Available online at www.sciencedirect.com

SCIENCE DIRECT®

Virus Research 101 (2004) 155–161

Virus Research

www.elsevier.com/locate/virusres

Analysis of synonymous codon usage in SARS *Coronavirus* and other viruses in the *Nidovirales*

Wanjun Gu, Tong Zhou, Jianmin Ma, Xiao Sun, Zuhong Lu*

Key Laboratory of Molecular and Biomolecular Electronics, Southeast University, Ministry of Education, Nanjing, Jiangsu 210096, China

Received 25 October 2003; received in revised form 9 January 2004; accepted 9 January 2004

Abstract

In this study, we calculated the codon usage bias in severe acute respiratory syndrome *Coronavirus* (SARSCoV) and performed a comparative analysis of synonymous codon usage patterns in SARSCoV and 10 other evolutionary related viruses in the *Nidovirales*. Although there is a significant variation in codon usage bias among different SARSCoV genes, codon usage bias in SARSCoV is a little slight, which is mainly determined by the base compositions on the third codon position. By comparing synonymous codon usage patterns in different viruses, we observed that synonymous codon usage pattern in these virus genes was virus specific and phylogenetically conserved, but it was not host specific. Phylogenetic analysis based on codon usage pattern suggested that SARSCoV was diverged far from all three known groups of *Coronavirus*. Compositional constraints could explain most of the variation of synonymous codon usage among these virus genes, while gene function is also correlated to synonymous codon usages to a certain extent. However, translational selection and gene length have no effect on the variations of synonymous codon usage in these virus genes.

Keywords: Synonymous codon usage; Mutational bias; Selection pressure; Base composition; SARS; *Coronavirus*

1. Introduction

Synonymous codons are not used equally both within and between genomes (Grantham et al., 1980; Martin et al., 1989; Lloyd and Sharp, 1992). Compositional constraints and natural selection are thought to be the two main factors accounting for codon usage variation among genes in different organisms (Karlin and Mrazek, 1996; Sharp et al., 1986; Lesnik et al., 2000). The diverse patterns of codon usage in mammals may arise from compositional constraints of the genomes (Karlin and Mrazek, 1996; Francino and Ochman, 1999; Majumdar et al., 1999; Ghosh et al., 2000). In contrast, in some unicellular organisms, such as *Escherichia coli* and *Saccharomyces cerevisiae*, high expressed genes have a strong selective preference for codons with a high concentration of the corresponding acceptor tRNA molecule, whereas low expressed genes displayed a more uniform pattern of codon usage (Gouy and Gautier, 1982; Grantham et al., 1981; Ikemura, 1981, 1985; Sharp et al., 1986; Lesnik et al., 2000). Moreover, mutational pressure rather than translational selection is the most important determinant of the codon bias in some human RNA viruses (Levin and Whittome, 2000; Jenkins et al., 2001; Jenkins and Holmes, 2003). Furthermore, replicational and transcriptional selection is responsible for the codon usage variation among the genes of *Borrelia burgdorferi* (McInerney, 1998). In some other researches, codon usage was also found to be related to gene function (Chiapello et al., 1998; Epstein et al., 2000; Ma et al., 2002), protein secondary structure (Chiusano et al., 1999, 2000; Oresic and Shalloway, 1998; Xie and Ding, 1998; Gupta et al., 2000), cellular location of gene products (Chiapello et al., 1999) and gene length (Coghlan and Wolfe, 2000; Marais and Duret, 2001; Moriyama and Powell, 1998).

Severe acute respiratory syndrome (SARS) is a respiratory disease that was recently reported in Asia, North America and Europe (Chan-Yeung and Yu, 2003; Drazen,

Abbreviations: bp, base pair; SARSCoV, severe acute respiratory syndrome *Coronavirus*; RSCU, relative synonymous codon usage; ENC, effective number of codons; CA, correspondence analysis; GC_{3S}, the frequency of G+C at the synonymous third position of sense codons; A_{3S}, T_{3S}, G_{3S} and C_{3S}, the adenine, thymine, guanine and cytosine content at synonymous third positions; ORF, open reading frame; PCR, polymerase chain reaction; S.D., standard deviation

* Corresponding author. Tel.: +86-25-83619983; fax: +86-25-83619983.
E-mail address: zhlu@seu.edu.cn (Z. Lu).

