that in

the other two sets. Similar was the observation for the N_c' values given in Figure 2(b). This observation suggested that the conserved codons are under stronger selection than un-conserved codons.

Figure 2 (a) Distribution of N_c values in sars-cov-2 genes (b) distribution of N_c' values in SARS-CoV-2 genes (see online version for colours)



(a)



(b)

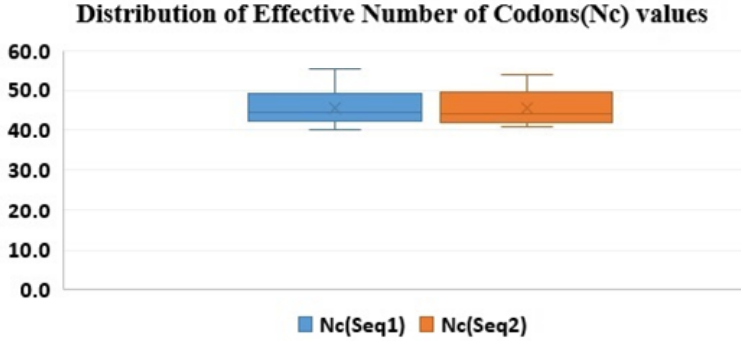
Notes: The figure presents the distribution of N_c values in three different regions of the SARS-CoV-2 genes. Genes with a size of more than 100 codons are considered in this figure. The Y-axis represents the distribution of N_c values, and the X-axis represents the three categories. It is evident from the figure that in the whole gene and variable codons, N_c values are significantly higher compared to conserved codons. The figure presents N_c' values distribution in three different regions of the SARS-CoV-2 genes with a size of more than 100 codons. The Y-axis represents the distribution of N_c' values, and the X-axis represents the three categories. It is evident from the figure that in the whole gene and variable codons, N_c' values are significantly higher compared to conserved codons.

Table 2 Codon usage bias measures N_c and N_c' values of conserved and variable regions in SARS-CoV-2 genes

