Synonymous Codon Usage Bias of E2 Genes of Classical Swine Fever Virus

Cao, H.W.,[#] Zhang, H.[#*] and Cui, Y.D.

College of Biological Science and Technology, HeiLongJiang BaYi Agricultural University, DaQing 163319, China
[#]Both authors contributed equally to this work.

[*] **Corresponding author:** Zhang, H., College of Bioscience and Technology, HeiLongJiang BaYi Agricultural University, 1 Xinyang road, Daqing 163319, PR China. E-mail: huazi8541@sina.com, tel.: +86 459 681 9299; fax: +86 459 681 9299.

ABSTRACT

In this study, synonymous codon usage bias in 44 E2 genes of classical swine fever virus (CSFV) was analyzed. The relative synonymous codon usage (RSCU) and effective number of codons (ENC) values were used to estimate codon usage variation in each gene. Correspondence analysis (COA) was used to study the major trend in codon usage variation. 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 CSFV E2 genes. Moreover, correlation analysis indicated that aromaticity of E2 genes also influenced the codon usage variation in a minor way. This study represents a comprehensive analysis to date of CSFV E2 genes' codon usage patterns and provides a basic understanding of the mechanisms for codon usage bias.

Key words: Classical swine fever virus, Envelope glycoprotein E2, Relative synonymous codon usage, Effective number of codons, Correspondence analysis.

INTRODUCTION

Synonymous codons are not used randomly as has been previously shown in many prokaryotes and some lower eukaryotes (1, 2, 3). Studies of the synonymous codon usage can reveal information about the molecular evolution of individual genes and provide data to prepare genome-specific gene recognition algorithms (4), which detect protein coding regions in uncharacterized genomic DNA (5). In addition to mutational pressure, translational selection is also influenced by non-random codon usage (6). To date, codon usage bias and nucleotide composition has been studied in great detail for many organisms such as bacteria (7), yeast (8), Drosophila (9, 10), and mammals (1). However, there are only a few reports on factors determining synonymous codon usage bias and nucleotide composition in viruses, especially in animal viruses.

Classical swine fever virus (CSFV) an enveloped positive-stranded RNA virus belongs to the *Pestivirus* genus of the *Flaviviridae* family. The genome contains a large open reading frame (ORF) which encodes for a unique polyprotein of about 3898 amino acids that give rise to 12 final cleavage proteins (11). Envelope protein E2 is the major envelope glycoprotein exposed on the outer surface of the virion and represents an important target for induction of the immune response during infection (12). Furthermore, the E2 gene is extensively used for evolutionary analysis (13, 14). Phylogenetic analysis indicates CSFV could be classified into 3 groups (Group 1, 2 and 3) and 10 subgroups (15). Recently, Tao *et al.* have analyzed the positive selection pressure acting on the CSFV envelope protein genes, and identified several specific codons subject to diversifying positive selection (16, 17). In order to better understand the characteristics of the

E2 gene of CSFV and to reveal more information about its evolution, we have analyzed the synonymous codon usage of E2 genes.

MATERIALS AND METHODS

Virus sequences

The available 44 complete coding sequence (CDS) of E2 gene of CSFV were downloaded from GeneBank website (http://www.ncbi.nlm.nih.gov/) and European Molecular Biology Library (EMBL) website (http://www.ebi.ac.uk/embl/). Sequences with > 99% sequence identities were excluded. All information are 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 (3). The preferred codon usage for each gene was analyzed using GCUA software package (version 1.0) (http://bioinf.may.ie/downloads.html) (18). The effective number of codons (ENC) was used to quantify the codon usage bias of each gene (19). The GC index (G+C content) was used to calculate the overall GC content in each genes, while the index GC3's (at synonymous third codon position) was used to calculate the fraction of GC nucleotides at the synonymous third codon position (excluding Met (Methionine), Trp (Tryptophan), and the termination codons) (20). The general average hydrophobicity (GRAVY) score and the frequency of aromatic amino acids (AROMO) in the hypothetical translated gene product were also computed (21).

