International Scholars Journals

Advanced Journal of Microbiology Research ISSN 2241-9837 Vol. 13 (5), pp. 001-006, May, 2019. Available online at www.internationalscholarsjournals.org © International Scholars Journals

Full Length Research Paper

Synonymous codon usage bias of spike genes of porcine epidemic diarrhea virus

Cao H. W.[1,2], Zhang H.[2], Liu, Y.[1] and Li D. S.[1]*

[1]College of Life Science, Northeast Agricultural University, Haerbin 150030, China.
[2]College of Life Science and Technology, HeiLongJiang BaYi Agricultural University, Daqing 163319, China.

Accepted 19 March, 2019

Synonymous codons are not used randomly. Rather, some codons are used more frequently than others. Investigating codon usage bias is essential to the understanding of viral evolution. However the factors shaping synonymous codon usage bias and nucleotide composition in animal viruses have been studied only to a limited extent. In our study, using the 19 complete CDS sequences of Spike (S) genes of porcine epidemic diarrhea virus (PEDV), we analyzed synonymous codon usage bias. The relative synonymous codon usage (RSCU) was used to estimate codon usage variation in each gene, and the results indicated that preferentially used codons were A-ended, G-ended, and U-ended codons. Effective number of codons (ENC) values varies from 48.15 to 49.52, which suggested that the codon usage bias in PEDV S genes was very slightly. Correspondence analysis (COA) was further performed to study the major trend in codon usage variation, and the plot of ENC values against GC3s (at synonymous third codon position) revealed that mutational pressure rather than translational selection was the main factor determining the codon usage bias in PEDV S genes. Moreover, correlation analysis indicated that aromaticity and hydrophobicity of S genes also influenced the codon usage variation in a minor way. Our study provides the most comprehensive analysis to date of PEDV S genes' codon usage patterns and provides a basic understanding of the mechanisms for codon usage bias.

Key words: Porcine epidemic diarrhea virus, Spike gene, relative synonymous codon usage, effective number of codons, correspondence analysis, correlation analysis.

INTRODUCTION

Synonymous codons are not used randomly and it has been previously studied in many prokaryotes and some lower eukaryotes (Aota and Ikemura 1986; Sharp et al., 1986; Marin et al., 1989). Previous studies proved that some codons are used more frequently than others (Tao et al., 2009). Studies of the synonymous codon usage can reveal information about the molecular evolution of individual genes and provide data to train genome-specific gene recognition algorithms (Bulmer, 1991), which detect protein coding regions in uncharacterized genomic DNA (Zhong et al., 2007).

Porcine epidemic diarrhea virus (PEDV), first recognized in 1977 (Pensaert and de Bouck, 1978), is an enveloped, single-stranded RNA virus belonging to the family *Coronaviridae*. The PEDV genome contains genes for the following proteins: pol1 (P1), spike (S) (180–220 kDa), envelope (E), membrane (M) (27–32 kDa), and nucleocapsid (N) (55–58 kDa) (Saif, 1993). Among the proteins, S, a glycoprotein peplomer (surface antigen) on the viral surface, plays an important role in the attachment of viral particles to the receptors of host cells with subsequent penetration into the cells by membrane fusion. The S glycoprotein also stimulates the induction of neutralizing antibodies in the host (Duarte and Laude, 1994). Furthermore, S gene is extensively used for evolutionary analysis. Phylogenetic analysis indicates PEDV is mainly classified into Groups 1 (Park et al., 2007). Understanding the extent and causes of

*Corresponding author. E-mail: caohongwei1974@gmail.com or deshanli@163.com. Tel: +86-451-55190645. Fax: +86-451-55190645.

Table 1. Data information of porcine epidemic diarrhea virus S genes used in this study.

