Journal of Southeast University (English Edition) Vol. 21, No. 2, pp. 203 - 206 June 2005 ISSN 1003—7985

Site discrepancy of synonymous codon usage in SARS coronavirus and other viruses in *Coronaviridae*

Zhou Tong Gu Wanjun Ma Jianmin Sun Xiao Lu Zuhong

(Key Laboratory of Molecular and Bio-Molecular Electronics of Ministry of Education, Southeast University, Nanjing 210096, China)

Abstract: The synonymous codon usage in the translational initiation and termination regions of genes of severe acute respiratory syndrome (SARS) coronavirus and five other viruses in *Coronaviridae* was systematically analyzed. The results indicate that most minor codons for these coronaviruses are preferentially used in the initial and terminal region. The minor codons preferentially used in the initial region are thought to have a negative effect on gene expression, which can be explained by the minor codon modulator hypothesis. It also indicates that the minor codons preferentially used in the terminal region may regulate the level of gene expression. The proposed results strongly imply that the minor codon modulator hypothesis can be applied to both some bacteria and some viruses.

Key words: codon usage; severe acute respiratory syndrome (SARS); coronavirus; gene expression; site discrepancy

Due to the degeneracy of genetic code, most amino acids are coded by more than one codon (synonymous codon). 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 which recognize protein coding regions in uncharacterized genomic DNA. It has also been reported that synonymous codons are not used equally both within and between genomes[1]. Codon usage bias may result from various factors. Base composition constraints and translation selection are thought to be the main factors accounting for codon usage variation among genes in different organisms[2-4]. The diverse patterns of codon usage in mammals may arise from compositional constraints of the genomes. In contrast, in some unicellular organisms, such as *Escherichia coli* and *Saccharomyces cerevisiae*, high expressed genes have a strong selective preference for codons which are recognized by most abundant tRNAs, whereas low expressed genes display a more uniform pattern of codon usage. In some recent research, codon usage was found to be related to gene function[5] and protein secondary structure[6]. Further analysis found that synonymous codon usage pattern varied at different sites along a coding sequence[7]. In many bacterial species, such as *Deinococcus radiodurans, Haemophilus influenzae*, and *Methanobacterium thermoautotrophicum*, some minor codons are preferentially used near the initiation codon[8]. These minor codons are thought to play an important role in gene expression.

Severe acute respiratory syndrome (SARS) is a respiratory disease that has been reported in Asia, North America, and Europe. The whole genome of SARS coronavirus has been sequenced, which contains 29 727 nucleotides and the genome organization is similar to that of other coronaviruses[9]. Sequence and phylogenetic analysis have revealed that this coronavirus is only moderately related to other known coronaviruses[10].

Although genome sequence of SARS coronavirus has been published and many studies have been performed on SARS coronavirus in recent months, little genomic analysis is available on this virus. In this study we have used the available complete gene sequences and analyzed the codon usage patterns in the translational initiation and termination regions of SARS and other coronaviruses. Codon usage data of SARS coronavirus and the comparison results might give some clues to the features of SARS coronavirus genome.

1 Materials and Methods

1.1 Nucleic acid data set

SARS coronavirus (SARSCoV) is a large, enveloped, positive-stranded RNA virus, which belongs to order *Nidovirales*, family *Coronaviridae*, genus *Coronavirus* in virus taxonomy[9]. The complete genome and coding sequences of SARSCoV TOR2 isolation were obtained from GenBank (accession NC_004718). To compare the codon usage bias among

Received 2004-12-24.

Foundation item: The National Natural Science Foundation of China (No. 60121101).

Biographies: Zhou Tong (1977—), male, graduate; Lu Zuhong (corresponding author), male, doctor, professor, zhlu@seu.edu.cn.

different viruses, coding genes of five other viruses belonging to genus *Coronavirus* were also parsed from GenBank. They were bovine coronavirus (BCoV, accession AF220295), avian infectious bronchitis virus (AIBV, accession NC_001451), human coronavirus 229E (HCoV-229E, accession NC_002645), porcine epidemic diarrhea virus (PEDV, accession NC_003436) and transmissible gastroenteritis virus (TGV, accession NC_002306).

1.2 Relative synonymous codon usage

To examine synonymous codon usage without the confounding influence of amino acid composition of different gene samples, the values of relative synonymous codon usage f_{RSCU} of different codons in each genome have been calculated. The f_{RSCU} value of the j-th codon for the i-th amino acid is calculated as[2]

$$f_{\mathrm{RSCU}_{ij}} = \frac{s_{ij} \Big/ \sum_{j=1}^{n_i} s_{ij}}{1/n_i} \qquad (1)$$

where s_{ij} is the observed number of the j-th codon for the i-th amino acid which has n_i type of synonymous codons. It is obvious that f_{RSCU} values close to 1.0 indicate a lack of bias for the corresponding codon.