Correspondence analysis

The relationships between variables and samples were explored using multivariate statistical analysis. Correspondence analysis (COA) was used to study the major trend in codon usage variation (22). Each dimension corresponded to the RSCU value of one sense codon (excluding AUG, UGG, and termination codons). Major trends within this dataset were determined using measures of relative inertia and genes ordered according to their positions along the axis of major inertia (23).

Table 1: Synonymous codon usage in CSFV E2 gene

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	402	1.08	Ser	UCU	50	0.40
	UUC	342	0.92		UCC	171	1.37
Leu	UUA	125	0.53		UCA	179	1.43
	UUG	**295**	**1.26**		**UCG**	**4**	**0.03**
Tyr	UAU	269	0.61	Cys	UGU	142	0.43
	UAC	**611**	**1.39**		**UGC**	**516**	**1.57**
ter	UAA	0	0.00	ter	UGA	0	0.00
ter	UAG	1	0.00	Trp	UGG	301	1.00
Leu	CUU	91	0.39	Pro	CCU	202	1.02
	CUC	178	0.76		CCC	248	1.25
	CUA	**383**	**1.63**		**CCA**	**219**	**1.11**
	CUG	**387**	**1.44**		**CCG**	**123**	**0.62**
His	CAU	55	0.32	Arg	CGU	1	0.01
	CAC	**294**	**1.68**		**CGC**	**6**	**0.04**
Gln	**CAA**	**220**	**1.44**		**CGA**	**1**	**0.01**
	CAG	85	0.56		CGG	46	0.34
Ile	AUU	137	0.65	Thr	ACU	362	0.92
	AUC	132	0.62		ACC	628	1.60
	AUA	**368**	**1.73**		**ACA**	**446**	**1.13**
Met	AUG	182	1.00		ACG	138	0.35
Asn	AAU	280	1.12	Ser	AGU	99	0.79
	AAC	218	0.88		AGC	246	1.97
Lys	AAA	322	0.72	Arg	AGA	341	2.49
	AAG	**575**	**1.28**		**AGG**	**427**	**3.12**
Val	GUU	162	0.42	Ala	GCU	204	1.06
	GUC	425	1.10		GCC	117	0.61
	GUA	344	0.89		GCA	294	1.53
	GUG	**614**	**1.59**		**GCG**	**155**	**0.81**
Asp	GAU	363	0.83	Gly	GGU	350	0.96
	GAC	511	1.17		GGC	311	0.85
Glu	GAA	450	0.94		GGA	274	0.75
	GAG	510	1.06		GGG	530	1.45

The preferentially used 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.

Statistical analysis

Correlation analysis was carried out using Spearman's rank correlation analysis method. All statistical analyses were carried out using the statistical analysis software SPSS Statistics (Version 17.0). Statistical significance was considered at $p \leq 0.05$.

Table 2: Data information of classical swine fever virus E2 genes used in this study

