Biomacromol. J., Vol. 7, No. 3, 114-125, December 2021.

Biomacromolecular Journal
www.bmmj.org

Molecular Modeling, Codon Usage, Rare Codon and Phylogenetic Relations Analysis of Spike Glycoprotein in SARS CoV-2 Virus

M. Mortazavi[a], M. Forouzesh[b], A. Malekpour[b,*], A. Keshavarzi[c], R. Mohammadi[d], F. Kargar[e] and R. Deghani[f]

[a]*Department of Biotechnology, Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran*

[b]*Legal Medicine Research Center, Legal Medicine Organization of Iran, Tehran, Iran*

[c]*Burn and Wound healing Research Center, Shiraz University of Medical Sciences, Shiraz, Iran*

[d]*Genetic Laboratory, Shiraz Fertility Center, Shiraz, Iran*

[e]*Department of Medical Biotechnology, School of Advanced Medical Sciences and Technologies Tabriz University of Medical Sciences Tabriz Iran*

[f]*Department of Pharmacology, Bam University of Medical Sciences, Bam, Iran*

(Received 20 May 2022, Accepted 15 July 2022)

ABSTRACT

The 2019 novel coronavirus (2019-nCoV/SARS-CoV-2) initially appeared as part of an important prevalence of respiratory disease centered in Hubei province, China. Now, it is a pandemic globally and is a significant public health concern. Taxonomically, SARS-CoV-2 was revealed to be a Beta coronavirus (lineage B) related to SARS-CoV and SARS-related bat coronaviruses closely, and it has been stated to have a similar receptor with SARS-CoV (ACE-2). Here, we carried out the codon usage bias (CUB) by analyzing the codon bias index (CBI), codon adaptation index (CAI), and an effective number of codons (ENC) besides phylogenetic analysis of Coronaviridae and also structural modeling of the SARS-CoV-2 spike glycoprotein. We observed that 2019-nCoV has a rich composition of AT nucleotides that strongly affects its codon usage, which seems to be not optimized for the human hosts. Moreover, a close evolutionary phylogenetic relationship was detected between SARS-CoV-2/human/IRIN/ and SARS-CoV-2/human/CHN/WH-09/2020. By in silico modeling of spike glycoprotein, an I-TASSER server, the 3Dstructure of it was also evaluated. This type of analysis would be beneficial for exploring a virus adaptation to host, evolution and is therefore of value to developing vaccine design and pharmaceutical agents.

Keywords: Computational biology, SARS-CoV-2, Beta coronavirus, Codon usage, Modeling

INTRODUCTION

Coronaviruses (CoVs) are recognized as enveloped, positive-stranded RNA viruses with a ~30 kb genome length and also four structural proteins, containing spike (S), envelope (E), membrane (M), and nucleocapsid (N) [1]. The S protein regulates attachment of the virus to the target host cell receptor [2]; the E protein acts to gather the virions and functions as an ion channel [3]; the M protein, besides the E protein, has a role in the assembly of virus and is associated with new virus particles biosynthesis [4]; and the N protein constructs the ribonucleoprotein complex along with the virus RNA [5]. Severe acute respiratory syndrome CoV 2 (2019-nCoV), which was first reported in Wuhan (China) in December of 2019, results in a severe acute respiratory disease with a 3% to 6% mortality rate [6].

The recently sequenced virus genome encodes two open reading frames (ORFs), ORF1a and ORF1ab, the second encodes an RNA-dependent RNA polymerase (RdRP) or viral replicase, besides four structural proteins [7]. CoV uses spike glycoprotein (S), a major target of antibody neutralization, for receptor binding, and mediating membrane fusion and virus entry [8] providing a plan for designing vaccines and viral entry inhibitors. The codon

*Corresponding author. E-mail: immurasoul@gmail.com

usage bias (CUB) phenomenon exists in numerous genomes containing RNA genomes and indeed, it is specified by mutation and selection [9]. Recent studies have shown that synonymous codons are not used with the same frequency in organisms [10]. Therefore, it is important to evaluate patterns of common codon usage in coronaviruses since CUB can be related to the driving forces, which construct the small RNA viruses' evolutions. Mutational bias has been specified as the major indicator of codon usage variation between RNA viruses [11]. Actually, RNA viruses indicate an effective number of codons (ENC) that is completely high (ENC > 45), pointing to completely random codon usage, while the adaptive index CAI shows that the viral CUB is in line with that of the host, as detected in the Equine infectious anemia virus (EIAV) or Zaire Ebola virus(ZEBOV) [12].

Characterization of gene Coronaviridae family and their possible differences are likely to facilitate and contribute to the development of effectively preventing and treatment protocols against coronavirus infection. The purpose of this bioinformatics study was to study the gene features of spike glycoprotein in the Coronaviridae family. In the present work, the purpose was to study the CUB by analyzing the codon bias index (CBI), codon adaptation index (CAI), and an effective number of codons (ENC). For further understanding of the rare codons role, the 3D structure of this enzyme was modeled in the I-TASSER [13] and the situation of these rare codons was visualized and studied using Swiss PDB Viewer software [14] and PyMOL Molecular Graphics System [15]. More precise approaches can be chosen for treatment regimens using the findings of this study.

MATERIALS AND METHODS

For bioinformatics analysis, the nucleotide sequences and their features of the Coronaviridae family were obtained from NCBI (http://www.ncbi.nlm.nih.gov/) (Table 1). In the next phase, the frequency, number, and fraction of 61 codons for each amino acid were evaluated and the preferred codons were extracted using the Gene Infinity website (https://www.bioinformatics.org/sms2/codon_usage.html) [16].

Sequence Information and Annotation

The variation in codon usage bias was quantified by the ENC and CBI in the ACUA software [17]. To evaluate the codon usage bias the ENC (value ranges from 20-61) is generally used [18]. The CBI also calculated the number of preferred codons that are used [19]. Thus, the CBI value of zero means random choices are used, the 1 means only preferred codons, and less than zero implies greater use of non-preferred codons [19]. In highly expressed genes, the CAI evaluates the degree of bias towards the codons [20]. In the gene, in which uniformly all synonymous codons were used, the CAI would be 0 indicating no bias. On the other hand, in the gene that the optimal codons were used, the CAI would be 1 for the strongest codon bias [21]. AT and GC contents at three codon positions *i.e.* AT1, AT2, AT3, GC1, GC2, and GC3 was calculated. The GC3 content is assumed to be an excellent index of base composition bias [22].