1.3 Analyzing codon usage in given regions

To analyze the codon usage of given regions in coding sequences, we define a variant of position balance of codon usage f_{PBCU}. f_{PBCU} values are calculated as[8]

$$f_{\mathrm{PBCU}} = \log\left(\frac{n_k/N_k}{n/N}\right) = \log\left(\frac{n_k N}{n N_k}\right) \qquad (2)$$

where k is the width of the given region, n_k is the total number of certain codons in the given region for all the coding sequences in one genome, N_k is the total number of corresponding amino acids in the given region, n is the total number of certain codon in all selected sequences of the genome, and N is the total number of corresponding amino acids in these sequences. Codons with f_{RSCU} values much lower than 1.0 are thought to be minor codons. It is obvious that f_{PBCU} values close to zero indicate a lack of bias for the corresponding codon used in the given region.

We focused on the initial region from the translational initiation site to the 30-th downstream codon and the terminal region from the translational termination site to the 30-th upstream codon. The f_{PBCU} values of these regions were calculated. The f_{RSCU} values of the corresponding genome were also calculated to compare with the f_{PBCU} values.

2 Results and Discussion

2.1 Codon usage near translational initiation site

Fig. 1 shows the relationship between f_{RSCU} (using the whole genomes) and f_{PBCU} for SARSCoV and the other five coronavirus species. Each dot represents a codon. A high f_{PBCU} value indicates that the corresponding codon is more preferentially used in the translational initiation region.

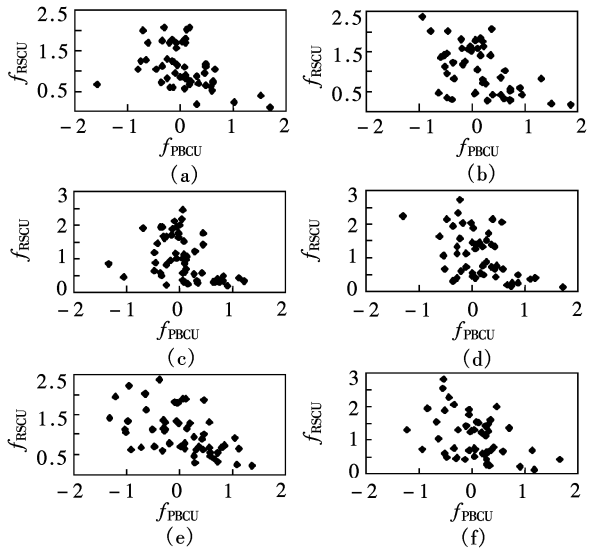

Fig. 1 The relationship between f_{RSCU} and f_{PBCU} of the initial translation region. (a) SARSCoV; (b) AIBV; (c) BCoV; (d) HCoV-229E; (e) PEDV; (f) TGV

Our study shows that f_{PBCU} values for some codons are large while their f_{RSCU} values are small. This means that these codons are minor codons in the whole genome, while they are preferentially used in the translational initiation site rather than in the rest of the gene in various viruses. These results are partially consistent with the previous report where several bacteria genomes are studied[8].

Codons with high f_{PBCU} values (larger than 1) in various organisms are listed in Tab. 1. All listed co-

Tab. 1 Preferentially used codons in initial region in various organisms

Organism	Amino acid	Codon	f_{RSCU}	f_{PBCU}
SARSCoV	Pro	CCC	0.40	1.52
	Ser	UCG	0.23	1.02
	Arg	CGG	0.09	1.69
AIBV	Gly	GGA	0.84	1.30
	Ser	UCC	0.19	1.50
	Arg	CGG	0.19	1.87
TGV	Ala	GCG	0.11	1.18
	Gly	GGA	0.71	1.13
	Ser	UCC	0.43	1.66
BCoV	Ser	UCC	0.44	1.11
	Arg	CGG	0.32	1.21
HCoV-229E	Gly	GGG	0.11	1.72
	Leu	CUA	0.39	1.18
	Phe	UUC	0.37	1.10
PEDV	Cys	UGC	0.65	1.12
	Ser	UCG	0.26	1.08
	Arg	CGG	0.22	1.37
	Arg	AGG	0.91	1.04

dons are minor. In other words, the f_{RSCU} values of these codons are lower than 1. This phenomenon is especially evident in codons CCC coding for Pro and CGG coding for Arg in SARSCoV, UCC coding for Ser and CGG coding for Arg in AIBV, UCC coding for Ser in TGV and GGG coding for Gly in HCoV-229E. They exhibit extremely high f_{PBCU} values and considerablely low f_{RSCU} values indicating that these minor codons are thirty- to sixty-fold more preferentially used in the initiation site than in the rest of the coding sequence. Similarly, codons GGA coding for Gly in AIBV, CGG coding for Arg in BCoV and CGG coding for Arg in PEDV also have very high f_{PBCU} values that correspond to more than fifteen-fold higher frequency in the initiation site than in the rest of the gene.