CSFV strains	Accession No.	ENC	GC3s	GC	GRAVY	AROMO
gi|311990282|	HQ317681	52.22	0.533	0.487	-0.165416	0.117962
gi|219964344|	FJ456876	52.68	0.537	0.492	-0.141019	0.115281
gi|219964342|	FJ456875	52.84	0.539	0.492	-0.140751	0.117962
gi|152032049|	EF683605	52.73	0.539	0.494	-0.110724	0.117962
gi|238627772|	FJ977628	53.14	0.562	0.508	-0.135389	0.115281
gi|152032077|	EF683619	52.16	0.564	0.505	-0.164611	0.117962
gi|152032082|	EF683622	51.63	0.564	0.502	-0.117426	0.117962
gi|152032059|	EF683610	51.05	0.566	0.506	-0.159786	0.117962
gi|219964322|	FJ456865	51.84	0.566	0.503	-0.142359	0.117962
gi|152032063|	EF683612	51.83	0.566	0.507	-0.139410	0.115281
gi|152032069|	EF683615	52.13	0.564	0.503	-0.152011	0.115281
gi|219964324|	FJ456866	52.08	0.565	0.505	-0.135389	0.117962
gi|152032079|	EF683620	51.45	0.568	0.504	-0.122312	0.118280
gi|152032075|	EF683618	52.24	0.564	0.501	-0.132976	0.120643
gi|219964326|	FJ456867	51.39	0.564	0.503	-0.122252	0.117962
gi|219964332|	FJ456870	52.09	0.561	0.502	-0.135657	0.117962
gi|219964334|	FJ456871	51.46	0.557	0.502	-0.151206	0.112601
gi|219964330|	FJ456869	51.41	0.564	0.502	-0.141555	0.117962
gi|152032065|	EF683613	51.84	0.555	0.499	-0.141555	0.117962
gi|219964328|	FJ456868	51.60	0.558	0.499	-0.149866	0.117962
gi|219964336|	FJ456872	52.51	0.554	0.498	-0.148526	0.117962
gi|152032073|	EF683617	53.44	0.552	0.498	-0.160322	0.117962
gi|223049419|	FJ607780	51.08	0.552	0.498	-0.152279	0.117962
gi|223049417|	FJ607779	50.48	0.552	0.496	-0.139142	0.117962
gi|221063259|	FJ582644	51.74	0.558	0.502	-0.146649	0.117962
gi|221063257|	FJ582643	50.12	0.558	0.501	-0.131904	0.117962
gi|221063263|	FJ598610	50.83	0.552	0.495	-0.186595	0.117962
gi|221063261|	FJ598609	51.61	0.552	0.497	-0.171314	0.115281
gi|221063255|	FJ582642	51.26	0.547	0.493	-0.141019	0.117962
gi|152032084|	EF683623	52.15	0.550	0.500	-0.121716	0.117962
gi|152032080|	EF683621	51.75	0.541	0.497	-0.130027	0.117962
gi|152032051|	EF683606	52.85	0.536	0.492	-0.146381	0.117962
gi|152032061|	EF683611	52.02	0.546	0.496	-0.152011	0.120643
gi|219964340|	FJ456874	52.17	0.551	0.497	-0.149062	0.117962
gi|152032057|	EF683609	52.02	0.552	0.497	-0.150402	0.117962
gi|152032055|	EF683608	52.34	0.547	0.496	-0.148526	0.117962
gi|152032053|	EF683607	52.21	0.541	0.492	-0.151475	0.117962
gi|152032067|	EF683614	52.29	0.550	0.496	-0.150402	0.117962
gi|219964338|	FJ456873	52.05	0.547	0.495	-0.127614	0.117962
gi|152032071|	EF683616	51.78	0.543	0.495	-0.151207	0.115281
gi|15283988|	AY027673	51.48	0.554	0.493	-0.128418	0.115281
gi|15283986|	AY027672	51.64	0.545	0.491	-0.100536	0.112601
gi|221063267|	FJ598612	50.49	0.523	0.485	-0.129759	0.117962
gi|221063265|	FJ598611	54.69	0.530	0.488	-0.164879	0.112601