Sl	*ORF*	*Gene*	*Size*	*GC*	*GC3*	N_c	Nc'	*Conserved*					*Variable*				
								Size	*GC*	*GC3*	N_c	N_c'	*Size*	*GC*	*GC3*	N_c	N_c'
1	ORF1ab	nsp1	180	0.49	0.39	42.66	47.63	47	0.40	0.21	27.16	35.22	133	0.51	0.45	37.72	41.16
2	ORF1ab	nsp2	638	0.40	0.30	46.35	53.06	165	0.29	0.15	30.96	39.65	473	0.44	0.35	48.01	52.94
3	ORF1ab	nsp3	1945	0.36	0.23	42.23	55.25	684	0.27	0.09	32.21	52.36	1261	0.41	0.30	46.61	54.17
4	ORF1ab	nsp4	500	0.37	0.24	41.56	52.12	178	0.32	0.14	32.57	47.06	322	0.39	0.29	42.57	49.35
5	ORF1ab	nsp5	306	0.38	0.22	39.57	48.49	120	0.31	0.07	27.19	40.91	186	0.43	0.32	42.92	47.82
6	ORF1ab	nsp6	290	0.36	0.28	41.08	49.05	98	0.34	0.14	30.68	41.86	192	0.38	0.35	41.15	45.94
7	ORF1ab	nsp7	83	0.38	0.31	34.03	37.27	25	0.29	0.13	16.00	23.31	58	0.42	0.39	28.01	29.21
8	ORF1ab	nsp8	198	0.38	0.25	38.20	46.93	70	0.31	0.24	30.23	36.67	128	0.42	0.26	35.60	42.44
9	ORF1ab	nsp9	113	0.40	0.25	39.17	44.22	39	0.38	0.11	23.52	35.30	74	0.41	0.32	33.25	33.72
10	ORF1ab	nsp10	139	0.42	0.25	42.75	47.90	67	0.38	0.14	28.98	34.16	72	0.47	0.34	37.53	39.06
11	ORF1ab	nsp11	13	0.46	0.39	12.00	16.65	9	0.33	0.33	8.00	12.63	4	0.75	0.50	4.00	2.00
12	ORF1ab	nsp12	932	0.37	0.27	45.13	54.52	405	0.33	0.15	37.17	52.33	527	0.41	0.35	47.59	52.35
13	ORF1ab	nsp13	601	0.38	0.23	43.51	53.74	233	0.31	0.13	35.74	49.70	368	0.43	0.30	47.00	51.64
14	ORF1ab	nsp14	527	0.38	0.25	43.35	53.60	194	0.34	0.12	33.46	47.17	333	0.41	0.32	47.14	52.31
15	ORF1ab	nsp15	346	0.34	0.19	38.98	52.00	129	0.27	0.09	29.57	40.85	217	0.38	0.25	40.87	49.11
16	ORF1ab	nsp16	298	0.36	0.20	39.05	51.57	134	0.29	0.08	28.69	43.88	164	0.41	0.29	41.90	48.55
17	ORF2	S	1274	0.37	0.25	43.64	54.92	452	0.32	0.15	37.22	51.54	822	0.41	0.31	46.33	53.45
18	ORF3a	ORF3a	276	0.40	0.33	45.56	50.52	38	0.28	0.16	21.58	28.09	238	0.42	0.36	44.51	48.35
19	ORF4	E	76	0.39	0.34	33.57	37.18	29	0.27	0.30	20.07	22.79	47	0.45	0.36	24.70	27.10
20	ORF5	M	223	0.43	0.36	47.05	50.09	71	0.39	0.22	32.84	38.58	152	0.45	0.42	45.80	47.91
21	ORF6	ORF6	62	0.28	0.25	27.67	32.10	19	0.24	0.18	11.67	15.21	43	0.30	0.28	26.20	28.57
22	ORF7a	ORF7a	122	0.38	0.28	39.94	44.02	25	0.29	0.30	16.40	17.30	97	0.41	0.27	37.84	42.26
23	ORF7b	ORF7b	44	0.32	0.30	23.37	26.16	9	0.25	0.17	8.00	10.38	35	0.33	0.32	21.45	24.29
24	ORF8	ORF8	122	0.36	0.25	38.23	45.94	28	0.24	0.07	18.52	25.27	94	0.40	0.30	37.38	42.48
25	ORF9	N	420	0.47	0.36	46.89	49.08	88	0.33	0.17	29.64	30.14	332	0.51	0.41	47.24	47.70
26	ORF10	ORF10	39	0.34	0.33	24.99	26.22	22	0.30	0.21	17.60	19.69	17	0.39	0.47	13.00	14.22

CUB difference observed between conserved and variable genic regions might be an artefact without significance. To counter this argument, we did a simulation study. We grouped codons of a gene sequence randomly into two sets, seq1, and seq2, and found N_c values for the two sets of codons. The distribution of the two sets of N_c values are plotted in Figure 3. It is evident from Figure 3 that the distribution of the two sets of N_c values are not different. The distribution of N_c values for both sequences are on the higher side of the theoretical range of 20.0 to 61.0. The distributions are similar to that when the N_c values of gene sequences were calculated as a whole. This simulation study result supports our finding that the conserved codons are under stronger selection.

Figure 3 Distribution of N_c values in simulation study (see online version for colours)



Notes: The figure presents the distribution of N_c values in the simulation study. Y-axis represents the distribution of N_c values, and X-axis represents the two sets of codons. It is evident from the figure that there is no difference in the two distributions.

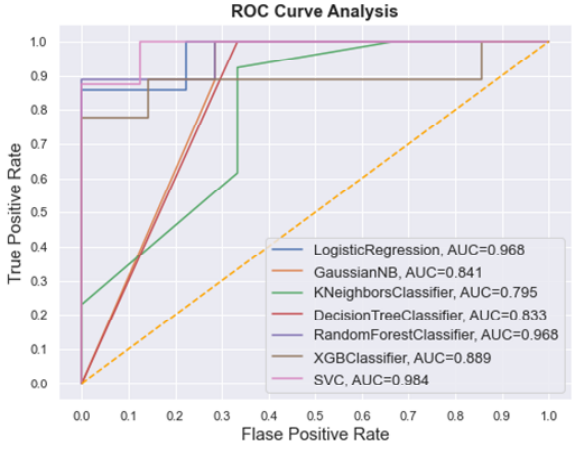
3.3 Codons with low G and C bases are preferred more in conserved than the variable sites in the SARS-CoV-2 genes for all the amino acids