It has been previously reported that codon usage bias in SARSCoV is slight, which is mainly determined by the base composition on the third codon position. Compositional constraints can explain most of the variation of synonymous codon usage among the coronavirus genes, whereas translational selection may have little effect on the codon usage pattern[11]. However, the present study shows that translational selection may also have effect on codon usage in SARSCoV and other viruses in *Coronaviridae*, because minor codons near the initiation codon may play a role in regulating gene expression. Since even a single minor codon in the initiation site can reduce gene expression as a result of limited availability of tRNAs depending on the host cell, minor codons listed in Tab. 1 probably have a negative effect on gene expression. This can be explained by the minor codon modulator hypothesis. When the concentration of tRNA for minor codons becomes extremely limited, ribosomes of the host cell stall at minor codon sites, inhibiting the effective entry of a ribosome at the initiation site, thereby resulting in a decrease in the rate of translation. When the distance between the initiation codon and the minor codon is greater than 50 to 60 codons, a queue of ribosomes at the minor codon does not block the entry of a ribosome in the translation process. Thus, minor codons within this critical limit from the initiation codon are thought to play an important role in regulating gene expression[8].

2.2 Codon usage near translational termination site

Fig. 2 shows the relationship between f_{RSCU} (using the whole genomes) and f_{PBCU} (near the terminal codon) for SARSCoV and the other five coronavirus species. A high f_{PBCU} value indicates that the corresponding codon is more preferentially used in the translational termination region.

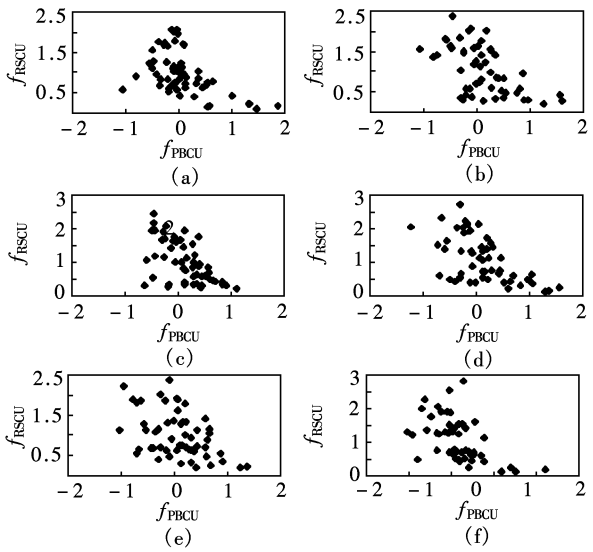

Fig. 2 The relationship between f_{RSCU} and f_{PBCU} of the terminal translation region. (a) SARSCoV; (b) AIBV; (c) BCoV; (d) HCoV-229E; (e) PEDV; (f) TGV

Resembling the codon usage in the initial region, codons with small f_{RSCU} value are also preferentially used in the translational termination site in these viruses. These results are partially consistent with the previous report where the genome of *Escherichia coli* was studied[12]. In that study, some codons, such as AGA and AGG coding for Arg, GGA and GGG coding for Gly, and GAG coding for Glu were found to be preferentially used in the terminal region and these codons were thought to be relative to the gene expression.

Codons with high f_{PBCU} values (larger than 1) in the terminal regions of these coronavirus genomes are listed in Tab. 2. It is obvious that the f_{RSCU} values of these codons are no higher than 0.63, which especially

Tab. 2 Preferentially used codons in terminal region in various organisms

Organism	Amino acid	Codon	f_{RSCU}	f_{PBCU}
SARSCoV	Ala	GCG	0.22	1.31
	Ser	UCG	0.23	1.33
	Arg	CGA	0.44	1.00
	Arg	CGG	0.09	1.46
	Thr	ACG	0.18	1.86
BCoV	Ala	GCG	0.22	1.09
TGV	Ala	GCG	0.11	1.50
	Gly	GGG	0.12	1.18
	Pro	CCC	0.25	1.40
	Arg	CGA	0.17	2.20
AIBV	Gly	GGG	0.21	1.25
	Pro	CCC	0.42	1.57
	Thr	ACG	0.27	1.59
HCoV-229E	Gly	GGG	0.11	1.29
	Phe	UUC	0.37	1.05
	Arg	CGA	0.25	1.56
	Arg	CGG	0.15	1.36
	Arg	AGG	0.63	1.05
PEDV	Pro	CCG	0.20	1.23
	Arg	CGG	0.22	1.37

applies to codons ACG coding for Thr in SARSCoV, CCC coding for Pro and ACG coding for Thr in AIBV, GCG coding for Ala and CGA coding for Arg in TGV and CGA coding for Arg in HCoV-229E. They exhibit extremely high f_{PBCU} values and considerablely low f_{RSCU} values indicating that these minor codons are more preferentially used in the termination site than in the rest of the coding sequence. This phenomenon is similar to that in the initial region and may also relate to the concentration of tRNA molecules in the host cell. Thus, minor codons listed in Tab. 2 probably also have a negative effect on gene expression.