SN	Strains	Accession No.	ENC	GC3s	GC	GRAVY	Aromo
1	Korean	AF500215	48.77	0.337	0.419	0.094581	0.111994
2	SM98	GU937797	49.43	0.338	0.414	0.116811	0.107971
3	JS-2004-2	AY653204	48.15	0.330	0.412	0.117860	0.111352
4	KNU-0905	GU180148	48.53	0.339	0.418	0.101804	0.111833
5	KNU-0904	GU180147	48.49	0.334	0.418	0.112482	0.112554
6	KNU-0903	GU180146	48.52	0.339	0.418	0.098990	0.112554
7	KNU-0902	GU180145	49.33	0.345	0.421	0.099928	0.111111
8	KNU-0901	GU180144	48.55	0.331	0.418	0.109091	0.112554
9	KNU-0802	GU180143	49.09	0.339	0.418	0.111183	0.111111
10	KNU-0801	GU180142	48.87	0.335	0.416	0.105628	0.112554
11	CV777	AF353511	49.45	0.338	0.416	0.123427	0.108460
12	DR13	DQ862099	49.49	0.338	0.415	0.117787	0.109183
13	DR13	DQ462404	49.20	0.340	0.414	0.122431	0.112156
14	Chinju99	AY167585	49.14	0.338	0.416	0.091251	0.114244
15	DX	EU031893	48.20	0.330	0.414	0.097036	0.112798
16	CV777	NC_003436	49.45	0.338	0.416	0.123427	0.108460
17	LZC	EF185992	49.52	0.338	0.416	0.127910	0.108460
18	Br1/87	Z25483	49.42	0.338	0.416	0.125163	0.109183
19	LJB/03	DQ985739	48.69	0.338	0.416	0.104049	0.111352

SN: serial number; accession No.: GenBank accession numbers; ENC: effective number of codons; GC3s: synonymous third codon position; GRAVY: hydrophobicity; Aromo; aromatic amino acids.

biases in codon usage of PEDV S gene is essential to the understanding of viral evolution, particularly the interplay between viruses and the immune response.

It is well known that mutational pressure and translational selection were thought to be the main factors that account for codon usage variation among genes in different organisms (Zhou et al., 2005; Zhang et al., 2010). Up to date, many organisms such as bacteria (Wright and Bibb, 1992), yeast (Sharp and Lloyd., 1993), drosophila (Hiroshi et al., 1998; Rubin, 1998), and mammals (Marin et al., 1989), where codon usage bias and nucleotide composition have been studied in great details, however, the factors shaping synonymous codon usage bias and nucleotide composition in animal viruses, especially in PEDV, have been few reported. In our study, we aimed to better understand the characteristics of the S gene of PEDV and further to reveal more information about its evolution. Synonymous codon usage of S genes has been analyzed and resulted into slight codon usage bias.

MATERIALS AND METHODS

Virus sequences

The available 19 complete CDS of S gene of PEDV were downloaded from GeneBank website (http://www.ncbi.nlm.nih.gov/), EMBL website (http://www.ebi.ac.uk/embl/) and DDBJ (http://www.ddbj.nig.ac.jp/searches-e.html). Sequences with >99% sequence identities were excluded. Clustal X (version 1.83) (Thompson et al., 1997) was used to align gene sequences. The serial number (SN), strains, and GenBank accession numbers (accession No.) were listed in Table 1.

Codon usage indices analysis

Relative synonymous codon usage (RSCU) values of each codon in each genes were used to measure the synonymous codon usage (Sharp et al., 1986). RSCU values are largely independent of amino acid composition and are particularly useful in comparing codon usage between genes, or sets of genes that differ in their size and amino acid composition. The preferred codon usage of each gene was analyzed using software package GCUA (version 1.0) (http://bioinf.may.ie/downloads.html) (Sharp et al., 1986). The effective number of codons (ENC) was used to quantify the codon usage bias of each gene, which is the best overall estimator of absolute synonymous codon usage bias (Wright, 1990). The GC index was used to calculate the overall GC content in each genes, while the index GC3s was used to calculate the fraction of GC nucleotides at the synonymous third codon position (excluding Met, Trp, and the termination codons) (Richard et al., 2000). The general average hydrophobicity (GRAVY) score and the frequency of aromatic amino acids (Aromo) in the hypothetical translated gene product were also computed (Kyte and Doolittle, 1982). All the indices mentioned above were calculated using the program CodonW (version 1.4) (Sharp et al., 1986).