Molecular Modeling of Spike Glycoprotein

To investigate the position of these codon usages in spike glycoprotein, the 3D structure of it was modeled by submission of spike glycoprotein sequence in I-TASSER web server [13]. I-TASSER web server was used to generate a total of five most suitable models based on multiple-threading alignment by LOMETS [23]. The model which showed the best confidence and Z-score was selected and visualized using swiss PDB viewer [24] and PyMOL molecular graphics system [15]. Hydrogen bonds were also calculated by WHAT IF web server [25] and PIC web server [26].

Evolutionary Relationship

In the following step, the evolutionary relationship and phylogenetic analysis of Coronaviridae were studied using the MEGA 7 software [29]. This analysis was performed by the construction of a phylogenetic tree with the maximum parsimony tree in MEGA 7. The frequency of used codons was reported as descriptive statistics.

RESULTS

Codon Usage and Nucleotide Composition Analysis

The codon bias and nucleotide composition relationship

Table 1. Genetic Properties of Coronaviridae

Source	Locus	DEFINITION	Version	Protein ID
SARS-CoV-2/human/IRN/	MT320891	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/IRN/HGRC-1.1-IPI-8206/2020, complete genome	MT320891.2	QIX12195.1
SARS-CoV-2/human/CHN/WH-09/2020	MT093631	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/WH-09/2020, complete genome.	MT093631.2	QIC53213.1
SARS-CoV-2/human/COL/79256_Antioquia/2020	MT256924	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/COL/79256_Antioquia/2020, complete genome	MT256924.2	QIS30054.2
SARS-CoV-2/human/CHN/Yunnan-01/2020	MT049951	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/Yunnan-01/2020, complete genome	MT049951.1	QIA20044.1
Coronavirus 2 isolate Wuhan-Hu-1	NC_045512	Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.	NC_045512.2	YP_009724390.1
SARS coronavirus TJF	AY654624	spike glycoprotein [SARS coronavirus TJF]	AY654624.1	AAT76147.1
BtRs-BetaCoV/YN2013	KJ473816	BtRs-BetaCoV/YN2013, complete genome	KJ473816.1	AIA62330.1
BtRs-BetaCoV/GX2013	KJ473815	BtRs-BetaCoV/GX2013, complete genome	KJ473815.1	AIA62320.1
BtRs-BetaCoV/HuB2013	KJ473814	BtRs-BetaCoV/HuB2013, complete genome	KJ473814.1	AIA62310.1
BtRf-BetaCoV/SX2013	KJ473813	BtRf-BetaCoV/SX2013, complete genome	KJ473813.1	AIA62300.1
BtRf-BetaCoV/HeB2013	KJ473812	BtRf-BetaCoV/HeB2013, complete genome	KJ473812.1	AIA62290.1
BtRf-BetaCoV/JL2012	KJ473811	BtRf-BetaCoV/JL2012, complete genome	KJ473811.1	AIA62277.1
Bat coronavirus	DQ648790	Bat coronavirus (BtCoV/A434/2005) spike (S) gene, complete cds	DQ648790.1	ABG11962.1
Bat coronavirus (BtCoV/A701/2005) spike (S) gene	DQ648793	Bat coronavirus (BtCoV/A701/2005) spike (S) gene, complete cds	DQ648793.1	ABG11965.1
Bat coronavirus (BtCoV/A515/2005) spike (S) gene	DQ648791	Bat coronavirus (BtCoV/A515/2005) spike (S) gene, complete cds	DQ648791.1	ABG11963.1
Bat coronavirus (BtCoV/A527/2005) spike (S) gene	DQ648792	Bat coronavirus (BtCoV/A527/2005) spike (S) gene, complete cds	DQ648792.1	DQ648792.1
Canine coronavirus strain TN-449	JQ404410	Canine coronavirus strain TN-449, complete genome	JQ404410.1	AFG19738.1

were evaluated by comparing the values of A, T, G, C, and GC with the A3, T3, G3, C3, and GC3 values, respectively. The AT1, AT2, At3, GC1, GC2, and GC3 values were calculated for each gene. The GC3% of the cds was in the range between 15.668 to 16.534 and GC3 Skewness from 0.299 to 0.34 (Table 2). The GC content at the first codon position (GC1) and second codon position (GC2) was compared with that of the third codon position (GC3) and showed that the patterns of base compositions are most likely the result of mutation pressure rather than that of natural selection since at all codon positions its effects are present [27].

Prevalence of Preferred (Used) Codons

The number and frequency of each codon type of spike glycoprotein were evaluated in the Sequence Manipulation Suite [16]. Table 3 shows the frequency, number, and fraction of 61 codons for each amino acid in the orf1a polyprotein (pp1a). The results of the codon usage analysis showed that some codon usages had generally different