To know if the identities of favoured codons in conserved region correspond to the nucleotide composition in the SARS-CoV-2 genome sequence, we did a comparative study on codon usage between conserved and variable sites. We calculated the RSCU values of the codons for these two sites considering all the genes together (Table 3).

RSCU values of the synonymous codons vary both in conserved and variable sites. However, the variation in RSCU values is higher in conserved sites compared to variable sites. For example, in the variable site, RSCU values of the Leu amino acid codons range between 1.82 and 0.44, whereas these values in the conserved site range between 2.63 and 0.04. Similar differences in RSCU values between the two sites were observed for all the amino acids. Because of these differences in RSCU values between the conserved and variable datasets, AUC values in the machine learning analysis were very high across all the seven classifiers (Figure 4). This highly biased codon usage in the conserved site was also reflected in the N_c value results presented in the earlier section.

We further did a comparative study between A/T and G/C rich codons of variable and conserved sites. For each amino acid, RSCU values for codons with A/U and G/C bases at 3rd codon positions are combined and given in Figure 5. RSCU values for the six-fold degenerate amino acid codons are shown in two groups, the four-fold degenerate family box, and the remaining two split box codons. It can be observed from Figure 5 that the U/A ending codons are used more compared to C/G ending codons both in conserved and variable sites. This observation suggests that the A/U ending codons are preferred compared to G/C ending codons in this virus genome. However, the difference between A/U ending and G/C ending codons in conserved sites is significantly (p-value 0.0002) higher than in variable sites. This difference between conserved and variable sites was consistently observed for all the amino acids. Genome G+C% for the virus is low; accordingly, U/A-ending codons are used more than the G/C-ending codons. However, in the conserved sited, the higher difference between the two sets of codons suggests that the conserved sites are under selection, and the selection on codon usage is acting in the direction of A/U-ending codons.

Figure 4 Classifier performance in predicting variable and conserved sited in terms of ROC curve (see online version for colours)