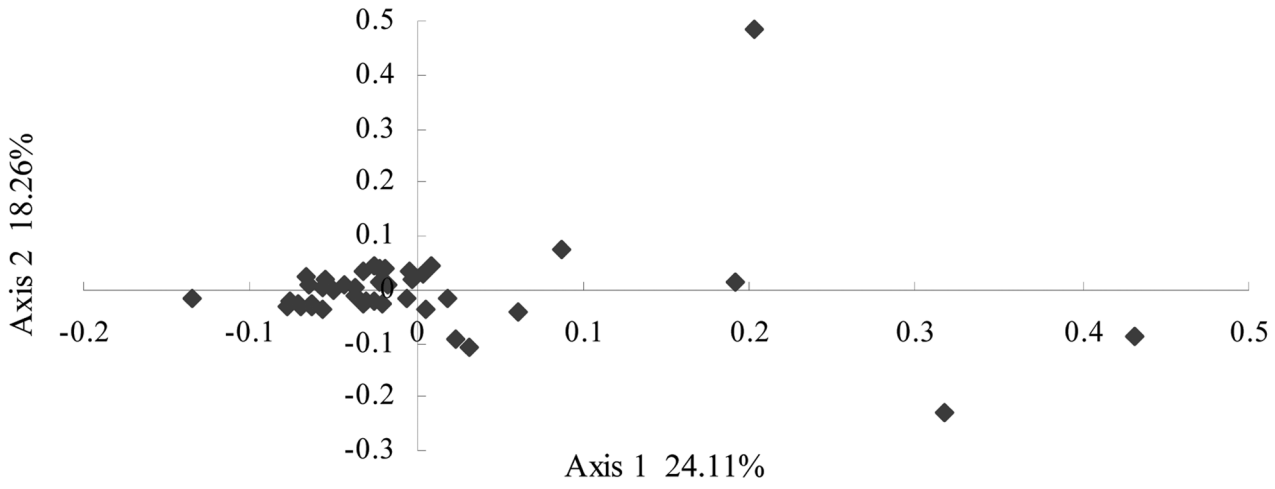

Fig 1: A plot of value of the first and second axis of each CSFV E2 gene in COA. The first axis accounts for 24.11% of all variation and the second axis accounts for 18.26% of total vibrations.

RESULTS AND DISCUSSION

Synonymous codon usage variation in E2 genes

In order to investigate the extent of codon bias in CSFV E2 genes, the (Relative Synonymous Codon Usage) RSCU values of different codon in E2 genes were calculated. The details of cumulative codon usage of 59 codons in 44 CSFV E2 genes are displayed in Table 1. The preferentially used codons were A-ended, C-ended, and G-ended codons. It was interesting to note that no U-ended codons were used as preferential codons. Effective number of 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 was used for each amino acid, the ENC value equalled 20. Conversely, in a gene exhibiting no bias, the value was 61 (19). Our data showed that the ENC values of different CSFV genes vary from 50.12 to 54.69, with a mean of 51.93 and S.D. of 0.8031, which indicated that the codon usage bias in CSFV E2 genes was small. Moreover, GC and GC3s values were calculated and are listed in Table 2. The average GC content of CSFV E2 genes was 0.4978 (mean values varying from 0.485 to 0.508, with a S.D. of 0.0054), while average GC3s content was 0.552 (mean values varying from 0.523 to 0.568, with a S.D. of 0.011). This results are consistent with previous observations that CSFV are not GC-poor genomes (17).

Correspondence analysis of codon usage

To investigate synonymous codon usage variation, Correspondence analysis (COA) was studied for 44 CSFV E2 genes selected for this study. Figure 1 depicts the position of each E2 gene on the plane defined by the first and second principal axes generated by COA on RSCU values. The first principal axis accounted for 24.11% of the total variation, and the next three axes accounted for 18.26%, 13.45%, and 11.34% of the variation, respectively. This observation indicated 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.

Mutational bias as 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 (24). In order to investigate whether codon usage variation of different genes is determined by mutational bias, correlation analysis was employed to correlate the first two axes of COA with codon usage indices. Correlation analysis showed that axis 1 of COA and axis 2 were both correlated with GC (r = -0.62, P<0.001), GC3s (r = -0.653, P<0.001), GC (r = -0.312, P<0.05), GC3s (r = -0.307, P<0.05), respectively, which indicated that the patterns of base composi-

tion were most likely the results of mutational pressure, and not natural selection, since the effects were present at all codon positions.