Correspondence analysis

Multivariate statistical analysis was performed for the relationships between variables and samples. We used correspondence analysis (COA) to study the major trend in codon usage variation (Gupta and Doolittle, 2001). Each dimension corresponds to the RSCU value of one sense codon (excluding AUG, UGG, and termination codons). Major trends within this dataset can be determined using measures of relative inertia and genes ordered according to their positions along the axis of major inertia (Grantham et al., 1981).

Statistical analysis

Correlation analysis was carried out using Spearman's rank

Table 2. Synonymous codon usage in porcine epidemic diarrhea virus S genes.

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	1145	1.18	Ser	UCU	634	1.71
	UUC	789	0.82		UCC	229	0.62
Leu	UUA	655	1.62		UCA	591	1.60
	UUG	541	1.34		UCG	106	0.29
Tyr	UAU	703	0.92	Cys	UGU	1348	1.19
	UAC	821	1.08		UGC	909	0.81
ter	UAA	1253	0.00	ter	UGA	1273	0.00
ter	UAG	667	0.00	Trp	UGG	1093	1.00
Leu	CUU	408	1.01	Pro	CCU	298	1.34
	CUC	328	0.81		CCC	125	0.56
	CUA	268	0.66		CCA	366	1.65
	CUG	219	0.54		CCG	98	0.44
His	CAU	678	1.02	Arg	CGU	249	0.96
	CAC	649	0.98		CGC	106	0.41
Gln	CAA	683	1.40		CGA	213	0.82
	CAG	292	0.60		CGG	179	0.69
Ile	AUU	315	1.40	Thr	ACU	444	1.52
	AUC	227	1.01		ACC	315	1.08
	AUA	133	0.59		ACA	360	1.23
Met	AUG	130	1.00		ACG	48	0.16
Asn	AAU	170	0.75	Ser	AGU	256	0.69
	AAC	281	1.25		AGC	405	1.09
Lys	AAA	253	1.42	Arg	AGA	452	1.73
	AAG	104	0.58		AGG	365	1.40
Val	GUU	484	1.66	Ala	GCU	332	1.24
	GUC	313	1.08		GCC	241	0.90
	GUA	205	0.70		GCA	327	1.22
	GUG	162	0.56		GCG	170	0.64
Asp	GAU	178	0.92	Gly	GGU	398	1.33
	GAC	209	1.08		GGC	322	1.08
Glu	GAA	133	1.02		GGA	242	0.81
	GAG	128	0.98		GGG	235	0.79

The preferentially used codons (A-ended, G-ended, and U-ended codons, RSCU>1.2) for each amino acid are displayed in bold. AA, amino acids; N, number of codons; RSCU, cumulative relative synonymous codon usage; ter, termination codon.

correlation analysis method. All statistical analyses were carried out using the statistical analysis software SPSS Statistics (Version 17.0).

RESULTS AND DISCUSSION

Synonymous codon usage variation in S genes

We computed the RSCU values of different codon in S genes to investigate the extent of codon bias in PEDV S genes. All data information of cumulative codon usage of 59 codons in 19 PEDV S genes were displayed in Table 2. The preferentially used codons were A-ended, G-ended, and U-ended codons. It was interesting to note that few U-ended codons were used as preferential codons. ENC values range from 20 to 61; the larger the extent of codon preference in a gene, the smaller the corresponding ENC value. In a highly biased gene where only one codon is used for each amino acid, the ENC value = 20. Conversely, in a gene exhibiting no bias, the value will be 61 (Wright, 1990). Our data showed that the ENC values of different PEDV genes vary from 48.15 to 49.52, with a mean of 48.96 and standard deviation (S.D.) of 0.4636, which indicated that the codon usage bias in PEDV S genes was very slightly. In addition, GC and GC3s values were

Table 3. Explanation of the variation by axis.