Comparing Tab. 1 with Tab. 2, it can be found that some minor codons, such as UCG coding for Ser and CGG coding for Arg in SARSCoV, GGG coding for Gly and UUC coding for Phe in HCoV-229E, CGG coding for Arg in PEDV and GCG coding for Ala in TGV, are preferentially used not only in the translational initiation site but also in the termination site. It is suggested that these codons play a quite important role in the regulation of gene expression.

3 Conclusion

In summary, the codon usage in the translational initiation site and termination site greatly relates to the regulation of gene expression. Most minor codons in SARSCoV and the other viruses in *Coronaviridae* are preferentially used in the initial region, which can be explained by the minor codon modulator hypothesis. These codons within this critical region are thought to play a negative role in regulating gene expression. At the same time, some minor codons are also preferentially used in the terminal region in these viruses, which also probably has a relation with gene expression. Our results strongly imply that the minor codon modulator hypothesis can be applied not only to some bacteria but also to some viruses. Further, such information might be helpful to understand the pathogenesis and the origin of SARSCoV.

References

[1] Lloyd A T, Sharp P M. Evolution of codon usage patterns: the extent and nature of divergence between Candida albicans and Saccharomyces cerevisiae [J]. *Nucleic Acids Research*, 1992, **20**(20): 5289 − 5295.

[2] Sharp P M, Tuohy T, Mosurski K. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes [J]. *Nucleic Acids Research,* 1986, **14**(13): 5125 − 5143.

[3] Francino H P, Ochman H. Isochores result from mutation not selection [J]. *Nature*, 1999, **400**(6739): 30 − 31.

[4] Gupta S K, Ghosh T C. Expressivity is the main factor in dictating the codon usage variation among the genes in Pseudomonas aeruginosa [J]. *Gene,* 2001, **273**(1): 63 − 70.

[5] Ma J, Zhou T, Gu W, et al. Cluster analysis of the codon use frequency of MHC genes from different species [J]. *Biosystems*, 2002, **65**(2): 199 − 207.

[6] Gu W, Zhou T, Ma J, et al. The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens [J]. *Biosystems,* 2004, **73**(2): 89 − 97.

[7] Hooper S D, Berg O G. Gradients in nucleotide and codon usage along Escherichia coli genes [J]. *Nucleic Acids Research,* 2000, **28**(18): 3517 − 3523.

[8] Ohno H, Sakai H, Washio T, et al. Preferential usage of some minor codons in bacteria [J]. *Gene*, 2001, **276**(1, 2): 107 − 115.

[9] Marra M A, Jones S J, Astell C R, et al. The genome sequence of the SARS-associated coronavirus [J]. *Science*, 2003, **300**(5624): 1399 − 1404.

[10] Paul A R, Steven O M, Stephan S M, et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome [J]. *Science*, 2003, **300**(5624): 1394 − 1399.

[11] Gu W, Zhou T, Ma J, et al. Analysis of synonymous codon usage in SARS coronavirus and other viruses in the Nidovirales [J]. *Virus Research,* 2004, **101**(2): 155 − 161.

[12] Li W. Position dependence of synonymous codon usage [J]. *Acta Biophysica Sinica*, 2001, **17**(3): 529 − 534. (in Chinese)

SARS 及其他几种冠状病毒密码子使用偏性的位点差异

周 童 顾万君 马建民 孙 啸 陆祖宏

(东南大学分子与生物分子电子学教育部重点实验室, 南京 210096)

摘要: 分析了 SARS 冠状病毒以及其他 5 种冠状病毒基因编码起始区与终止区的密码子使用偏性. 结果表明, 冠状病毒基因组的稀有密码子倾向于出现在编码起始区和终止区附近. 起始区的这种倾向性对冠状病毒基因的表达具有负性调控作用, 可以用"稀有密码子调控假说"解释. 终止区的这种倾向性表明, 这些出现在终止区的稀有密码子对基因的表达也有负性调节作用. 研究结果同时暗示了"稀有密码子调控假说"不仅适用于细菌, 而且还适用于某些病毒基因组.

关键词: 密码子使用偏性; SARS; 冠状病毒; 基因表达; 位点差异

中图分类号: Q617