Table 2. Compositional Analysis of Coronaviridae Genome Sequence

Gene	A	T	G	C	Total bp	AT%	GC%	AT Skewness	GC Skewness	A1	T1	G1	C1	AT1%	GC1%	AT1 Skewness	GC1 Skewness
> MT320891.2	1125	1271	703	723	3822	62.69	37.31	-0.061	-0.014	393	372	198	311	20.016	13.31	0.027	-0.222
>MT093631.2	1125	1271	703	723	3822	62.69	37.31	-0.061	-0.014	393	372	198	311	20.016	13.31	0.027	-0.222
> MT256924.2	1125	1271	703	723	3822	62.69	37.31	-0.061	-0.014	393	372	198	311	20.016	13.31	0.027	-0.222
>MT049951.1	1126	1270	703	723	3822	62.69	37.31	-0.06	-0.014	393	372	198	311	20.01	13.31	0.027	-0.222
>NC_045512.2	1120	1250	697	717	3784	62.63	37.36	-0.055	-0.014	386	304	365	206	18.23	15.09	0.119	0.278
>AY654624.1	1055	1253	703	757	3768	61.25	38.74	-0.086	-0.037	382	370	184	320	19.95	13.37	0.016	-0.27
>KJ473816.1	1009	1216	735	742	3702	60.10	39.89	-0.093	-0.005	358	374	189	313	19.77	13.56	-0.022	-0.247
>KJ473815.1	1069	1187	702	771	3729	60.49	39.50	-0.052	-0.047	382	359	189	313	19.87	13.46	0.031	-0.247
>KJ473814.1	1064	1159	708	795	3726	59.66	40.33	-0.043	-0.058	378	368	186	310	20.02	13.31	0.013	-0.25
>KJ473813.1	1050	1196	698	776	3720	60.37	39.62	-0.065	-0.053	369	369	187	315	19.83	13.49	0	-0.255
> KJ473812.1	1053	1199	699	775	3726	60.44	39.56	-0.065	-0.052	370	372	187	313	19.91	13.41	-0.003	-0.252
>KJ473811.1	1036	1185	719	771	3711	59.84	40.15	-0.067	-0.035	367	365	189	316	19.72	13.60	0.003	-0.251
>DQ648790.1	1080	1256	758	865	3959	59.00	40.99	-0.075	-0.066	404	362	210	344	19.34	13.99	0.055	-0.242
>DQ648793.1	1031	1463	830	757	4081	61.11	38.88	-0.173	0.046	399	425	225	312	20.19	13.15	-0.032	-0.162
>DQ648791.1	1076	1314	890	829	4109	58.16	41.83	-0.1	0.035	398	403	229	342	19.49	13.89	-0.006	-0.198
>DQ648792.1	1061	1300	888	816	4065	58.08	41.91	-0.101	0.042	393	408	229	335	19.70	13.87	-0.019	-0.188
> JQ404410.1	1319	1429	872	745	4365	62.95	37.04	-0.04	0.079	445	430	260	320	20.04	13.28	0.017	-0.103

Gene	A2	T2	G2	C2	AT2 Percent	GC2 Percent	AT2 Skewness	GC2 Skewness	A3	T3	G3	C3	AT3 Percent	GC3 Percent	AT3 Skewness	GC3 Skewness
> MT320891.2	344	590	137	203	24.437	8.896	-0.263	-0.194	387	309	368	209	18.21	15.097	0.112	0.276
>MT093631.2	344	590	137	203	24.437	8.896	-0.263	-0.194	387	309	368	209	18.21	15.097	0.112	0.276
> MT256924.2	344	590	137	203	24.437	8.896	-0.263	-0.194	387	309	368	209	18.21	15.097	0.112	0.276
>MT049951.1	344	590	137	203	24.437	8.896	-0.263	-0.194	388	308	368	209	18.21	15.097	0.115	0.276
>NC_045512.2	393	362	197	309	19.952	13.372	0.041	-0.221	341	583	135	202	24.419	8.906	-0.262	-0.199
>AY654624.1	294	584	150	228	23.301	10.032	-0.33	-0.206	378	299	369	209	17.967	15.34	0.117	0.277
>KJ473816.1	292	541	175	226	22.501	10.832	-0.299	-0.127	358	301	371	203	17.801	15.505	0.086	0.293
>KJ473815.1	307	541	147	248	22.741	10.593	-0.276	-0.256	379	287	366	210	17.86	15.447	0.138	0.271
>KJ473814.1	310	505	157	270	21.873	11.46	-0.239	-0.265	375	286	365	215	17.74	15.566	0.135	0.259
>KJ473813.1	314	521	151	254	22.446	10.887	-0.248	-0.254	366	306	360	207	18.065	15.242	0.089	0.27
> KJ473812.1	315	522	151	254	22.464	10.87	-0.247	-0.254	367	305	361	208	18.035	15.271	0.092	0.269
>KJ473811.1	301	528	167	241	22.339	10.994	-0.274	-0.181	367	292	363	214	17.758	15.548	0.114	0.258
>DQ648790.1	284	573	190	272	21.647	11.67	-0.337	-0.177	391	321	358	249	17.984	15.332	0.098	0.18
>DQ648793.1	237	715	193	215	23.328	9.998	-0.502	-0.054	394	323	412	230	17.569	15.731	0.099	0.283
>DQ648791.1	261	591	253	264	20.735	12.582	-0.387	-0.021	416	320	408	223	17.912	15.357	0.13	0.293
>DQ648792.1	252	577	251	259	20.394	12.546	-0.392	-0.016	415	315	408	222	17.958	15.498	0.137	0.295
> JQ404410.1	394	659	177	225	24.124	9.21	-0.252	-0.119	479	340	435	200	18.763	14.548	0.17	0.37