0168-1702/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.virusres.2004.01.006

Table 1
Identified ORFs (length > 150 bps) in the SARSCoV (TOR2 isolation) genome[a,b]

Gene product	L[a]	ENC	GC_{3S} (%)	f'_1[b]
Putative orf1ab polyprotein	21222	48.47	32.20	−0.60
Orf1a polyprotein	13149	48.24	33.10	−0.57
Putative spike glycoprotein	3468	45.73	28.30	−0.85
Putative uncharacterized protein	825	47.66	34.50	−0.37
Putative uncharacterized protein	465	42.80	45.10	1.34
Putative small envelope protein E	231	59.06	38.70	0.34
Putative protein M	666	59.04	42.50	0.51
Putative uncharacterized protein	192	42.19	28.80	−1.08
Putative uncharacterized protein	269	43.05	30.60	−0.55
Putative nucleocapsid protein	1269	54.16	37.60	0.49
Putative uncharacterized protein	297	46.62	58.10	1.87

[a] L represents the length of identified ORF.
[b] f'_1 represent the first axis values of each gene in CA.

2003). Although genome sequence of severe acute respiratory syndrome *Coronavirus* (SARSCoV) has been published and many studies have been performed on SARSCoV in recent months (Paul et al., 2003; Qin et al., 2003; Marra et al., 2003; Snijder et al., 2003), little genomic analysis is available on this virus. Codon usage data of SARSCoV might give some clues to the features of SARSCoV genome and some evolutionary information of this virus. Here, we analyzed the codon usage data of this virus and other viruses in the order *Nidovirales*. The key evolutionary determinants of codon usage bias in these viruses were also investigated.

2. Materials and methods

2.1. Materials

SARSCoV is a large, enveloped, positive-stranded RNA virus, which belongs to order *Nidovirales*, family Coronaviridae, genus *Coronavirus* in virus taxonomy (Marra et al., 2003). The complete genome and coding sequences of SARSCoV TOR2 isolation were obtained from GenBank (Version 134.0). To keep the statistical significance of codon usage bias, only sequences with length above 150 bps were analyzed (Table 1). To compare the codon usage pattern among different viruses, coding genes of 10 other viruses belonging to order *Nidovirales* (six viruses in the genus *Coronavirus*, four viruses in the genus *Arterivirus*) were also parsed from GenBank (Version 134.0) (Table 2).

2.2. Methods

2.2.1. Synonymous codon usage measures (RSCU)

Relative synonymous codon usage values of each codon in a gene were used to examine the synonymous codon usage without the confounding influence of amino acid composition (Sharp and Li, 1986). N_{3S}, the frequency of base N at synonymous third codon positions, was also used to calculate the extent of base composition bias. Additionally, the effective number of codons of a gene (ENC) was used to quantify the codon usage bias of a gene (Wright, 1990), which is the best overall estimator of absolute synonymous codon usage bias (Comeron and Aguade, 1998). ENC value ranges from 20 (when only one codon is used per amino acid) to 61 (when all synonymous codons are equally used for each amino acid).

Table 2
Phylogenetic breakdown, accession number, GC_{3S} and the first two axis values in CA of 11 selected viruses in order *Nidovirales*[a,b]

Organism[a]	Accession number	GC_{3S}(%)	f'_1[b]	f'_2[b]
Coronavirus				
HCoV 229E	NC_002645	30.89	−0.84	−0.16
PEDV	NC_003436	37.32	−0.04	0.42
TGV	NC_002306	27.02	−0.99	−0.08
BCoV	NC_003045	29.43	−0.75	0.48
MHV	NC_001846	38.30	−0.16	0.27
AIBV	NC_001451	26.09	−0.90	−1.30
SARSCoV	NC_004718	37.23	0.05	0.36
Arterivirus				
EAV	NC_002532	47.28	0.80	0.47
LDEV	NC_002534	45.18	0.53	0.43
PRRSV	NC_001961	53.76	1.31	0.55
SHFV	NC_003092	48.43	1.09	−0.14