Moreover, ENC-plot (ENC plotted against GC3's) was used as part of a general strategy to investigate patterns of synonymous codon usage (22). Genes, whose codon choice is constrained only by a G + C mutation bias, should lie on or just below the curve of the predicted values (5, 19). All of the spots were located below the expected curve as in Figure 2, indicating that the codon usage bias in these 44 CSFV E2 genes was greatly influenced by the GC compositional constraints. In addition, a significantly negative correlation (r = -0.327, 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 (caused by mutation bias) is the main determinant of the variation in synonymous codon usage.

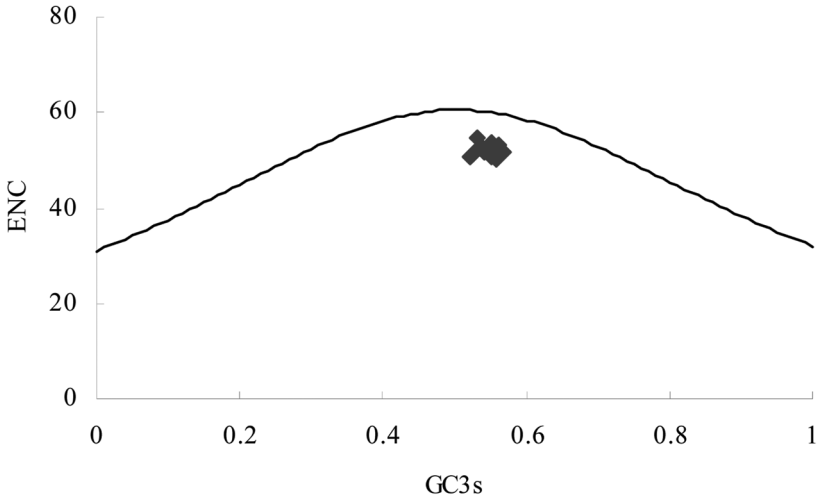

Fig 2:. Effective number of codons used in each gene plotted against the GC3s. The continuous curve plots the relationship between GC3's and ENC in the absence of selection. All of spots lie below the expected curve.

Aromaticity and hydrophobicity affect codon usage

To test whether selection pressure contributes to the codon usage variation among E2 genes, we performed a correlation analysis to evaluate whether GRAVY and AROMO values were related to first two axes of COA, ENC and GC3s (25). Our results showed that only AROMO was correlated with axis 1 (r =-0.306, P<0.05) (Table 3), while GRAVY was not correlated with two axes, ENC and GC3s. The results indicated that the degree of hydrophobicity was not associated with condon usage variation, whereas, the aromatic amino acids (Phe, Tyr, Trp) were associated with the codon usage variation to some extent.

Table 3: Summary of correlation analysis between GRAVY, AROMO, ENC, and the first two axes in COA

		Axis 1	Axis2	ENC	GC3s
GRAVY	*r*	-0.257	-0.135	-0.133	0.098
	P	0.092	0.381	0.389	0.528
AROMO	*r*	-0.306	0.050	-0.63	0.13
	P	0.043*	0.745	0.684	0.402

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

CONCLUSION

Synonymous codon usage biases in 44 CSFV E2 genes were analyzed, and the results showed that CSFV E2 genes had low codon usage bias. Mutational pressure is the main factor determining the codon usage biases. In addition, aromaticity could partially account for the codon usage variation.

ACKNOWLEDGEMENT

The study was supported by the Technology Research Foundation of Education Department of HeiLongJiang Province, China (12511352).

REFERENCES

1. Marin, A., Bertranpetit, J., Oliver, J.L. and Medina, J.R.: Variation in G+C content and codon choice: differences among synonymous codon groups in vertebrate genes. Nucleic Acids Res. 17: 6181-6189, 1989.
2. Aota, S. and Ikemura, T.: Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. 14: 6345-6355, 1986.
3. Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R.: Codon usage in yeast cluster-analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14: 5125-5143, 1986.