Table 3. The Nucleotide Compositional Properties of the Coronaviridae

		MERS		Covid		Sars	
Amino acids	Codon	Number	Fraction	Number	Fraction	Number	Fraction
Ala	GCG	30	0.09	13	0.04	14	0.04
	GCA	89	0.26	83	0.27	95	0.29
	GCT	174	0.51	166	0.54	171	0.53
	GCC	50	0.15	47	0.15	45	0.14
Cys	TGT	73	0.54	113	0.82	94	0.65
	TGC	61	0.46	25	0.18	50	0.35
Asp	GAT	145	0.61	133	0.63	137	0.62
	GAC	94	0.39	87	0.37	74	0.38
Glu	GAG	82	0.46	67	0.26	116	0.48
	GAA	95	0.54	176	0.74	124	0.52
Phe	TTT	136	0.62	156	0.75	118	0.61
	TTC	83	0.38	52	0.25	77	0.39
Gly	GGG	21	0.08	6	0.02	12	0.04
	GGA	37	0.15	44	0.17	49	0.18
	GGT	138	0.54	172	0.65	154	0.57
	GGC	59	0.23	42	0.16	54	0.20
His	CAT	55	0.66	53	0.71	55	0.64
	CAC	28	0.34	22	0.29	31	0.36
Ile	ATA	44	0.22	68	0.32	36	0.17
	ATT	112	0.57	109	0.51	126	0.58
	ATC	41	0.21	38	0.18	53	0.25
Lys	AAG	121	0.51	102	0.37	130	0.50
	AAA	118	0.49	174	0.63	128	0.50
Leu	TTG	117	0.27	87	0.20	84	0.19
	TTA	88	0.21	117	0.27	74	0.17
	CTG	32	0.08	18	0.04	48	0.11
	CTA	27	0.06	45	0.10	40	0.09
	CTT	118	0.28	129	0.30	135	0.30
	CTC	44	0.10	39	0.09	63	0.14
Met	ATG	95	1.00	105	1.00	111	1
Asn	AAT	142	0.68	161	0.69	130	0.61
	AAC	67	0.32	72	0.31	84	0.39
Pro	CCG	9	0.05	4	0.02	5	0.03
	CCA	51	0.30	63	0.38	69	0.42
	CCT	81	0.48	86	0.52	77	0.46
	CCC	29	0.17	11	0.07	15	0.09
Gln	CAG	57	0.39	54	0.36	69	0.47
	CAA	88	0.61	97	0.64	78	0.53

Table 3. Continued

Arg	AGG	22	0.15	20	0.15	19	0.13
	AGA	32	0.22	55	0.42	50	0.34
	CGG	12	0.08	3	0.02	1	0.01
	CGA	11	0.08	7	0.05	8	0.05
	CGT	40	0.27	34	0.26	53	0.36
	CGC	29	0.20	12	0.09	15	0.10
Ser	AGT	78	0.24	76	0.26	63	0.21
	AGC	22	0.07	12	0.04	24	0.08
	TCG	10	0.03	5	0.02	10	0.03
	TCA	68	0.21	82	0.28	76	0.26
	TCT	114	0.35	98	0.33	102	0.34
	TCC	32	0.10	21	0.07	23	0.08
Thr	ACG	11	0.04	12	0.03	11	0.03
	ACA	99	032	142	0.41	127	0.40
	ACT	144	0.46	158	0.46	127	0.40
	ACC	60	0.19	33	0.10	55	0.17
Val	GTG	84	0.20	57	0.15	69	0.19
	GTA	74	0.17	78	0.21	71	0.20
	GTT	185	0.43	193	0.52	162	0.45
	GTC	87	0.20	43	0.12	55	0.15
Trp	TGG	50	1.00	46	1.00	45	1.00
Tyr	TAT	132	0.67	112	0.57	103	0.56
	TAC	64	0.33	83	0.43	81	0.44
Terminal codon	TGA	0.00	0.00	0.00	0.00	0.00	0.00
	TAG	0.00	0.00	0.00	0.00	0.00	0.00
	TAA	1.00	1.00	1.00	1.00	1.00	1.00

frequencies in the MERS and SARS in relation to 2019-nCoV/SARS-CoV-2 (Table 3). This result is very important, as these residues may have a critical role in determining the final structure of the orf1a polyprotein. However, it is essential to confirm this conclusion with more experimental evidence.

Molecular Modeling of Spike Glycoprotein

For the detailed study of these codons, a 3D model of this enzyme is needed. In this context, the 3D structure of spike glycoprotein was modeled by the I-TASSEAR web server that created five models and the best model was selected based on C-score, shown in Fig. 1. This model showing a 1.52 value of overall C-score, Exp. RMSD equal to 13.3 ± 4.1 Å and 0.53 ± 0.15 value of TM-Score. The physicochemical properties of spike glycoprotein that were calculated by the ProtParam tool were shown in Table 4. The first value is based on the assumption that both cysteine residues form cystine and the second assumes that both cysteine residues are reduced.

Evolutionary Relationship

In the following step, the evolutionary relationship nucleotide sequences of spike glycoprotein and phylogenetic analysis of Coronaviridae spike glycoprotein were studied using the MEGA 7 software (Fig. 2) [29]. This analysis was performed by the construction of a phylogenetic tree with the maximum parsimony tree in

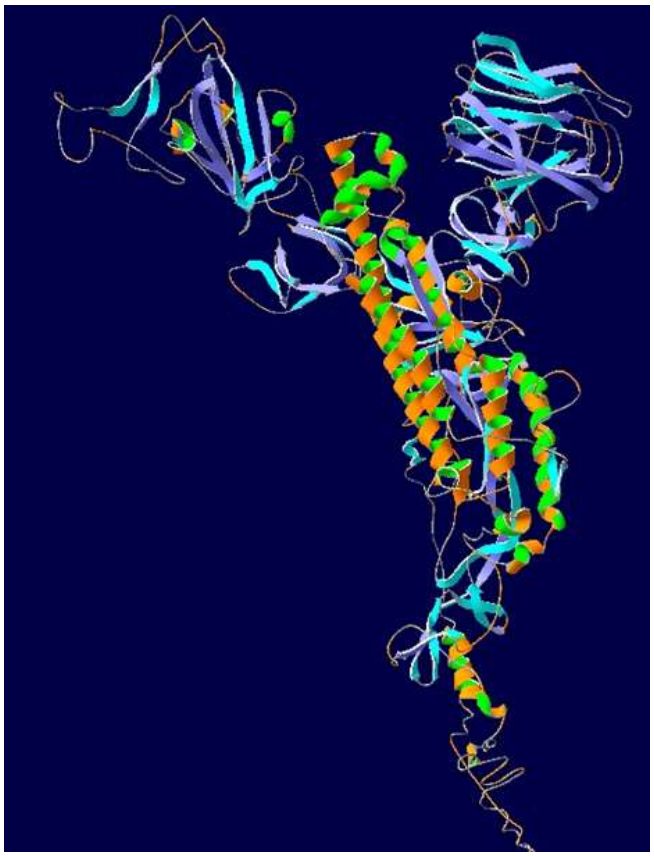

Fig. 1. The ribbon diagram of spike glycoprotein of SARSCoV-2.