[a] *Organism abbreviation*: HCoV 229E, human *Coronavirus* 229E; PEDV, porcine epidemic diarrhea virus; TGV, transmissible gastroenteritis virus; BCoV, bovine *Coronavirus*; MHV, murine hepatitis virus; AIBV, avian infectious bronchitis virus; SARSCoV, SARS *Coronavirus*; EAV, equine arteritis virus; LDEV, lactate dehydrogenase elevating virus; PRRSV, porcine reproductive and respiratory syndrome virus; SHFV, simian hemorrhagic fever virus.
[b] f'_1 and f'_2, respectively, represent the first axis mean value and the second axis mean value in CA of each genome.

2.2.2. Correspondence analysis (CA)

Correspondence analysis was used to investigate the major trend in codon usage variation among genes. Each gene is represented as a 59 dimensional vector, and each dimension corresponds to the RSCU value of one sense codon (excluding AUG, UGG and three stop codons).

CA based on RSCU values relies on two main steps (Mardia et al., 1979). The first step is to measure the similarities in codon usage using the squared Euclidean distance among all genes, and the resulting distance table will be used to compute the coordinates of the genes in a multidimensional space. The second step provides the visualization of these Euclidean distances through positioning genes by successive orthogonal projections of the cloud of points. Essentially, this process consists in finding the linear transformations $f'_1, f'_2, \ldots, f'_{58}$ of the original variables $f'_1, f'_2, \ldots, f'_{59}$. The f'-variables are calculated and ordered according to the values of relative variance. f'_1 is the maximum value; f'_2 is the next value and is by construction not correlated with f'_1. The same applies to f'_3, f'_4, and so on, until f'_{58}. So, genes with similar codon usage are neighbors on the components of projection.

2.2.3. Statistical methods

Linear regression analysis was used to find the correlation between codon usage bias and nucleotide composition. One tailed t-test was used to compare the variation of codon usage between different gene groups (Ewens and Grant, 2001). As a null hypothesis, it is assumed that mean values of codon usage indices in different gene groups is statistically the same. Under the null assumption, t-statistic could be calculated. Then, P-value is derived and it is taken as significance when P-value is below 0.05.

A C++ program was developed to calculate the codon usage indices for each gene. CA and other statistical analysis were performed with statistical software SPSS 11.0.

3. Results

3.1. Synonymous codon usage in SARSCoV

The details of coding genes in SARSCoV and the overall RSCU values of 61 sense codons in SARSCoV were, respectively, shown in Tables 1 and 3. All preferentially used codons in SARSCoV are all A-ended or U-ended codons (Table 3). SARSCoV is a GC poor genome with GC content of 37.52%. Due to compositional constraints, it is expected that A-ended and/or U-ended codons should be preferentially used in this genome. To study the codon usage variation among different SARSCoV genes, ENC and GC$_{3S}$ values of different SARSCoV genes were calculated (Table 1). ENC values of different SARSCoV genes vary from 42.19 to 59.06, with a mean value of 48.99 and S.D. of 6.41. Because all ENC values of SARSCoV genes are much higher (ENC $>$ 40), codon usage bias in SARSCoV genome is a little slight. However, there is a marked variation in codon usage pattern among different SARSCoV genes (S.D. = 6.41). Similarly, GC$_{3S}$ values of each SARSCoV gene also confirm the heterogeneity of synonymous codon usage among different SARSCoV genes, which range from 28.3 to 58.1% with a mean of 37.23 and S.D. of 8.78%.

3.2. Synonymous codon usage in different viruses is virus specific, but not host specific

CA was implemented for all identified ORFs from each of the 11 virus genomes as a single dataset, which consists of 103 coding sequences. CA detected one major trend in the first axis which accounted for 15.40% of the total variation, and none of the other axes individually accounted for more than 7.60% of the total variation. A plot of the first axis and the second axis of each gene was shown in Fig. 1. Although this graph is a little complex with some overlap among genes from different genomes, it is clear that genes from a particular genome tend to cluster together. The separation of one virus genome from other virus genomes is determined to be significant on both axes (t-test, P-value $<10^{-15}$ on the first axis and P-value $<10^{-3}$ on the second axis). So, similar to codon usage in mammals and bacteria, synonymous codon usage in these viruses is also virus specific.