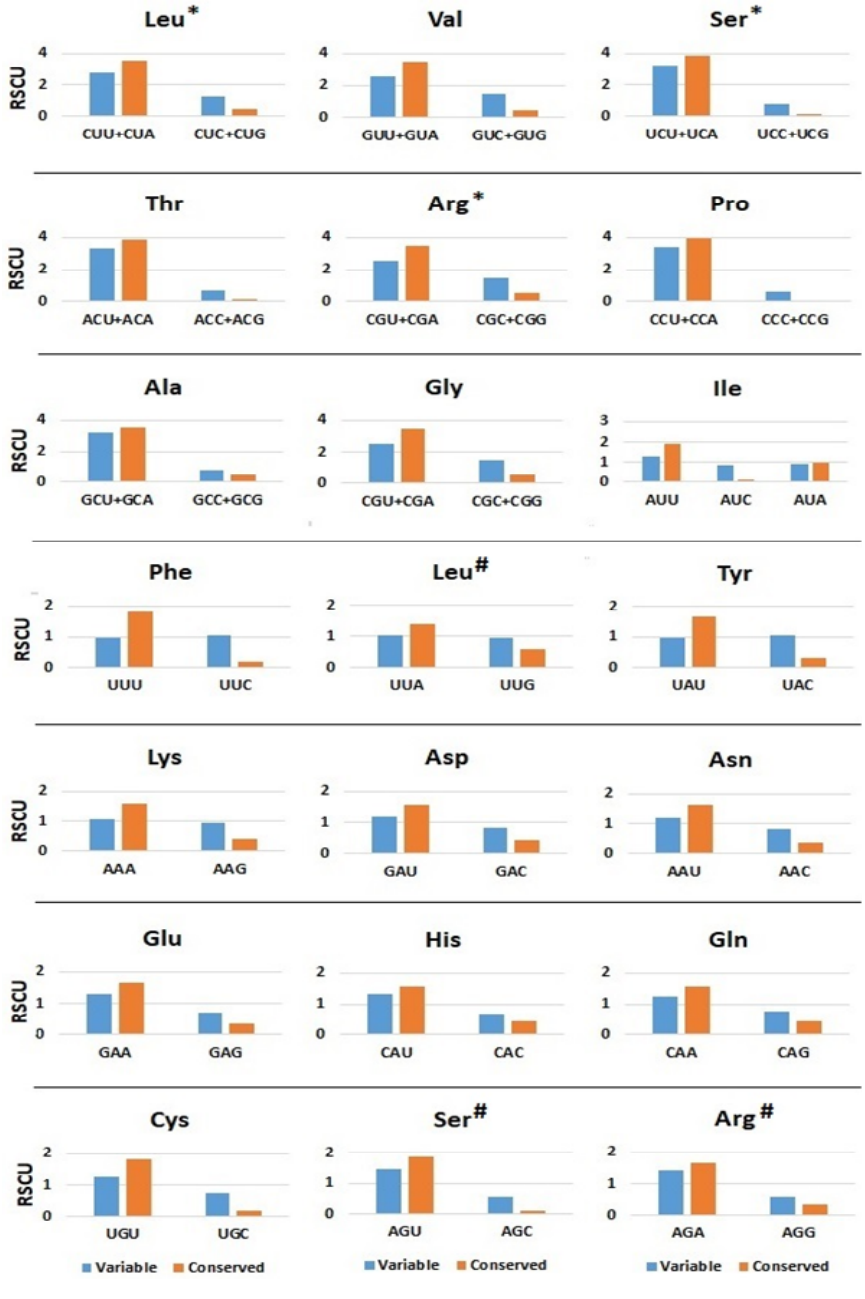
Notes: Figure presents ROC curves with AUC values to exhibit the classifier performance in the two sets of variable and conserved features. A consistently higher AUC values observed across classifiers can be observed.

Recent research reports suggested that the CUB in SARS-CoV-2 appears to be incompatible with the codon usage in the human genome (Chen and Yang, 2022) and gradually dissimilating from humans (Mogro et al., 2022). Though selection has been reported in some of the genes in the human genome (Plotkin et al., 2004; Dhindsa et al., 2020), selection on codon usage is generally low and resembles genome composition (Satapathy et al., 2015). In this context, our findings on the influence of selection on codon usage in the SARS-CoV-2 genome will be engaging in understanding the evolution of this deadly human pathogen in the coming future.

Table 3 Codon usage (RSCU) values in conserved and variable sites of SARS-CoV2 genome