4. Bulmer, M.: The selection-mutation-drift theory of synonymous codon usage. Genetics. 129: 897-907, 1991.
5. Zhong, J.C., Li, Y.M., Zhao, S., Liu, S.G. and Zhang, Z.D.: Mutation pressure shapes codon usage in the GC-Rich genome of foot-and-mouth disease virus. Virus Genes. 35: 767-776, 2007.
6. Zhou, T., Gu, W.J., Ma, J.M., Sun, X. and Lu, Z.H.: Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. BioSystems. 81: 77-86, 2005.
7. Wright, F. and Bibb, M.J.: Codon usage in the G+C rich *Streptomyces* genome. Gene 113: 55-65, 1992.
8. Sharp, P.M. and Lloyd, A.T.: Regional base composition variation along yeast chromosome III evolution of chromosome primary structure. Nucleic Acids Res. 21: 179-183, 1993.
9. Rubin, G.M.: The *Drosophila* genome project: a progress report. Trends in Genetics. 14: 340-343, 1998.
10. Hiroshi, A., Richard, M.K. and Adam, E.W.: Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. Genetica. 102: 49-60, 1998.
11. Fan, Y.F., Zhao, Q., Zhao, Y., Wang, Q., Ning, Y.B. and Zhang, Z.Q.: Complete genome sequence of attenuated low-temperature Thiverval strain of classical swine fever virus. Virus Genes. 36: 531-538, 2008.
12. Montesino, R., Toledo, J.R., Sanchez, B., Zamora, Y., Barrera, M., Royle, L., Rudd, P.M., Dwek, R.A., Harvey, D.J. and Cremata, J.A.: N-Glycosylation Pattern of E2 Glycoprotein from classical swine fever virus. J. Proteo. Res. 8: 546-555, 2009.
13. Paton, D.J., McGoldrick, A., Greiser-Wilke, I., Parchariyanon, S., Song, J.Y., Liou, P.P., Stadejek, T., Lowings, J.P., Bjorklund, H. and Belak, S.: Genetic typing of classical swine fever virus. Vet. Micr. 73: 137-157, 2000.
14. Chen, N., Hu, H.X., Zhang, Z.F., Shuai, J.B., Jiang, L.L. and Fang, W.H.: Genetic diversity of the envelope glycoprotein E2 of classical swine fever virus: Recent isolates branched away from historical and vaccine strains. Vet. Micr. 127: 286-299, 2008.
15. Paton, D.J. and Greiser-Wilke, I.: Classical swine fever - an update. Res.Vet. Sci. 75: 169-178, 2003.
16. Zhang, H., Wang, Y.H., Cao, H.W. and Cui, Y.D.: Phylogenetic analysis of E2 genes of classical swine fever virus in China. Isr. J. Vet. Med. 65: 151-155, 2010.
17. Tao, P., Dai, L., Luo, M.C., Tang, F.Q., Tien, P. and Pan, Z.S.: Analysis of synonymous codon usage in classical swine fever virus. Virus Genes. 38: 104-112, 2009.
18. Sharp, P.M. and Li, W.H.: Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for rare codons. Nucleic Acids Res. 14: 7737-7749, 1986.
19. Wright, F.: The effective number of codons used in a gene. Gene. 87: 23-29, 1990.
20. Richard, J.E., Lin, K. and Tan, T.: A functional significance for codon third bases. Gene. 245: 291-298, 2000.
21. Kyte, J. and Doolittle, R.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157: 105-132, 1982.
22. Gupta, S.K. and Ghosh, T.C.: Expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. Gene. 273: 63-70, 2001.
23. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R.: Codon catalogue usage is a genome strategy for genome expressivity. Nucleic Acids Res. 9: 43-75, 1981.
24. Gareth, M.J. and Edward, C.H.: The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92: 1-7, 2003.
25. Lobry, J.R. and Gautier, C.: Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 *Escherichia coli* chromosome encoded genes. Nucleic Acid Res. 22: 3174-3180, 1994.