Synonymous Codon Usage Bias in Porcine Epidemic Diarrhea Virus

Cao, H.W. and Zhang, H.*

College of Biological Science and Technology, HeiLongJiang BaYi Agricultural University, DaQing 163319, China.

* **Corresponding author:** Zhang, H. Tel: (086) 0459-6819290, Fax: (086) 0459-6819290. E-mail: huazi8541@sina.com

ABSTRACT

In this study, we analyzed the synonymous codon usage bias in porcine epidemic diarrhea virus genome. The effective number of codons (ENC) and the relative synonymous codon usage (RSCU) values are used to estimate codon usage variation. The plot of ENC values against GC3s and correlation analysis revealed that mutational pressure rather than translational selection is the main factor determining the codon usage bias in porcine epidemic diarrhea virus. In addition, other factors, such as aromaticity, hydrophobicity of genes also influenced the codon usage variation among virus genomes in a minor way. Our work might contribute to the understanding of molecular evolution of codon usage variation in porcine epidemic diarrhea virus genome.

Keywords: Synonymous codon usage, Mutational bias, GC content, Porcine epidemic diarrhea virus, Correspondence analysis.

INTRODUCTION

Synonymous codons are not used randomly, and synonymous codon usage bias exists in a wide range of biological systems from prokaryote to eukaryote. The usage frequencies of alternative synonymous codons vary among organisms (1). A variety of factors have been found to cause or correlate with the codon usage bias, such as mutational bias, replicational and translational selection, dinucleotide bias, gene length, tRNA abundance, codon-anticodon interaction, and tissue or organ specificity (2). Synonymous codon usage analysis has several aspects of potential applications, including protein structure prediction, gene classification and function prediction. Recently, codon usage biases in viruses have been documented with increasing frequency. Thus, understanding codon patterns and the relevant underlying factors can offer insights into the molecular evolution of viruses.

Porcine epidemic diarrhea virus (PEDV), first recognized in 1977, 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) (3, 4). 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 (5, 6). In our previous study, we found that mutational pressure rather than translational selection was the main factor determining the codon usage bias in PEDV S genes (7, 8). Codon usage bias and nucleotide composition have been studied in great detail in many organisms such as bacteria, yeast, drosophila, and mammals. However, the factors shaping synonymous codon usage bias and nucleotide composition in PEDV genome have been infrequently reported. In our study, we aimed to better understand the characteristics of the PEDV genome and further to reveal more information about its evolution. Synonymous codon usage of PEDV genome has been analyzed indicating slight codon usage bias.

MATERIALS AND METHODS

Data sets

A total of 410 publicly available complete coding sequences representing a set of deferent typical subtypes of porcine epidemic diarrhea virus were randomly retrieved from GenBank (http://ncbi.nlm.hih.gov/). The open reading frame (ORF) length of each gene and GenBank accession numbers are aligned using Clustal X (version 1.83) (9).

Codon usage indices

Relative synonymous codon usage (RSCU) values of each codon in each ORF were used to measure the synonymous codon usage. RSCU values are largely independent of amino acid composition and are particularly useful in comparing codon usage between genes. In our study, the preferred codon usage of each gene was analyzed using software package GCUA version 1.0 (http://bioinf.may.ie/downloads.html). The effective number of codons (ENC) were used to quantify the codon usage bias of each ORF (10, 11). ENC values ranged 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 (12). The GC index was used to calculate the overall GC content in the gene, 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). At the amino acid level, the general average hydrophobicity score (GAH) and the frequency of aromatic amino acids (Aromaticity) in the putative gene product were also analyzed. All the indices mentioned were calculated using the analysis program CodonW, version 1.4 (13).

Correspondence analysis

Correspondence analysis (COA) was used to study the major trend in codon usage variation among ORFs. In order to minimize the effects of amino acid composition on codon usage, each ORF is represented as a 59-dimensional vector; each dimension corresponds to the RSCU value of one sense codon (excluding AUG, UGG, and stop 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.

Correlation analysis

Spearman's rank correlation analysis was performed to determine the role of different factors in shaping the codon usage biases in the various subtypes of porcine epidemic diarrhea virus.