Table 3
Synonymous codon usage in SARSCoV[a,b,c]

AA[a]	Codon	RSCU	*N*[b]	AA[a]	Codon	RSCU	*N*[b]
Ala	**GCU**	**2.08**	531	Ile	**AUU**	**1.72**	410
	GCC	0.58	147		AUC	0.67	159
	GCA	1.13	288		AUA	0.62	148
	GCG	0.22	55	Cys	**UGU**	**1.27**	280
Gly	GGG	0.17	37		UGC	0.73	160
	GGA	0.85	182	Thr	**ACU**	**1.66**	427
	GGC	0.95	202		ACC	0.59	153
	GGU	**2.02**	431		ACG	0.18	46
Val	**GUU**	**1.71**	479		ACA	1.57	406
	GUC	0.67	188	Asn	**AAU**	**1.24**	449
	GUA	0.83	232		AAC	0.76	277
	GUG	0.78	219	Gln	**CAA**	**1.16**	298
Leu	UUA	1.04	238		CAG	0.84	214
	UUG	1.10	251	Tyr	**UAU**	**1.12**	345
	CUU	**1.79**	409		UAC	0.88	270
	CUC	0.83	191	His	**CAU**	**1.29**	187
	CUA	0.64	147		CAC	0.71	103
	CUG	0.60	138	Asp	**GAU**	**1.24**	463
Phe	UUC	0.77	260		GAC	0.76	282
	UUU	**1.23**	414	Glu	**GAA**	**1.04**	354
Pro	**CCU**	**1.74**	247		GAG	0.96	326
	CCC	0.40	57	Lys	**AAA**	**1.04**	421
	CCA	1.70	241		AAG	0.96	388
	CCG	0.16	22	Arg	CGU	1.77	153
Ser	**UCU**	**1.96**	310		CGC	0.72	62
	UCC	0.42	67		CGA	0.44	38
	UCA	1.70	270		CGG	0.09	8
	UCG	0.23	36		**AGA**	**2.08**	180
	AGU	1.17	186		AGG	0.90	78
	AGC	0.52	82				

[a] AA is the abbreviation of amino acid.
[b] *N* represents the number of occurrence of each sense codon.
[c] The preferentially used codons for each amino acid are displayed in bold.

To show whether there is a correlation between virus codon usage and its host, these 103 virus genes were divided into several groups according to the virus host. For example, because both SARSCoV genes and human *Coronavirus* 229E infect human, genes in these two viruses were incorporated as a group. Next, t-test was also used to test whether the separation of different viral genes which infect different hosts is significant. The P-value is 0.57 on the first axis and is 0.08 on the second axis, which suggested that codon usage in different virus genes was not host specific.

3.3. Phylogenetic analysis of these viruses based on codon usage pattern

In Fig. 1, all virus genes in the genus *Coronavirus* were plotted in red. At the same time, all viral genes in the genus *Arterivirus* were plotted in blue. *Coronavirus* genes are mainly located on the left side of the plot, while a majority