AA	*Codon*	*All*	*Conserved*	*Variable*	*AA*	*Codon*	*All*	*Conserved*	*Variable*	*AA*	*Codon*	*All*	*Conserved*	*Variable*	*AA*	*Codon*	*All*	*Conserved*	*Variable*
Phe	UUU	1.53	1.80	0.96	Ser	UCU	2.24	2.21	1.83	Tyr	UAU	1.48	1.68	0.95	Cys	UGU	1.40	1.83	1.24
	UUC	0.47	0.20	1.04		UCC	0.73	0.08	0.65		UAC	0.52	0.32	1.05		UGC	0.60	0.17	0.76
Leu	UUA	1.56	2.63	1.13		UCA	1.58	1.80	1.61	TER	UAA	X	X	X	TER	UGA	X	X	X
	UUG	1.11	1.12	1.04		UCG	0.12	0.05	0.15		UAG	X	X	X	Trp	UGG	1.00	1.00	1.00
	CUU	2.00	1.60	1.82	Pro	CCU	2.00	2.25	1.85	His	CAU	1.53	1.57	1.32	Arg	CGU	1.29	1.30	1.55
	CUC	0.67	0.23	0.78		CCC	0.28	0.00	0.38		CAC	0.47	0.43	0.68		CGC	0.14	0.14	0.84
	CUA	0.50	0.38	0.79		CCA	1.72	1.71	1.56	Gln	CAA	1.48	1.59	1.22		CGA	0.00	0.28	0.30
	CUG	0.17	0.04	0.44		CCG	0.00	0.04	0.21		CAG	0.52	0.41	0.78		CGG	0.29	0.09	0.24
Ile	AUU	1.74	1.93	1.29	Thr	ACU	1.81	1.77	1.78	Asn	AAU	1.23	1.65	1.19	Ser	AGU	1.03	1.75	1.27
	AUC	0.55	0.10	0.82		ACC	0.41	0.13	0.44		AAC	0.77	0.35	0.81		AGC	0.30	0.11	0.49
	AUA	0.71	0.97	0.89		ACA	1.65	2.07	1.52	Lys	AAA	1.25	1.59	1.07	Arg	AGA	2.86	3.49	2.20
Met	AUG	1.00	1.00	1.00		ACG	0.12	0.03	0.25		AAG	0.75	0.41	0.93		AGG	1.43	0.70	0.87
Val	GUU	1.98	2.49	1.73	Ala	GCU	2.13	2.36	2.12	Asp	GAU	1.39	1.58	1.18	Gly	GGU	2.29	2.75	2.07
	GUC	0.87	0.21	0.71		GCC	0.41	0.39	0.62		GAC	0.61	0.42	0.82		GGC	0.73	0.34	0.95
	GUA	0.62	1.04	0.85		GCA	1.37	1.21	1.07	Glu	GAA	1.42	1.65	1.31		GGA	0.83	0.89	0.78
	GUG	0.54	0.26	0.71		GCG	0.10	0.03	0.19		GAG	0.58	0.35	0.69		GGG	0.15	0.02	0.19

Figure 5 RSCU values for codons with A/U and G/C bases at 3rd codon positions (see online version for colours)



Notes: Vertical bars in the figure represent RSCU values for codons with A/U and G/C bases at 3rd codon positions separately for variable and conserved codon sets for individual amino acids. RSCU values for the six-fold degenerate amino acid codons are shown in two groups, family box* and split box# codons.

References

Akashi, H. (2003) 'Translational selection and yeast proteome evolution', *Genetics*, Vol. 164, No. 4, pp.1291–1303, DOI: 10.1093/genetics/164.4.1291.

Albers, S. and Czech, A. (2016) 'Exploiting tRNAs to boost virulence', *Life*, Vol. 6, No. 1, Basel, Switzerland, DOI: 10.3390/life6010004.

Angeletti, S. et al. (2020) 'COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis', *Journal of Medical Virology*, Vol. 92(6), pp. 584–588. doi: 10.1002/jmv.25719.

Aziz, R. et al. (2022) 'Incorporation of transition to transversion ratio and nonsense mutations, improves the estimation of the number of synonymous and non-synonymous sites in codons', *DNA Research*, Vol. 29, No. 4, p.dsac023, DOI: 10.1093/dnares/dsac023.

Baek, S. and Sung, K.M. (2000) 'Fast K-nearest-neighbour search algorithm for nonparametric classification', *Electronics Letters*, Vol. 36, pp.1821–1822, DOI: 10.1049/el:20001249.

Bahir, I. et al. (2009) 'Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences', *Molecular Systems Biology*, Vol. 5, No. 1, p.311.

Belalov, I.S. and Lukashev, A.N. (2013) 'Causes and implications of codon usage bias in RNA viruses', *PloS One*, Vol. 8, No. 2, p.e56642.

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 1, pp.5–32, DOI: 10.1023/A:1010933404324.

Bulmer, M. (1991) 'The selection-mutation-drift theory of synonymous codon usage', *Genetics*, Vol. 129, No. 3, pp.897–907, DOI: 10.1093/genetics/129.3.897.

Burns, C.C. et al. (2006) 'Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region.', *Journal of Virology*, Vol. 80, No. 7, pp.3259–3272, DOI: 10.1128/JVI.80.7.3259-3272.2006.