RESULTS AND DISCUSSION

Synonymous codon usage variation in porcine epidemic diarrhea virus

The overall RSCU values of 59 sense codons for the porcine epidemic diarrhea virus are shown in Table 1. The preferentially used codons were A-ended (8 ones) and U-ended (4 ones) codons. The average GC content of all porcine epidemic diarrhea virus was 45.34% (From 38.5% to 51.8%, with a Standard Deviation (S.D.) of 2.56%), while average GC3s content in codons was 42.9% (From 33.3% to 54.3%, with a S.D. of 4.12%). This is consistent with previous observations that porcine epidemic diarrhea virus is GC-poor genomes, and so it is expected that A-ended and/or U-ended codons are preferentially used. In addition, the ENC values vary from 44.03 to 61, with a mean of 52.58 and S.D. of 11.24. All of the ENC values are higher than 40, suggesting the codon usage biases among these subtypes and the genes of porcine epidemic diarrhea virus vary considerably. Besides, the average codon usage bias value (= 0.1) calculated by the CodonO online server (http://www.sysbiology.org/CodonO/index.php) for the data sets also suggested that the porcine epidemic diarrhea virus as a whole had low codon usage bias. It should be noted here that for CodonO program, the average codon usage bias value calculated is ranging from 0 to 1, higher values denotes higher bias, and vice versa.

Correspondence analyses of codon usage variations

Figure 1 depicts the position of each ORF on the plane defined by the first and second principal axes generated by COA on RSCU values of ORFs. A preliminary COA identified a major trend in the axis 1, which accounted for 25.1% of the total variation. The axis 2 explained 18.7% of total variation. This suggested that although the axis 1 explained a substantial amount of variation in codon usage, the axis 2 also played an appreciable role. If not specifically mentioned, the values of the first two axes of this COA were used for correlation analysis hereafter.

Table 1: Synonymous codon usage in PEDV viruses.

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	1183	0.86	Ser	UCU	974	1.05
	UUC	**1528**	**1.14**		UCC	723	0.83
Leu	UUA	524	0.55		**UCA**	**1264**	**1.37**
	UUG	**889**	**0.92**		UCG	372	0.35
Tyr	**UAU**	**993**	**1.05**	Cys	UGU	626	0.94
	UAC	917	0.95		**UGC**	**773**	**1.08**
Leu	**CUU**	**1237**	**1.26**	Pro	CCU	798	1.13
	CUC	1069	1.08		CCC	543	0.72
	CUA	961	0.99		**CCA**	**1124**	**1.61**
	CUG	1189	1.22		CCG	373	0.51
His	**CAU**	**747**	**1.24**	Arg	CGU	180	0.22
	CAC	464	0.77		CGC	262	0.31
Gln	**CAA**	**1474**	**1.01**		**CGA**	**505**	**0.54**
	CAG	1458	0.98		CGG	453	0.52
Ile	**AUU**	**1625**	**1.08**	Thr	ACU	1254	1.05
	AUC	1375	0.82		ACC	1012	0.82
	AUA	1632	1.05		**ACA**	**2076**	**1.74**
					ACG	402	0.32
Asn	**AAU**	**2144**	**1.17**	Ser	AGU	1032	1.23
	AAC	1650	0.87		**AGC**	**1088**	**1.31**
Lys	**AAA**	**2318**	**1.14**	Arg	**AGA**	**2307**	**2.74**
	AAG	1778	0.84		AGG	1385	1.74
Val	GUU	878	0.87	Ala	GCU	1204	1.08
	GUC	875	0.88		GCC	973	0.85
	GUA	831	0.91		**GCA**	**1998**	**1.76**
	GUG	**1485**	**1.45**		GCG	382	0.32
Asp	**GAU**	**1934**	**1.19**	Gly	GGU	676	0.52
	GAC	1419	0.85		GGC	663	0.54
Glu	**GAA**	**2832**	**1.12**		GGA	2256	1.87
	GAG	2222	0.88		GGG	1256	1.17

AA) Amino acids; N) number of codons; RSCU) cumulative relative synonymous codon usage.
The preferentially used codons for each amino acid are displayed in bold.