No.	R. Iner	R. Sum
1	+0.3081	+0.3081
2	+0.2288	+0.5369
3	+0.1634	+0.7003
4	+0.0874	+0.7877

No.: serial number of axis; R. Iner: each principal axis accounting variation; R. Sum: total variation.

Table 4. The position of each gene by two axis.

Strains	Axis1	Axis2
Korean	-0.03441	-0.00680
SM98	-0.03441	-0.00680
JS-2004-2	-0.03352	-0.00969
KNU-0905	-0.03320	-0.00788
KNU-0904	-0.03569	-0.00455
KNU-0903	-0.05208	0.01414
KNU-0902	-0.00460	0.01691
KNU-0901	-0.01706	0.02486
KNU-0802	-0.04631	0.01765
KNU-0801	-0.05480	0.05638
CV777	0.02702	-0.07463
DR13	0.00286	-0.06723
DR13	0.01071	-0.07483
Chinju99	0.04113	-0.01594
DX	0.07738	0.03149
CV777	0.07526	0.03419
LZC	0.02858	0.01082
Br1/87	0.06659	0.02498
LJB/03	0.01656	0.03693

calculated and listed in Table 1. The average GC content of PEDV S genes was 0.4164 (from 0.412 to 0.421, with a S.D. of 0.0021), while average GC3s content in codons was 0.337 (from 0.330 to 0.345, with a S.D. of 0.0037). Based on these findings, we confirmed that are PEDV S genes aren't GC-poor genomes.

Correspondence analysis of codon usage

In order to investigate synonymous codon usage variation, COA was implemented for 19 PEDV S genes selected in this study. We observed that the first principal axis accounted for 30.81% of the total variation, and the next three axes accounted for 22.88%, 16.34%, and 8.74% of the variation, respectively (Table 3). This observation demonstrated that although the first major axis could explains a substantial amount of variation in trends in codon usage, the second major axis also had an appreciable impact on total variation in synonymous codon usage. The position of each gene by axis was displayed in Table 4. Figure 1 depicted the position of each S gene on the plane defined by the first and second principal axes generated by COA on RSCU values. 19 S genes were random distributed on the plane, which supported the evidence that codon usage bias in PEDV S genes were very slightly.

Mutational bias is the main factor determining codon usage variation

Mutational pressure and translational selection are thought to be the main factors accounting for codon usage variation in genes (Gareth and Edward, 2003). In order to investigate whether codon usage variation of PEDV S genes is determined by mutational bias, correlation analysis to correlate the first two axes of COA with codon usage indices were employed. Correlation analysis showed that axis 1 of COA and axis 2 were not

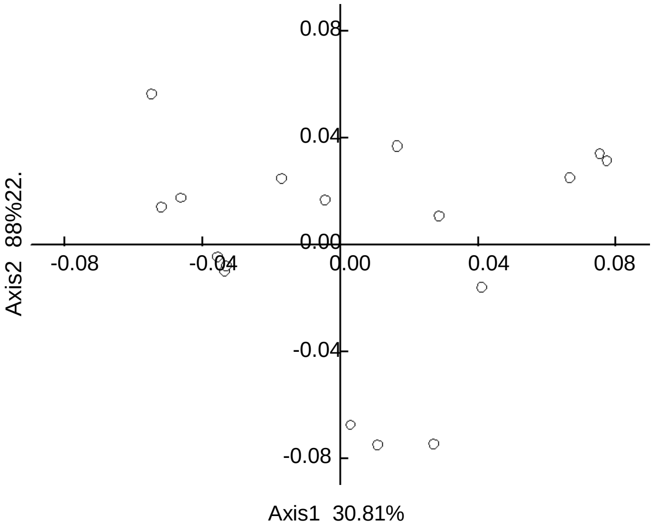

Figure 1. A plot of value of the first and second axis of each PEDV S gene in COA. The first axis accounts for 30.81% of total variation and the second axis accounts for 22.88% of total variations. 19 S genes were random distributed on the plane, which supported the evidence that codon usage bias in PEDV S genes were very slightly.