MEGA 7. The frequency of used codons was reported as descriptive statistics (Table 5). The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree with the sum of branch length = 3.93120359 is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree.

Table 4. The Physicochemical Properties of Spike Glycoprotein Obtained from ProtParam Tool*

Parameters	Spike glycoprotein
Theoretical pI	
Molecular weight	6.24
Sequence length	141178.47
Extinction coefficients	1273
(M^{-1} cm^{-1} at 260 nm)*	146460-148960
Asp + Glu	110
Arg + Lys	103
Instability index	33.01
Grand average of	-0.079
hydropathicity	84.67
Aliphatic index	

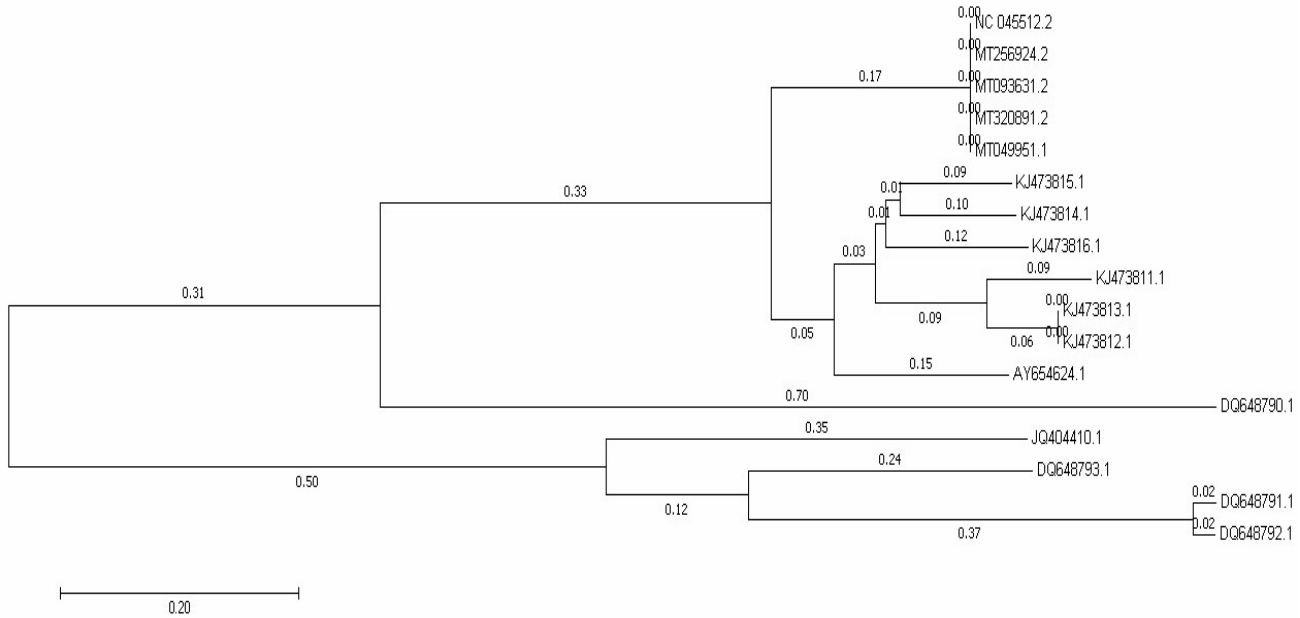

Fig. 2. Phylogenetic tree of the Coronaviridae spike glycoprotein nucleotide sequences.

Table 5. Molecular Evolution and Phylogenetic Diagram Coronaviridae

MT320891.2																	
MT093631.2	0.000																
MT256924.2	0.000	0.000															
MT049951.1	0.000	0.000	0.000														
NC_045512.2	0.000	0.000	0.000	0.000													
AY654624.1	0.362	0.362	0.362	0.362	0.362												
KJ473816.1	0.388	0.388	0.388	0.389	0.388	0.312											
KJ473815.1	0.373	0.373	0.373	0.373	0.373	0.301	0.218										
KJ473814.1	0.375	0.375	0.375	0.375	0.375	0.291	0.234	0.190									
KJ473813.1	0.413	0.413	0.413	0.414	0.413	0.347	0.283	0.239	0.257								
KJ473812.1	0.413	0.413	0.413	0.414	0.413	0.348	0.282	0.238	0.256	0.001							
KJ473811.1	0.430	0.430	0.430	0.429	0.430	0.343	0.321	0.314	0.297	0.147	0.147						
DQ648790.1	1.201	1.201	1.201	1.201	1.201	1.258	1.182	1.204	1.203	1.231	1.232	1.340					
DQ648793.1	1.659	1.659	1.659	1.661	1.659	1.727	1.653	1.652	1.725	1.737	1.740	1.706	1.843				
DQ648791.1	1.859	1.859	1.859	1.862	1.859	1.873	1.790	1.794	1.841	1.908	1.908	1.836	1.995	0.625			
DQ648792.1	1.863	1.863	1.863	1.863	1.863	1.854	1.788	1.797	1.852	1.890	1.890	1.820	2.020	0.635	0.037		
JQ404410.1	1.625	1.625	1.625	1.628	1.625	1.674	1.707	1.695	1.712	1.771	1.766	1.767	1.886	0.790	0.78	0.78	

The evolutionary distances were computed using the Maximum Composite Likelihood method [2] and are in the units of the number of base substitutions per site. The analysis involved 17 nucleotide sequences. Codon positions included 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There was a total of 3294 positions in the final dataset.

The evolutionary history amino acid sequences were inferred using the Neighbor-Joining method [1]. The optimal tree with the sum of branch length = 3.27713782 is shown (Fig. 3). The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. The analysis involved 17 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 1135 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Table 6).

DISCUSSION

The present study was taken up to analyze several widely used parameters of codon usage bias namely CAI, CBI, Fop, and ENC along with the base composition of the coding sequences of some essential genes in Coronaviridae. The accurate coding sequences were retrieved using a program in Perl, developed by the researchers involved in the current study. After a preliminary analysis of the base composition, it was found that the cds of Coronaviridae are rich in AT like Severe Acute Respiratory Syndrome (SARS) [11,28].