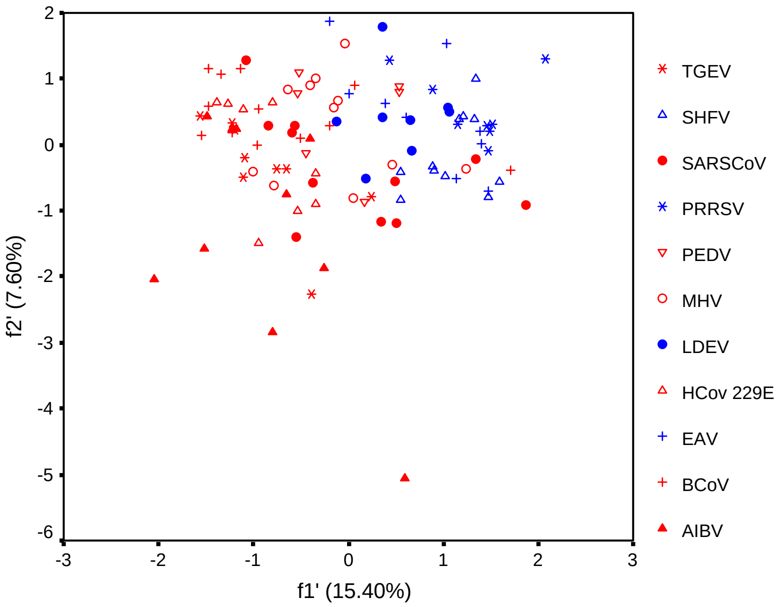

Fig. 1. A plot of the values of the first axis and the second axis of each gene in CA (abbreviations of the viruses: AIBV, avian infectious bronchitis virus; BCoV, bovine *Coronavirus*; EAV, equine arteritis virus; HCoV 229E, human *Coronavirus* 229E; LDEV, lactate dehydrogenase elevating virus; MHV, murine hepatitis virus; PEDV, porcine epidemic diarrhea virus; PRRSV, porcine reproductive and respiratory syndrome virus; SARSCoV, SARS *Coronavirus*; SHFV, simian hemorrhagic fever virus; TGV, transmissible gastroenteritis virus. f'_1 and f'_2, respectively, represent the values of the first and the second axis of each gene in CA).

of *Arterivirus* genes are located on the right side. The separation of *Coronavirus* genes and *Arterivirus* genes on the first axis is statistically significant (*t*-test, *P*-value $<10^{-15}$). Hence, synonymous codon usage appears to be conserved between phylogenetically related viruses.

Also, SARSCoV genes were widely extended in the first axis (Fig. 1). Six of eleven SARSCoV genes were located in the cluster of *Coronavirus* genes, while the other five SARSCoV genes were located in the cluster of *Arterivirus* genes. Therefore, SARSCoV might have been diverged far from all three known *Coronavirus* groups. Comparing with all other viruses in the genus *Coronavirus*, it might be more evolutionary related to the genus *Arterivirus*.

3.4. Mutational bias is the main factor determines the codon usage variation among different viruses

Linear regression analysis was implemented to find whether there is some correlation between synonymous codon usage bias and nucleotide compositions. The R^2 value and significance level of these regression analyses was listed in Table 4. The first axis value of each gene in CA is closely correlated with all the base compositions on the third codon position, while the second axis of each gene is correlated with some base compositions on the third codon position to a certain extent. Therefore, compositional constraint mainly determines the variation of synonymous codon usage among these virus genes.

Furthermore, we plotted the first axis values in CA and GC_{3S} values of each gene (Fig. 2). The GC_{3S} mean value of genes in coronaviruses ranges from 26.09 to 37.32, and it ranges from 45.18 to 53.76 in arteriviruses (Table 2). Although codon usage bias appears to be conserved between evolutionary related viruses (Section 3.3), the patterns of codon usage in different virus genes also appear to be a direct function of the GC content on the third codon position of these genes.

3.5. Gene function also drives the codon usage variation among different viruses

The plot of ENC and GC_{3S} is another effective way to explore codon usage variation among genes (Wright, 1990). ENC values of each virus gene were plotted against its

Table 4
Summary of linear regression analysis between the first two axes in CA and the nucleotide contents on the third codon position in all selected virus genes[a]

Base composition	f'^{b}_1	f'^{b}_2
A_{3S}	0.791^{****}	0.085^{*}
T_{3S}	0.239^{****}	0.444^{****}
G_{3S}	0.484^{****}	0.082^{*}
C_{3S}	0.720^{****}	0.0001^{NS}
GC_{3S}	0.936^{****}	0.018^{NS}

NS in superscript represent non-significant.

[a] Value in this table is the R^2 value of each linear regression analysis.

[b] f'_1 and f'_2, respectively, represent the values of the first and the second axis of each gene in CA.

* *P*-value <0.01.