Chen, F. and Yang, J-R. (2022) 'Distinct codon usage bias evolutionary patterns between weakly and strongly virulent respiratory viruses', *iScience*, Vol. 25, No. 1, p.103682, DOI: 10.1016/j.isci.2021.103682.

Chen, F. et al. (2020) 'Dissimilation of synonymous codon usage bias in virus–host coevolution due to translational selection', *Nature Ecology and Evolution*, Vol. 4, No. 4, pp.589–600.

Costafreda, M.I. et al. (2014) 'Hepatitis A virus adaptation to cellular shutoff is driven by dynamic adjustments of codon usage and results in the selection of populations with altered capsids', *Journal of Virology*, Vol. 88, No. 9, pp.5029–5041, DOI: 10.1128/JVI.00087-14.

Daron, J. and Bravo, I.G. (2021) 'Variability in codon usage in coronaviruses is mainly driven by mutational bias and selective constraints on CPG dinucleotide', *Viruses*, Vol. 13, No. 9, p.1800.

Denison, M.R. et al. (2011) 'Coronaviruses', *RNA Biology*, Vol. 8, No. 2, pp.270–279, DOI: 10.4161/rna.8.2.15013.

Dhindsa, R.S. et al. (2020) 'Natural selection shapes codon usage in the human genome', *American Journal of Human Genetics*, Vol. 107, No. 1, pp.83–95, DOI: 10.1016/j.ajhg.2020.05.011.

dos Reis, M., Wernisch, L. and Savva, R. (2003) 'Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome', *Nucleic Acids Research*, Vol. 31, No. 23, pp.6976–6985, DOI: 10.1093/nar/gkg897.

Grantham, R., Gautier, C. and Gouy, M. (1980) 'Codon frequencies in 119 individual genes confirm corsistent choices of degenerate bases according to genome type', *Nucleic Acids Research*, Vol. 8, No. 9, pp.1893–1912, DOI: 10.1093/nar/8.9.1893.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) 'Overview of supervised learning BT – the elements of statistical learning: data mining, inference, and prediction', in Hastie, T., Tibshirani, R. and Friedman, J. (Eds.): pp.9–41, Springer New York, New York, NY, DOI: 10.1007/978-0-387-84858-7_2.

Hershberg, R. and Petrov, D.A. (2009) 'General rules for optimal codon choice', *PLoS Genetics*, Vol. 5, No. 7, p.e1000556.

Holland, J.J. (2006) 'Transitions in understanding of RNA viruses: a historical perspective BT – quasispecies: concept and implications for virology', in Domingo, E. (Ed.): pp.371–401, Springer Berlin Heidelberg, Berlin, Heidelberg, DOI: 10.1007/3-540-26397-7_14.

Holmes, K.V (2003) 'SARS-associated coronavirus', *The New England Journal of Medicine*, Vol. 348, No. 20, pp.1948–1951, DOI: 10.1056/NEJMp030078.

Hussain, S., Shinu, P., Islam, M.M., Chohan, M.S. and Rasool, S.T. (2020) 'Analysis of codon usage and nucleotide bias in middle east respiratory syndrome coronavirus genes', *Evol. Bioinform. Online*, 4 May, Vol. 16, DOI: 10.1177/1176934320918861, PMID: 32425493; PMCID: PMC7218340.

Ikemura, T. (1981) 'Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes', *Journal of Molecular Biology*, Vol. 146, No. 1, pp.1–21, DOI: https://doi.org/10.1016/0022-2836(81)90363-6.

Ikemura, T. (1985) 'Codon usage and tRNA content in unicellular and multicellular organisms', *Molecular Biology and Evolution*, Vol. 2, No. 1, pp.13–34, DOI: 10.1093/oxfordjournals.molbev.a040335.

Jenkins, G.M. and Holmes, E.C. (2003) 'The extent of codon usage bias in human RNA viruses and its evolutionary origin', *Virus Research*, Vol. 92, No. 1, pp.1–7, DOI: 10.1016/s0168-1702(02)00309-x.

Jia, W. and Higgs, P.G. (2008) 'Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection', *Molecular Biology and Evolution*, Vol. 25, No. 2, pp.339–351, DOI: 10.1093/molbev/msm259.