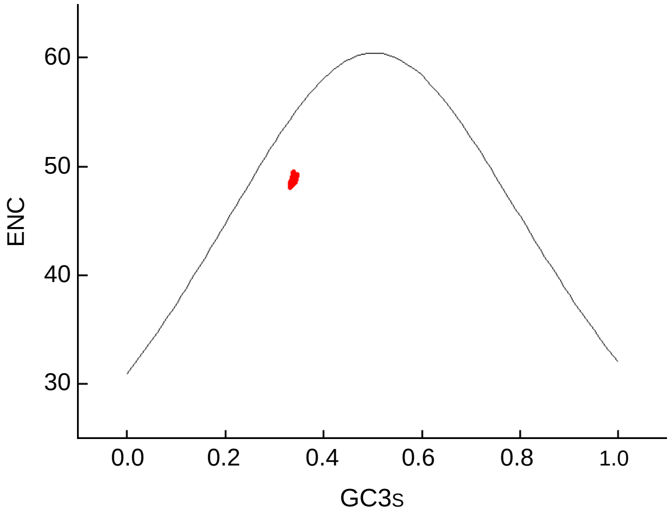

Figure 2. Effective number of codons used in each gene plotted against the GC3s. The continuous curve depicted the relationship between GC3s and ENC in the absence of selection. All of spots (red spots) lie below the expected curve, which indicated that codon choice is constrained only by a G + C mutation bias.

correlated with GC and GC3s. G+C content at the first and second codon positions (GC12s) and GC3s are correlated (*r*=0.386, *P* < 0.05). At the same time, GC content and GC3 s significantly correlated (*r*= 0.482, *P* < 0.05). This result implied that they are most likely the result of mutational pressure, because natural selection would be expected to act differently on different codon positions (Bulmer, 1991; Tang et al., 2008).

Moreover, ENC-plot (ENC plotted against GC3s) was used as part of a general strategy to investigate patterns of synonymous codon usage (Gupta and Ghosh, 2001). Genes, whose codon choice is constrained only by a G + C mutation bias, will lie on or just below the curve of the predicted values (Wright, 1990; Zhong et al., 2007). All of the spots lied below the expected curve in Figure 2, indicating that the codon usage bias in these 19 PEDV S genes was greatly influenced by the GC compositional constraints. In addition, a significantly positive correlation (*r*= 0.661, *P*<0.05) between GC3s and ENC values was observed, which indicated the patterns of condon usage also appear to be closely related to the GC content on the third codon position.

These results indicated that most of the codon usage bias among genes was directly related to the nucleotide composition. Therefore, it is concluded that the compositional constraint which mainly caused by mutation bias is the main determinant of the variation in synonymous codon usage.

Aromaticity and hydrophobicity affect codon usage to a low extent

Beyond the aforementioned factors, we were also concerned with whether other factors in PEDV S genes can explain their codon usage. As we all known, selection pressure was thought to be the other factor contributing to the codon usage variation among S genes, thus we performed a correlation analysis to evaluate whether GRAVY and Aromo values were related to first two axes of COA, ENC, GC and GC3s (Lobry and Gautier, 1994). Analysis results showed that only both GRAVY and Aromo was correlated with ENC (*r*=0.520, *P*<0.05; *r*=-0.692, *P*<0.01), while GRAVY and Aromo were not correlated with two axes, GC and GC3s (Table 5). The results indicated that the degree of hydrophobicity and the aromatic amino acids (Phe, Tyr, Trp) were associated with the codon usage variation to a low extent. These findings were in accordance with the above results that mutational bias rather than selection pressure is the main factor determining codon usage variation.

Conclusion

We analyzed the synonymous codon usage biases in 19 PEDV S genes, and found that PEDV S genes had low codon usage bias. Mutational pressure rather selection pressure is the main factor determining the codon usage biases. Furthermore, aromaticity and hydrophobicity could be partially accounting for the codon usage variation.

Table 5. Correlation analysis among GRAVY, Aromo, ENC, GC3s, GC and the first two axes in COA.