The coronavirus S-protein is the responsible structural protein for the CoV crown-like shape of the viral particles, from which the name “coronavirus” was originated [29]. The long S-protein with ~1200 aa is of the class-I viral fusion proteins and has roles in cell receptor binding, tissue tropism, and pathogenesis [30].

Transmembrane spike (S) glycoprotein mediated entry into host cells by forming homotrimers protruding from the viral surface [31]. S constructs two functional subunits

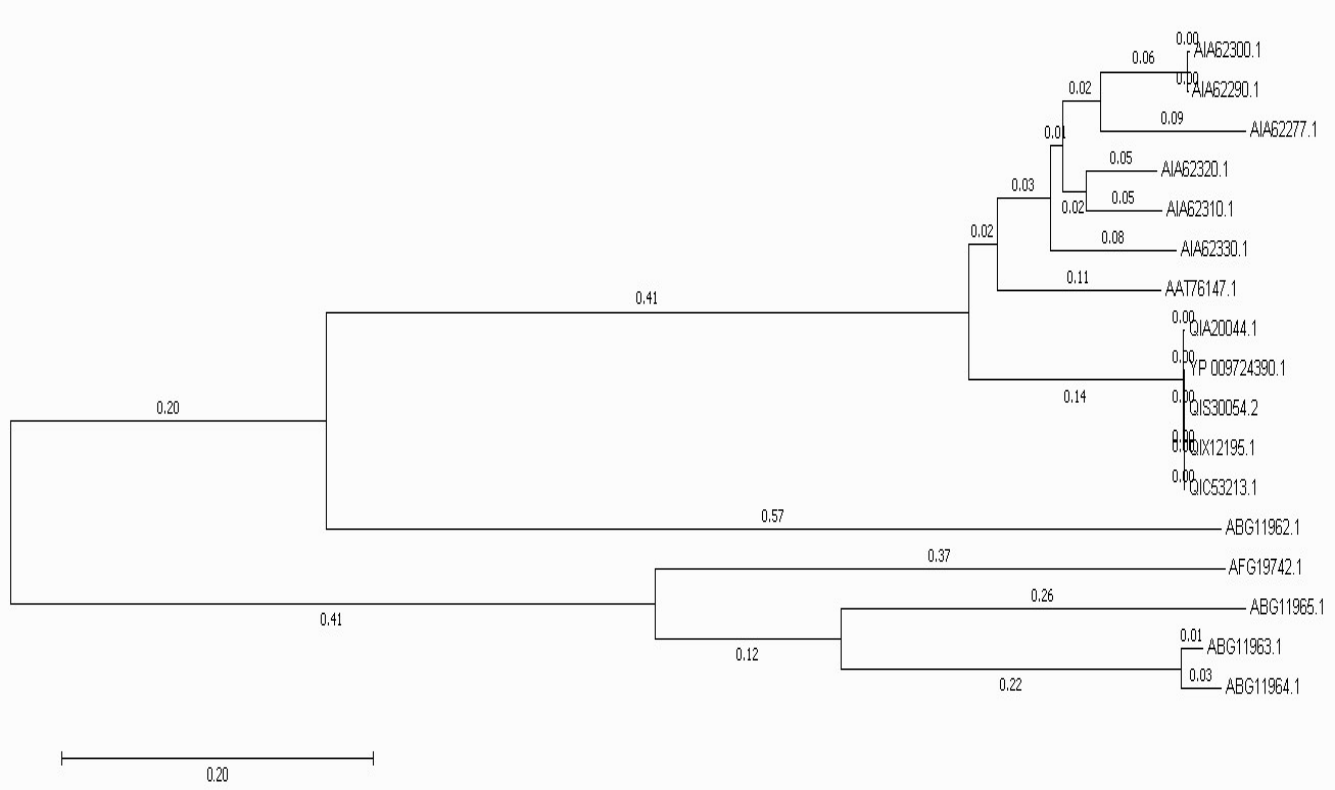

Fig. 3. Phylogenetic tree of the Coronaviridae spike glycoprotein amino acid sequences.

Table 6. Molecular Evolution and Phylogenetic Diagram Coronaviridae

QIX12195.1																	
QIC53213.1	0.000																
QIS30054.2	0.000	0.000															
QIA20044.1	0.001	0.001	0.001														
YP_009724390.1	0.000	0.000	0.000	0.001													
AAT76147.1	0.261	0.261	0.261	0.261	0.261												
AIA62330.1	0.272	0.272	0.272	0.272	0.272	0.215											
AIA62320.1	0.261	0.261	0.261	0.261	0.261	0.217	0.143										
AIA62310.1	0.262	0.262	0.262	0.262	0.262	0.210	0.148	0.094									
AIA62300.1	0.286	0.286	0.286	0.286	0.286	0.241	0.178	0.099	0.145								
AIA62290.1	0.284	0.284	0.284	0.284	0.284	0.240	0.177	0.098	0.145	0.002							
AIA62277.1	0.313	0.313	0.313	0.312	0.313	0.266	0.211	0.200	0.204	0.151	0.150						
ABG11962.1	1.121	1.121	1.121	1.121	1.121	1.105	1.132	1.105	1.110	1.134	1.137	1.132					
ABG11965.1	1.541	1.541	1.541	1.541	1.541	1.545	1.521	1.525	1.529	1.529	1.525	1.566	1.636				
ABG11963.1	1.509	1.509	1.509	1.509	1.509	1.489	1.525	1.497	1.485	1.521	1.517	1.545	1.596	0.496			
ABG11964.1	1.521	1.521	1.521	1.521	1.521	1.497	1.541	1.509	1.509	1.537	1.533	1.550	1.609	0.501	0.040		
AFG19742.1	1.550	1.550	1.550	1.550	1.550	1.541	1.529	1.521	1.509	1.521	1.517	1.529	1.497	0.747	0.721	0.724	

responsible for the host cell receptor binding (S1 subunit) and the viral fusion and cellular membranes (S2 subunit) [32]. For most CoVs, a cleavage has occurred at the boundary among the S1 and S2 subunits, which creates a non-covalently bound in the perfusion conformation [33]. The distal S1 subunit consists of the receptor-binding domain(s) and involves the stabilization of the perfusion state of the membrane-anchored S2 subunit which involves the fusion machinery [34]. For all CoVs, S is then cleaved via host proteases at the so-called S2′ site located instantly upstream of the fusion peptide [35]. This cleavage has been suggested to activate the protein for membrane fusion through extensive permanent conformational alternations [33]. Consequently, coronavirus entrance into susceptible cells is a sophisticated process requiring the concordant action of receptor-binding and proteolytic cleavage of the S protein for virus-cell fusion promotion [36].