**** *P*-value <0.00001.

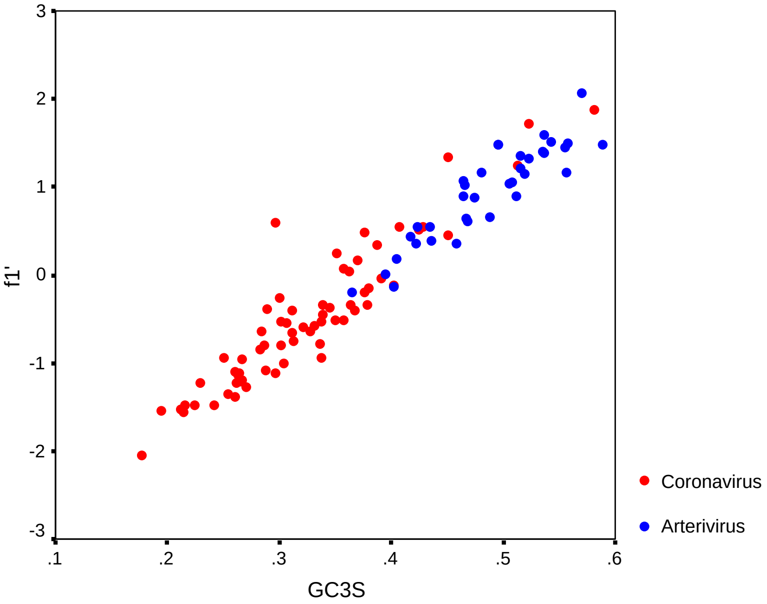

Fig. 2. A dot plot of the first axis value in correspondence analysis and GC_{3S} of each gene (f_1' denotes the first axis value in correspondence analysis of each gene, and GC_{3S} denotes the G + C content on the third synonymous codon position of each gene).

corresponding GC_{3S} (Fig. 3). The solid line represents the curve if codon usage is only determined by GC content on the third codon position. A large proportion of points lie near to the solid line on the left region of this distribution. It also suggests that mutational bias is the main factor determines the codon usage variation among these genes. However, there are also some points lying below the expected curve. Hence, other than mutational bias, there might be some additional factors drive the codon usage variation among these genes.

To show whether translational selection or gene function were correlated with the observed variation in codon bias, all virus genes were grouped into several classes according to gene function. Because most of these viruses contain genes coding for RNA polymerase, envelop protein and structural glycoprotein, these three gene groups were selected to find whether there is some correlation between codon usage and gene function. One tailed *t*-test was then performed on ENC values of these genes with the hypothesis that there is no correlation between codon usage bias and gene function. Some associations have been found. Average codon usage bias is higher in RNA polymerase gene group than in envelop gene group (*t*-test, *P*-value = 0.031), and it is higher in polymerase gene group than in structural glycoprotein gene group (*t*-test, *P*-value = 0.002). But, there is no association between codon usage in structural glycoprotein gene group and envelop protein gene group (*t*-test, *P*-value = 0.74). Because the structural glycoprotein and envelop protein are all structural proteins in these viruses and RNA polymerase is a nonstructural protein, it is clear that codon usage in structural genes is significantly diverged from that in nonstructural genes. On the other hand, structural genes are generally highly expressed than nonstructural genes. So, if translational selection was also contributed to codon usage bias in these genes, codon usage bias in structural genes should be higher than in RNA polymerase genes. However, RNA polymerase genes (ENC = 49.25) were found to have greater codon usage bias than structural genes (ENC = 54.60 for envelop gene and ENC = 55.33 for structural glycoprotein). Hence, codon usage bias in these virus genes is not related to gene expression level. Furthermore, we also performed a linear regression analysis on ENC value and gene length of each gene. But, there was no significant correlation between codon usage and gene length in these virus genes (*P*-value > 0.05). So, gene function, rather than translational selection

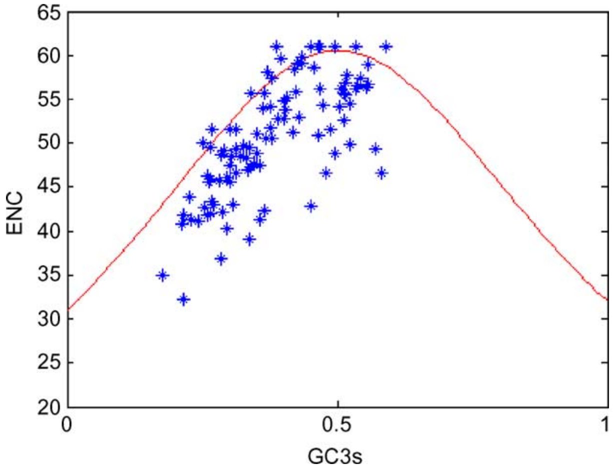

Fig. 3. ENC vs. GC_{3S} plot of all virus genes (ENC denotes the effective number of codon of each gene, and GC_{3S} denotes the G + C content on the third synonymous codon position of each gene. The solid line represents the relationship between GG3S and ENC under random codon usage assumption).

and gene length, is another factor accounting for codon usage variation among these virus genes.