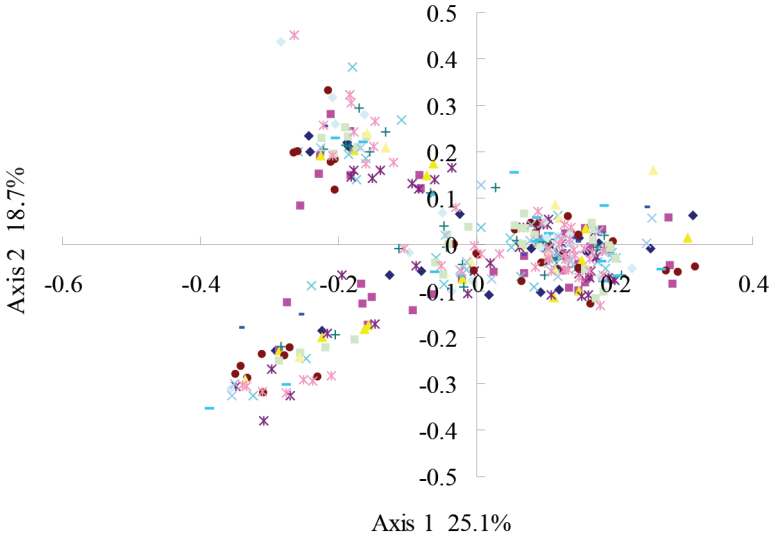

Figure 1: A plot of values of the first and second axis of each ORF in COA. The first axis accounts for 25.1% of all variation among ORFs, and the second axis accounts for 18.7% of total vibrations.

Mutational bias is the main factor determining codon usage in porcine epidemic diarrhea virus

In order to investigate whether codon usage variation of different genomes in porcine epidemic diarrhea virus is determined by mutational bias, we employed correlation analysis to correlate the first two axes of COA with codon usage indices. Our data showed that axis 1 of COA is correlated with GC (r= -0.65, p<0.001), while axis 2 is correlated with both GC3s (r= 0.51, p<0.001) and GC (r= 0.23, p<0.001) (Table 2), indicating that the patterns of base composition are most likely the results of mutational pressure, and not natural selection, since the effects are present at all codon positions.

Furthermore, Wright suggested that the ENC-plot (ENC plotted against GC3s) could be used as a general strategy to investigate patterns of synonymous codon usage. Genes, whose codon choice is solely constrained only by the GC mutation bias, will lie on or just below the curve of the expected ENC value. The ENC-plot (Figure 2) showed that the actual codon usage indices are close to the values expected from their GC composition. Most of the real points were lying below or on the expected curve, although only a few points lied above the expected curve. In addition, a significantly positive correlation (r= 0.35, p<0.001) between GC3s and ENC values was observed, which indicated the patterns of condon usage in different subtypes also appear to be closely related to the GC content on the third codon position. Most of the codon usage bias among these PEDV genomes is directly related to the nucleotide composition. Taken together, it could be concluded mutational bias is the main determinant of the variation in synonymous codon usage among the ORFs of different influenza A subtypes.

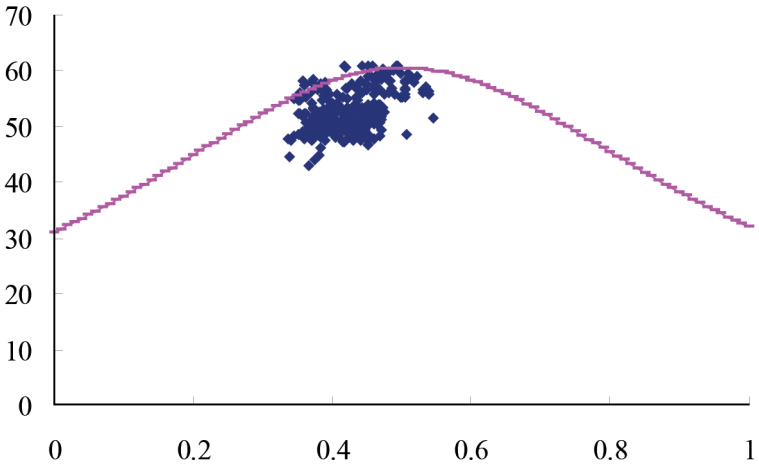

Figure 2: Effective number of codons used in each ORF plotted against the GC3s. The continuous curve plots the relationship between GC3s and NEC in the absence of selection. All of spots lie below the expected curve.

Table 2: Summary of correlation analysis between the first two axes in COA, GC3s, GRAVY, or aromaticity in the selected PEDV ORFs

		GRAVY	Aromaticity	GC	GC3s
Axis 1	*r*	0.431**	-0.114	-0.65**	-0.224
	P	0.002	0.524	p<0.001	0.184
Axis 2	*r*	0.436**	-0.553**	0.23	0.51**
	P	0.003	<0.001	p<0.001	p<0.001
ENC	*r*	-0.077	0.502**	0.42**	0.35**
	P	0.592	<0.001	p<0.001	p<0.001

* *P*-value ≤ 0.05; ** *P*-value ≤ 0.01.