Based on the virus strains and cell types, CoV S proteins may be cleaved by one or numerous host proteases, containing trypsin, furin, cathepsins, transmembrane protease serine protease-2 (TMPRSS-2), human airway trypsin-like protease (HAT), and TMPRSS-4 [8]. It has been reported that serine protease TMPRSS2 is essential for SARS-CoV-2 S activation. Moreover, SARS-CoV-2 S protein entrance to 293/hACE2 cells is mostly mediated via endocytosis, and that PIKfyve, TPC2, and cathepsin L are important for virus entry [8]. It was also demonstrated a special furin-like cleavage site in the Spike protein of the SARS-CoV-2, absent in the other SARS-like CoVs [29]. Because furin is expressed in the lungs highly, an enveloped virus that infects the respiratory tract may use this convertase successfully and activate its surface glycoprotein [37]. This furin-like cleavage site is suggested to be cleaved through virus egress for S-protein “priming” and may create a gain-of-function to the SARS-CoV-2 for effective spreading in the human population in comparison with other lineage b beta coronaviruses [29].

It has been found that the SARS-CoV-2 S2 subunit is highly conserved and has 99% identity with bat SARS-like CoVs (SL-CoV ZXC21 and ZC45) and human SARS-CoV. Therefore, the wide range of antiviral peptides against S2 would be a significant preventive and treatment modality for examining in animal models prior clinical trials [38]. In spite of sequence differences in the S1 domains, there are conserved residues presented in ternary folding. This reveals that the SARS-CoV-2 may have an interaction with some of the previously known host targets (ACE2, CD26, Ezrin, cyclophilins), however by few different molecular interactions. Recent researches also verify the possible SARS-CoV-2 and ACE-2 interaction [39].

Computational analyses are linked with numerous research studies like genomic analyses, evolution, and drug design [40]. Factors determined to have an effect on the CUB of an organism are nucleotide composition, the rate of synonymous substitution, tRNA abundance, codon hydropathy and initiation sites of the DNA replication, the length of a gene, and its expression level [41]. Therefore, it is important to evaluate the structures and compositions of the viral gene at the codon or nucleotide level to clarify the virus-host relationships mechanisms and virus evolution [42].

According to our obtained results, SARS-CoV-2 has a rich AT nucleotides composition that strongly affects its codon usage indicating mutation pressure rather than natural selection. In similar researches, the Nucleotide compositions and codon usage of different coronaviruses were compared and various patterns of both mutational bias and natural selection which influence the codon usage were found [43]. Exploring the codon usage in RNA viruses generally helps in finding the evolutionary history of viruses and the evolutionary forces, which form the viral genome, assisting in finding the features of novel appearing viruses. In addition, research on influenza A virus (IAV) [44] reported that finding codon usage and its nucleotide content in viruses may help in designing new vaccines using synthetic attenuated virus engineering (SAVE) could play a role in developing vaccines for IAV, through de-optimizing its codon it might be possible to reduce virus effects [45].

Moreover, we analyzed the evolution of the novel Iranian SARS-CoV-2 spike glycoprotein using phylogenetic analysis. The results showed that SARS-CoV-2/human/IRN/ and SARS-CoV-2/human/CHN/WH-09/2020 spike glycoprotein have the most similarities at nucleotide and amino acid sequences suggested their common ancestor. In previous studies, phylogenetic analyses were carried out among different SARS-CoV-2 isolates [42,46].

On the other hand, to better comprehend the structure of COVID-19, we modeled the homo-trimer structure of S

glycoprotein using the I-TASSER web server showing the most accurate fold and validation of the predicted structure. Similarly, in recent research, the structural modeling of the SARS-CoV-2 spike glycoprotein was performed [46]. The obtained model revealed an extended structural loop, which has basic amino acids at the receptor binding (S1) and fusion (S2) domains interface. It has been suggested that this loop confers fusion activation and entry features more consistent with beta coronaviruses in lineages A and C, and be a key member in the SARS-CoV-2 evolution with this structural loop influencing stability and transmission of the virus [46]. Several other studies also reported molecular modeling strategies to understand the exact protein structures [47-50].

Overall, such bioinformatic analyses can be applied for next practical experiments and clinical trials, and also a better comprehending of SARS-CoV-2 replication and pathogenesis. A similar analysis performed on other viral agents could also provide novel insights in the field of viral performance.

REFERENCES

[1] S.G. Siddell, J. Ziebuhr, E.J. Snijder, Coronaviruses, Toroviruses, and Arteriviruses. Topley & Wilson's Microbiology and Microbial Infections, 2010.

[2] D. Cavanagh, The Coronavirus Surface Glycoprotein. The Coronaviridae: Springer, 1995, pp. 73-113.

[3] T.R. Ruch, C.E. Machamer, Viruses 4 (2012) 363.

[4] B.W. Neuman, G. Kiss, A.H. Kunding, D. Bhella, M.F. Baksh, S. Connelly, *et al*. Journal of Structural Biology 174 (2011) 11.

[5] C. Risco, I.M. Antón, L. Enjuanes, J.L. Carrascosa Journal of Virology 70 (1996) 4773.

[6] D. Forni, R. Cagliani, M. Clerici, M. Sironi, Trends in Microbiology 25 (2017) 35.

[7] A. Sheikh, A. Al-Taher, M. Al-Nazawi, A.I. Al-Mubarak, M. Kandeel, Journal of Virological Methods 277 (2020) 113806.