4. Discussion

Our analysis revealed that synonymous codon usage bias in SARSCoV was less biased, which was mainly determined by the base compositions on the third codon position. Comparative analysis of codon usage bias in the order *Nidovirales* also suggested that codon usage in these viruses was virus specific and mutational bias was the main factor drives the codon usage variation among these viruses. Gene function was also related to codon usage bias in these viruses to some extent. But, translational selection and gene length might have no effect on the codon usage pattern in these viruses. Some published results has shown that the overall extent of codon usage bias in RNA viruses is low and there is little variation in bias between genes (Levin and Whittome, 2000; Jenkins et al., 2001; Jenkins and Holmes, 2003). Although SARSCoV is a newly detected RNA virus infecting human, the synonymous codon usage pattern in SARSCoV we described here is also in accordance with these published codon usage pattern of human RNA viruses (Jenkins and Holmes, 2003). Because mutation rates in RNA viruses are much higher than those in DNA viruses (Drake and Holland, 1999), it is understandable that mutation pressure is the main determinant of codon usage bias in SARSCoV. Our analysis also revealed that there was no host specific codon usage pattern in these viruses. So, host genome might have no obvious effect on the evolution of these viruses.

Some phylogenetic analysis of SARSCoV (Qin et al., 2003; Marra et al., 2003) has shown that SARSCoV does not closely resemble any of the three previously known groups in genus *Coronavirus*. But Snijder et al. (2003) has proposed that SARSCoV is most closely related to group 2 *Coronavirus*es. Based on different codon usage patterns in different coronaviruses, we revealed that codon usage patterns of each virus was phylogenetically distinct and SARSCoV might have been diverged far from all three known *Coronavirus* groups, which is in accordance with the results Qin et al. (2003) and Marra et al. (2003) proposed.

Codon usage patterns and the phylogenetic results we proposed here are useful to understand the processes governing the evolution of SARSCoV, especially the roles played by mutation pressure and natural selection. Further, such information might be helpful to understand the pathogenesis and the origin of SARSCoV.

Acknowledgements

This research is a part of Projects 60121101 and 60223002 supported by National Natural Science Foundation of China.

References

Chan-Yeung, M., Yu, W.C., 2003. Outbreak of severe acute respiratory syndrome in Hong Kong Special Administrative Region: case report. Brit. Med. J. 326, 850–852.

Chiapello, H., Lisacek, F., Caboche, M., Henaut, A., 1998. Codon usage and gene function are related in sequences of Arabidopsis thaliana. Gene 209, GC1–GC38.

Chiapello, H., Ollivier, E., Landes-Devauchelle, C., Nitschke, P., Risler, J.L., 1999. Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases. Nucleic Acids Res. 27, 2848–2851.

Chiusano, M.L., D'Onofrio, G., Alvarez-Valin, F., Jabbari, K., Colonna, G., Bernardi, G., 1999. Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. Gene 238, 23–31.

Chiusano, M.L., Alvarez-Valin, F., Di Giulio, M., D'Onofrio, G., Ammirato, G., Colonna, G., Bernardi, G., 2000. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. Gene 261, 63–69.

Coghlan, A., Wolfe, K.H., 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast 16, 1131–1145.

Comeron, J.M., Aguade, M., 1998. An evaluation of measures of synonymous codon usage bias. J. Mol. Evol. 47, 268–274.

Drake, J.W., Holland, J.J., 1999. Mutation rates among RNA viruses. Proc. Natl. Acad. Sci. U.S.A. 96, 13910–13913.

Drazen, J.M., 2003. Case clusters of the severe acute respiratory syndrome. New Engl. J. Med. 348, e6–e7.