Aromaticity and hydrophobicity affect codon usage

As showed in Figure 2, the majority of the actual ENC values are slightly lower than the expected ones, indicating that other factors may also influence the codon usage in porcine epidemic diarrhea virus. To test whether selection pressure contributes to the codon usage variation among porcine epidemic diarrhea virus, we performed a correlation analysis to evaluate whether Aromaticity and GRAVY values were related to first two axes of COA and ENC values (Table 2). Our results showed that GRAVY was correlated with ENC and both axis 1 (r= 0.431, p= 0.002) and axis 2 (r= 0.436, p= 0.003), while Aromaticity was correlated with ENC and axis 2 (r= -0.553, p<0.001), indicating that the degree of hydrophobicity and the frequency of aromatic amino acids (Phe, Tyr, Trp) were also associated with the codon usage variation.

CONCLUSION

Synonymous codon usage biases in porcine epidemic diarrhea virus were analyzed. Our results showed that porcine epidemic diarrhea virus had low codon usage bias. Mutational pressure is the main factor determining the codon usage biases in porcine epidemic diarrhea virus. Additional factors, such as aromaticity, hydrophobicity and gene length, could be partially accounted for the codon usage variation.

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (NSFC, Grant Nos. 31200121, 31310103018 and 31300145).

REFERENCES

1. Wright, F. and Bibb, M.J.: Codon usage in the G+C rich *Streptomyces* genome. Gene. 113: 55-65, 1992.
2. Cao, H.W., Zhang, H., and Cui, Y.D.: Synonymous Codon Usage Bias of E2 Genes of Classical Swine Fever Virus. Israel Journal of Vet. Med. 67: 253-258, 2012.
3. Song, D. and Park, B.: Porcine epidemic diarrhoea virus: a comprehensive review of molecular epidemiology, diagnosis, and vaccines. Virus Genes. 44: 167-175, 2012.
4. Chae, C., Kim, O., Choi, C., Min, K., Cho, W.S., Kim, J., and Tai, J.H.: Prevalence of porcine epidemic diarrhoea virus and transmissible gastroenteritis virus infection in Korean pigs. Vet. Rec. 147: 606-608, 2000.
5. Duarte, M. and Laude, H.: Sequence of the Spike Protein of the Porcine Epidemic Diarrhea Virus. J. Gen. Virol. 75: 1195-1200, 1994.
6. Sun, D.B., Feng, L., Shi, H.Y., Chen, J.F., Liu, S.W., Chen, H.Y., and Wang, Y.F.: Spike protein region (aa 636-789) of Porcine epidemic diarrhea virus is essential for induction of neutralizing antibodies. Acta Virologica. 51: 149-156, 2007.
7. Cao, H.W., Zhang, H., Liu, Y., and Li, D.S.: Synonymous codon usage bias of spike genes of porcine epidemic diarrhea virus. Afr. J. Microbiol. Res. 5: 3784-3789, 2011.
8. Chen, X., Yang, J.X., Yu, F.S., Ge, J.Q., Lin, T.L., and Song, T.Y.: Molecular characterization and phylogenetic analysis of porcine epidemic diarrhea virus (PEDV) samples from field cases in Fujian, China. Virus Genes. 45: 499-507, 2012.
9. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G.: The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic. Acids. Res.,. 25: 4876-82, 1997.
10. Banerjee, T., Gupta, S.K., and Ghosh, T.C.: Towards a resolution on the inherent methodological weakness of the "effective number of codons used by a gene". Biochem. Biophys. Res. Commun. 330: 1015-8, 2005.
11. Tang, Y., Cho, P.Y., Kim, T.I., and Hong, S.J.: Clonorchis sinensis: codon usage in nuclear genes. Exp. Parasitol. 115: 187-91, 2007.
12. Tao, P., Dai, L., Luo, M., Tang, F., Tien, P., and Pan, Z.: Analysis of synonymous codon usage in classical swine fever virus. Virus Genes. 38: 104-12, 2009.
13. Grocock, R.J. and Sharp, P.M.: Synonymous codon usage in Cryptosporidium parvum: identification of two distinct trends among genes. Int. J. Parasitol. 31: 402-12, 2001.