[8] X. Ou, Y. Liu, X. Lei, P. Li, D. Mi, L. Ren, *et al.* Nature Communications 11 (2020) 1.

[9] I.S. Belalov, A.N. Lukashev, PLoS One 8 (2013).

[10] T. Ikemura, Molecular Biology and Evolution 2 (1985) 13.

[11] G.M. Jenkins, E.C. Holmes, Virus Research 92 (2003) 1.

[12] Y. Chen, Q. Xu, X. Yuan, X. Li, T. Zhu, Y. Ma, *et al*. Oncotarget 8 (2017) 110337.

[13] Y. Zhang, BMC Bioinformatics 9 (2008) 40.

[14] W. Kaplan, T.G. Littlejohn, Briefings in Bioinformatics 2 (2001) 195.

[15] W.L. DeLano, The PyMOL Molecular Graphics System, 2002.

[16] P. Stothard, Biotechniques 28 (2000) 1102.

[17] U. Vetrivel, V. Arunkumar, S. Dorairaj, Bioinformation 2 (2007) 62.

[18] F. Wright, Gene 87 (1990) 23.

[19] J.L. Bennetzen, B.D. Hall, Journal of Biological Chemistry 257 (1982) 3026.

[20] H. Naya, H. Romero, N. Carels, A. Zavala, H. Musto, FEBS Letters 501 (2001) 127.

[21] M. Stenico, A.T. Lloyd, P.M. Sharp, Nucleic Acids Research 22 (1994) 2437.

[22] T. Zhou, W. Gu, J. Ma, X. Sun, Z. Lu, Biosystems 81 (2005) 77.

[23] S. Wu, Y. Zhang, Nucleic Acids Research 35 (2007) 3375.

[24] N. Guex, M. Peitsch, Protein Data Bank Quaterly Newsletter 77 (1996).

[25] G. Vriend, Journal of Molecular Graphics 8 (1990) 52.

[26] K. Tina, R. Bhadra, N. Srinivasan, Nucleic Acids Res. 35. Web Server issue) W473-W476, 2007.

[27] C. Supriyo, P. Prosenjit, T.H. Mazumder, Notulae Scientia Biologicae 6 (2014) 417.

[28] W. Gu, T. Zhou, J. Ma, X. Sun, Z. Lu, Virus Research 101 (2004) 155.

[29] B. Coutard, C. Valle, X. de Lamballerie, B. Canard, N. Seidah, E. Decroly, Antiviral Research 176 (2020) 104742.

[30] G. Lu, Q. Wang, G.F. Gao, Trends in Microbiology 23 (2015) 468.

[31] M.A. Tortorici, A.C. Walls, Y. Lang, C. Wang, Z. Li, D. Koerhuis, *et al.* Nature structural & molecular biology 26 (2019) 481.

[32] R.N. Kirchdoerfer, C.A. Cottrell, N. Wang, J.

Pallesen, H.M. Yassine, H.L. Turner, *et al.* Nature 531 (2016) 118.

[33] A.C. Walls, M.A. Tortorici, J. Snijder, X. Xiong, B.-J. Bosch, F.A. Rey, *et al*. Proceedings of the National Academy of Sciences 114 (2017) 11157.

[34] W. Song, M. Gui, X. Wang, Y. Xiang, PLoS Pathogens 14 (2018) e1007236.

[35] J.K. Millet, G.R. Whittaker, Virus Research 202 (2015) 120.

[36] A.C. Walls, Y.-J. Park, M.A. Tortorici, A. Wall, A.T. McGuire, D. Veesler, Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell, 2020.

[37] D.E. Bassi, J. Zhang, C. Renner, A.J. Klein-Szanto, Molecular Carcinogenesis 56 (2017) 11822017.

[38] J.F.-W. Chan, K.-H. Kok, Z. Zhu, H. Chu, K.K.-W. To, S. Yuan, *et al*. Emerging Microbes & Infections 9 (2020) 221.

[39] N. Vankadari, J.A. Wilce, Emerging Microbes & Infections 9 (2020) 601.

[40] M. Kandeel, T. Miyamoto, Y. Kitade, Biological and Pharmaceutical Bulletin 32 (2009) 1321.

[41] L. Wang, H. Xing, Y. Yuan, X. Wang, M. Saeed, J. Tao, *et al*. PloS One 13 (2018).

[42] A.M. Anwar, S.M. Khodary, Insights into The Codon Usage Bias of 13 Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Isolates from Different Geo-locations. bioRxiv, 2020.

[43] M. Dilucca, S. Forcelloni, A.G. Georgakilas, A. Giansanti, A. Pavlopoulou, Viruses 12 (2020) 498.

[44] N. Goñi, A. Iriarte, V. Comas, M. Soñora, P. Moreno, G. Moratorio, *et al.* Virology Journal 9 (2012) 263.

[45] J. Coleman, THE DISTILLERY.

[46] J.A. Jaimes, N.M. André, J.S. Chappie, J.K. Millet, G.R. Whittaker, Phylogenetic Analysis and Structural Modeling of SARS-CoV-2 Spike Protein Reveals an Evolutionary Distinct and Proteolytically-sensitive Activation loop. Journal of Molecular Biology, 2020.

[47] M. Mortazavi, S. Shakeri, M. Maleki, F. Khoshbasirat, Investigation and Bioinformatics Analysis of Squalene Synthase Gene and Protein in Native Strain of Aurantiochytrium, 2018.

[48] F. Kargar, M. Mortazavi, A. Savardashtaki, S. Hosseinkhani, M.T. Mahani, Y. Ghasemi, International Journal of Biological Macromolecules 124 (2019) 689.

[49] K. Jamshidi Goharrizi, F. Amirmahani, F. Fatehi, M. Nazari, S.S. Moosavi, Journal of Genetic Resources 5 (2019) 72.

[50] M. Fattahi, A. Malekpour, M. Mortazavi, A. Safarpour, N. Naseri, Middle East Journal of Digestive Diseases 6 (2014) 214.