Epstein, R.J., Lin, K., Tan, T.W., 2000. A functional significance for codon third bases. Gene 245, 291–298.

Ewens, W.J., Grant, G.R., 2001. Statistical Methods in Bioinformatics. Springer, New York.

Francino, H.P., Ochman, H., 1999. Isochores result from mutation not selection. Nature 400, 30–31.

Ghosh, T.C., Gupta, S.K., Majumdar, S., 2000. Studies on codon usage in Entamoeba histolytica. Int. J. Parasitol. 30, 715–722.

Gouy, M., Gautier, C., 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10, 7055–7074.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pave, A., 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, r49–r62.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., Mercier, R., 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9, r43–r74.

Gupta, S.K., Majumdar, S., Bhattacharya, T.K., Ghosh, T.C., 2000. Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochem. Biophys. Res. Commun. 269, 692–696.

Ikemura, T., 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. 151, 389–409.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol 2, 13–34.

Jenkins, G.M., Pagel, M., Gould, E.A., Zanotto, P.M.d.A., Holmes, E.C., 2001. Evolution of base composition and codon usage bias in the genus Flavivirus. J. Mol. Evol 52, 383–390.

Jenkins, G.M., Holmes, E.C., 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92, 1–7.

Karlin, S., Mrazek, J., 1996. What drives codon choices in human genes? J. Mol. Biol. 262, 459–472.

Lesnik, T., Solomovici, J., Deana, A., Ehrlich, R., Reiss, C., 2000. Ribosome traffic in *E. coli* and regulation of gene expression. J. Theor. Biol. 202, 175–185.

Levin, D.B., Whittome, B., 2000. Codon usage in nucleopolyhedroviruses. J. Gen. Virol. 81, 2313–2325.

Lloyd, A.T., Sharp, P.M., 1992. Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. Nucleic Acids Res. 20, 5289–5295.

Ma, J.M., Zhou, T., Gu, W.J., Sun, X., Lu, Z.H., 2002. Cluster analysis of the codon use frequency of MHC genes from different species. Biosystems 65, 199–207.

Majumdar, S., Gupta, S.K., Sundararaj, V.S., Ghosh, T.C., 1999. Compositional correlation studies among the three different codon positions in 12 bacterial genomes. Biochem. Biophys. Res. Commun. 266, 66–71.

Marais, G., Duret, L., 2001. Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans. J. Mol. Evol. 52, 275–280.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate analysis. Academic press, New York.

Marra, M.A., Jones, S.J., Astell, C.R., 2003. The genome sequence of the SARS-associated *Coronavirus*. Science 300, 1399–1404.

Martin, A., Bertranpetit, J., Oliver, J.L., Medina, J.R., 1989. Variation in G + C content and codon choice: differences among synonymous codon groups in vertebrate genes. Nucleic Acids Res. 17, 6181–6189.

McInerney, J.O., 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. U.S.A. 95, 10698–10703.

Moriyama, E.N., Powell, J.R., 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res. 26, 3188–3193.

Oresic, M., Shalloway, D., 1998. Specific correlations between relative synonymous codon usage and protein secondary structure. J. Mol. Biol. 281, 31–48.

Paul, A.R., Steven, O.M., Stephan, S.M., 2003. Characterization of a novel *Coronavirus* associated with severe acute respiratory syndrome. Science 300, 1394–1399.

Qin, E.D., Zhu, Q.Y., Yu, M., 2003. A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). Chin. Sci. Bull. 48, 941–948.

Sharp, P.M., Li, W.H., 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for ‘rare’ codons. Nucleic Acids Res. 14, 7737–7749.

Sharp, P.M., Tuohy, T.M., Mosurski, K.R., 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14, 5125–5143.

Snijder, E.J., Bredenbeek, P.J., Dobbe, J.C., 2003. Unique and conserved features of genome and proteome of SARS-*Coronavirus*, an early split-off from the *Coronavirus* group 2 lineage. J. Mol. Biol. 331, 991–1004.

Wright, F., 1990. The ‘effective number of codons’ used in a gene. Gene 87, 23–29.

Xie, T., Ding, D.F., 1998. The relationship between synonymous codon usage and protein structure. FEBS Lett. 434, 93–96.