		Axis 1	**Axis2**	**ENC**	**GC3s**	**GC**
GRAVY	*r*	0.249	-0.299	0.520*	0.017	-0.427
	P	0.304	0.213	0.022	0.944	0.068
Aromo	*r*	-0.271	0.187	-0.692**	-0.275	0.213
	P	0.262	0.444	0.001	0.255	0.382

r, correlation coefficient; **P*-value ≤ 0.05; ***P*-value ≤ 0.01.

ACKNOWLEDGEMENT

The study was supported by the Technology Research Foundation of Education Department of HeiLongJiang Province, China (No.12511352 and 11541251).

REFERENCES

Aota S, Ikemura T (1986). Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic. Acids. Res., 14: 6345-6355.

Bulmer M (1991). The selection-mutation-drift theory of synonymous codon usage. Gene, 129: 897-907.

Duarte M, Laude H (1994). Sequence of the spike protein of the porcine epidemic diarrhoea virus. J. Gen. Virol., 75: 1195-1200.

Gareth MJ, Edward CH (2003). The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res., 92: 1-7.

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981). Codon catalogue usage is a genome strategy for genome expressivity. Nucleic Acids Res., 9: 43-75.

Gupta SK, Ghosh TC (2001). Expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. Gene, 273: 63-70.

Hiroshi A, Richard MK, Adam EW (1998). Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. Gene., 102: 49-60.

Kyte J, Doolittle R (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol., 157: 105-132.

Lobry JR, Gautier C (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 *Escherichia coli* chromosome encoded genes. Nucleic. Acids Res., 22: 3174-3180.

Marin A, Bertranpetit J, Oliver JL, Medina JR (1989). Variation in G+C content and codon choice: differences among synonymous codon groups in vertebrate genes. Nucleic Acids Res., 17: 6181-6189.

Park SJ, Moon HJ, Yang JS, Lee CS, Song DS, Kang BK, Park BK (2007). Sequence analysis of the partial spike glycoprotein gene of porcine epidemic diarrhea viruses isolated in Korea. Virus Gene, 321-332.

Pensaert MB, de Bouck P (1978). A new coronavirus-like particles associated with diarrhea in swine. Arch. Virol., 58: 243-247.

Richard JE, Lin K, Tan T (2000). A functional significance for codon third bases. Gene, 245: 291-298.

Rubin GM (1998). The *Drosophila* genome project: a progress report. Trends. Gene, 14: 340-343.

Saif LJ (1993). Coronavirus immunogens. Vet. Microbiol., 37: 285-297.

Sharp PM, Lloyd AT (1993). Regional base composition variation along yeast chromosome III evolution of chromosome primary structure. Nucleic Acids Res., 21: 179-183.

Sharp PM, Tuohy TMF, Mosurski KR (1986). Codon usage in yeast cluster-analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res., 14: 5125-5143.

Tang FQ, Pan ZS, Zhang CY (2008). The selection pressure analysis of classical swine fever virus envelope protein genes E-rns and E2. Virus Res., 131: 132-135.

Tao P, Dai L, Luo MC, Tang FQ, Tien P, Pan ZS (2009). Analysis of synonymous codon usage in classical swine fever virus. Virus Gene, 38: 104-112.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res., 25: 4876-82.

Wright F (1990). The effective number of codons used in a gene. Gene, 87: 23-29.

Wright F, Bibb MJ (1992). Codon usage in the G+C rich *Streptomyces* genome. Gene, 113: 55-65.

Zhang H, Wang YH, Cao HW, Cui YD (2010). Phylogenetic analysis of E2 genes of classical swine fever virus in China. Isr. J. Vet. Med., 65: 151-155.

Zhong JC, Li YM, Zhao S, Liu SG, Zhang ZD (2007). Mutation pressure shapes codon usage in the GC-Rich genome of foot-and-mouth disease virus. Virus Genes, 35: 767-776.

Zhou T, Gu WJ, Ma JM, Sun X, Lu ZH (2005). Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. Biosystems, 81: 77-86.