Jitobaom, K. et al. (2020) 'Codon usage similarity between viral and some host genes suggests a codon-specific translational regulation', *Heliyon*, Vol. 6, No. 5, p.e03915.

Klitting, R., Gould, E.A. and de Lamballerie, X. (2016) 'G+C content differs in conserved and variable amino acid residues of flaviviruses and other evolutionary groups', *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, Vol. 45, pp.332–340, DOI: 10.1016/j.meegid.2016.09.017.

Knight, R.D., Freeland, S.J. and Landweber, L.F. (2001) 'A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes', *Genome Biology*, Vol. 2, No. 4, p.RESEARCH0010, DOI: 10.1186/gb-2001-2-4-research0010.

Kurmi, A., Sen, P., Dash, M., Patra, A. K., Ray, S. and Satapathy, S. S. (2023) 'Prediction of essential genes using single nucleotide compositional features in genomes of bacteria: a machine learning-based analysis', International Journal of Bioinformatics Research and Applications, 19(1), p. 1. doi: 10.1504/IJBRA.2023.10054584.

Lewis Jr, C. A. et al. (2016) 'Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history', Proceedings of the National Academy of Sciences, 113(29), pp. 8194–8199.

Lucks, J.B. et al. (2008) 'Genome landscapes and bacteriophage codon usage', *PLoS Computational Biology*, Vol. 4, No. 2, p.e1000001.

Mogro, E.G., Bottero, D. and Lozano, M.J. (2022) 'Analysis of SARS-CoV-2 synonymous codon usage evolution throughout the COVID-19 pandemic', *Virology*, Vol. 568, pp.56–71, DOI: 10.1016/j.virol.2022.01.011.

Mousavizadeh, L. and Ghasemi, S. (2021) 'Genotype and phenotype of COVID-19: Their roles in pathogenesis', *Journal of Microbiology, Immunology, and Infection = Wei Mian Yu Gan Ran Za Zhi*, Vol. 54, No. 2, pp.159–163, DOI: 10.1016/j.jmii.2020.03.022.

Mueller, S. et al. (2006) 'Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity', *Journal of Virology*, Vol. 80, No. 19, pp.9687–9696, DOI: 10.1128/JVI.00738-06.

Noble, W.S. (2006) 'What is a support vector machine?', *Nature Biotechnology*, Vol. 24, No. 12, pp.1565–1567, DOI: 10.1038/nbt1206-1565.

Novembre, J.A. (2002) 'Accounting for background nucleotide composition when measuring codon usage bias', *Molecular Biology and Evolution*, Vol. 19, No. 8, pp.1390–1394, DOI: 10.1093/oxfordjournals.molbev.a004201.

Pérez, A., Larranaga, P. and Inza, I. (2006) 'Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes', *International Journal of Approximate Reasoning*, Vol. 43, pp.1–25, DOI: 10.1016/j.ijar.2006.01.002.

Plotkin, J.B. and Kudla, G. (2011) 'Synonymous but not the same: the causes and consequences of codon bias', *Nature Reviews. Genetics*, Vol. 12, No. 1, pp.32–42, DOI: 10.1038/nrg2899.

Plotkin, J.B., Robins, H. and Levine, A.J. (2004) 'Tissue-specific codon usage and the expression of human genes', *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 101, No. 34, pp.12588–12591, DOI: 10.1073/pnas.0404957101.

Quinlan, J.R. (1986) 'Induction of decision trees', *Machine Learning*, Vol. 1, No. 1, pp.81–106, DOI: 10.1007/BF00116251.

Ramazzotti, D. et al. (2022) 'Large-scale analysis of SARS-CoV-2 synonymous mutations reveals the adaptation to the human codon usage during the virus evolution', *Virus Evolution*, Vol. 8, No. 1, p.veac026.

Roth, A., Anisimova, M. and Cannarozzi, G.M. (2012) 'Measuring codon usage bias', in Cannarozzi, G.M. and Schneider, A. (Eds.): *Codon Evolution: Mechanisms and Models*, Oxford University Press, DOI: 10.1093/acprof:osobl/9780199601165.003.0013.

Satapathy, S.S. et al. (2015) 'Codon usage bias is not significantly different between the high and the low expression genes in human', *Int. J. Mol. Genet. Gene. Ther.,* Vol. 1, No. 1, http://dx.doi.org/10.16966/2471- 4968.103.

Satapathy, S.S. et al. (2017) 'Codon degeneracy and amino acid abundance influence the measures of codon usage bias: improved Nc ($\hat{N}$(c)) and ENCprime ($\hat{N}$(')(c)) measures', *Genes to Cells : Devoted to Molecular and Cellular Mechanisms*, Vol. 22, No. 3, pp.277–283, DOI: 10.1111/
gtc.12474.

Sharp, P.M. and Li, W.H. (1986) 'An evolutionary perspective on synonymous codon usage in unicellular organisms', *Journal of Molecular Evolution*, Vol. 24, Nos. 1–2, pp.28–38, DOI: 10.1007/BF02099948.

Sharp, P.M. and Li, W.H. (1987) 'The codon adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications', *Nucleic Acids Research*, Vol. 15, No. 3, pp.1281–1295, DOI: 10.1093/nar/15.3.1281.

Sharp, P.M. et al. (1988) 'Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity', *Nucleic Acids Research*, Vol. 16, No. 17, pp.8207–8211, DOI: 10.1093/nar/16.17.8207.

Sharp, P.M. et al. (2005) 'Variation in the strength of selected codon usage bias among bacteria', *Nucleic Acids Research*, Vol. 33, No. 4, pp.1141–1153, DOI: 10.1093/nar/gki242.

Sharp, P.M., Emery, L.R. and Zeng, K. (2010) 'Forces that influence the evolution of codon bias', *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, Vol. 365, No. 1544, pp.1203–1212, DOI: 10.1098/rstb.2009.0305.

Sheridan, R.P. et al. (2016) 'Extreme gradient boosting as a method for quantitative structure-activity relationships', *Journal of Chemical Information and Modeling*, Vol. 56, No. 12, pp.2353–2360, DOI: 10.1021/acs.jcim.6b00591.

Simmonds, P. and Ansari, M.A. (2021) 'Extensive C->U transition biases in the genomes of a wide range of mammalian RNA viruses; potential associations with transcriptional mutations, damage- or host-mediated editing of viral RNA', *PLoS Pathogens*, Vol. 17, No. 6, p.e1009596, DOI: 10.1371/journal.ppat.1009596.

Tian, L. et al. (2018) 'The adaptation of codon usage of +ssRNA viruses to their hosts', *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, Vol. 63, pp.175–179, DOI: 10.1016/j.meegid.2018.05.034.

Wong, E. H. M. et al. (2010) 'Codon usage bias and the evolution of influenza a viruses. Codon usage biases of influenza virus', *BMC Evolutionary Biology*, Vol. 10, No. 1, p.253, DOI: 10.1186/1471-2148-10-253.

Wright, F. (1990) 'The 'effective number of codons' used in a gene', *Gene*, Vol. 87, No. 1, pp.23–29, DOI: 10.1016/0378-1119(90)90491-9.

Yang, P. and Wang, X. (2020) 'COVID-19: a new challenge for human beings', *Cellular and Molecular Immunology*, Vol. 17, No. 5, pp.555–557, DOI: 10.1038/s41423-020-0407-x.

Yao, X., Fan, Q., Yao, B., Lu, P., Rahman, S.U., Chen, D. and Tao, S. (2020) 'Codon usage bias analysis of bluetongue virus causing livestock infection', *Front Microbiol*., 19 May, Vol. 11, p.655, DOI: 10.3389/fmicb.2020.00655, PMID: 32508755; PMCID: PMC7248248.

Zheng, J. (2020) 'SARS-CoV-2: an emerging coronavirus that causes a global threat', *International Journal of Biological Sciences,* Vol. 16, No. 10, pp.1678–1685, DOI: 10.7150/ijbs.45053.