ATLANTIS PRESS

Advances in Biological Sciences Research (ABSR), volume 6

2017 2nd International Conference on Biological Sciences and Technology (BST 2017)

Analysis of Codon Bias on *Clostridium Perfringens* Del1 Genome

Lin CHEN[1,#], Ting BAI[1,#], Wei WANG[1,a,*], Tian-fei LIU[2], Li-li JI[1] and Jia-min ZHANG[1]

[1]Key Lab of Meat Processing of Sichuan Province, Chengdu University, Chengdu, 610106, China

[2]Institute of Animal Science, Guangdong Academy of Agricultural Sciences, Guangzhou, 510640, China

[a]wangwei8619@163.com

*Corresponding author

[#]Equal contributors

Keywords: *Clostridium perfringens*, Codon usage bias, Correspondence analysis, Optimal codon.

Abstract. Clostridium perfringens is a conditional pathogen, which widely exists in the nature water, soil, intestinal tract, and serious threat to public health and safety. In this study, the codon usage of Clostridium perfringens was examined. CDS sequences of 2744 high confidence protein coding genes were selected from Clostridium perfringens Genome Database. Their GC content of these genes ranged from 14.7% to 41.9%, with the average of 29.9% for whole genome. The frequency of G or C in the third position of synonymous codon was 25.6%, which was lower than that of A or T. The effective number of codon(ENC) of Clostridium perfringens genome ranged from 26.9 to 61.0,with the average of 39.9. Clostridium perfringens prefer codons ending in A/T. The ENC values of each gene significantly positive correlated with its C3s, G3s, GC3s and GC content (r=0.674, 0.738, 0.780 and 0.368, respectively, $p<0.01$). However, ENC showed significant negative correlations with T3s, A3s and CAI (r=-0.668, -0.688 and -0.255, respectively, $p<0.01$). A total of 27 optimal codons were found from Clostridium perfringens genome, all of which the third position of synonymous codon was A or T.

Introduction

Most amino acids can be coded by more than one triplet of nucleotides. Such codons are defined as synonymous codons. Synonymous codons are not used equally both within and between genomes [1, 2] many factors have been reported to influence codon usage in various organisms. Compositional constraints and translational selection are considered to be the main influences[3].

Studies of synonymous codon usage can help us to understanding the mechanisms of biased usage of synonymous codons[4], the selection of appropriate host expression systems[5], the design of degenerate primers[6], gene prediction from genomic sequences[7], and protein functional classification[8]. In addition, profiles of synonymous codon usage can reveal information about the molecular evolution of individual genes and provide data to train genome-specific gene recognition algorithms, which detect protein-coding regions in uncharacterized genomic DNA[9].

Clostridium perfringens (*C. perfringens*)is a Gram-positive, rod-shaped, anaerobic, spore-forming pathogenic bacterium[10]. In the United Kingdom and United States, *C.*

perfringens bacteria are the third most common cause of foodborne illness, with poorly prepared meat and poultry, or food properly prepared but left to stand too long, the main culprits in harboring the bacterium.

In this study, we analyzed the codon usage of C. perfringens from genome using the CodonW 1.4.2 program. The results may provide some insights into the basic characteristics of the C. perfringens genome as well as information about the evolution of the bacterium. Knowledge of the codon usage profile in C. perfringens can provide a basis for understanding the mechanisms of biased usage of synonymous codons.

Materials and Methods

Data

Data of genomic coding sequences were retrieved from the genome database in National Center for Biotechnology Information (NCBI; http:www.ncbi. nlm.nih.gov/nuccore/NZ_cp019576.1). Our own program developed in PERL script was used to extract the coding sequences.The dataset obtained was then manually checked to correct existing errors, and the 5'-or 3'- partial CDSs were removed. To minimize sampling errors, genes less than 300bp were excluded. Finally, A total of 2744 genes was further analyzed.

Parameters for Assessment of Codon Bias and Statistic Analysis

The following codon indices were examined: relative synonymous codon usage (RSCU), effective number of codons (ENC), codon adaptation index (CAI), frequency of optimal codon (FOP), frequencies of adenine (A3s), cytosine (T3s),guanine (G3s), and cytimidine (C3s) in the third position of synonymous codons, the average value of GC-content in the first and second position of the codons (GC12), the GC-content in the third position(GC3) and the content of either a G or C at the third codon position of synonymous codons (GC3s).

RSCU is a parameter for assessing usage bias of 59 synonymous codons except for the termination codons TAG, TGG, and TGA and the unique codons Met and Trp. The RSCU value is defined as the ratio of actual observed value of synonymous codons to the expected value when they were used with the same probability[3]. The RSCU value is 1.0 when the synonymous codons are completely used in random; however, this value will be larger than 1.0 when a codon has more frequent use than other codons, and vice versa.

ENC shows the magnitude of codon bias of a single gene,which ranges from 20 (each amino acid is encoded by only one codon) to 61 (all the codons are used randomly). Genes with lower ENC have stronger codon usage bias[11].

CAI is the match degree between synonymous codons of high confidence CDS and optimal codons, which ranges from 0.0 to 1.0. The genes with high expression levels exhibit high CAI values, and vice versa. CAI value has been widely used in predicting gene expression level because it is extremely approaching the observed value of gene expression [12, 13].

The frequency of optimal codon (FOP) is the percentage of optimal codons to all codons.

Software

CodonW 1.4.2 (http://codonw.sourceforge.net/) was used for calculating the indices of codon usage. SPSS16.0

(http://www01.ibm.com/software/analytics/spss/downloads.html) and excel 2007 were implemented for statistical analysis.

Results

Codon Composition Analysis

The G+C content is 14.7~41.9%. The average content of GC in whole genome was 29.9%. The GC content 3rd position of synonymous codons(25.6%) is lower than AT content 3rd position of synonymous codons. The content of G3s, C3s, A3s and T3s were 24.8%, 11.5%, 53.0% and 43.2% (Table 1).

Table 1 The constitution and usage parameters of condons

Parameter of condons	Variation range	X±SD
T3s	0.156~ 0.689	0.432±0.103
C3s	0.008~ 0.317	0.115±0.051
A3s	0.171 ~0.784	0.530±0.082
G3s	0.000 ~0.659	0.248±0.144
CAI	0.055 ~0.308	0.141±0.031
ENC	26.940~ 61.000	39.896±7.005
GC3s	0.037~ 0.639	0.256 ±0.126
GC	0.147~ 0.419	0.299±0.037
FOP	0.112~ 0.558	0.306±0.048

Usage Preference Analysis of Synonymous Codon

The whole genome of *C. perfringens* was analyzed and processed with CodonW software. Data on the preferences and frequency of synonymous codons for all amino acids of the *C. perfringens* genome are listed (Table 2). In codons with RSCU values greater than 1, codons ending with A, U, G, and C accounted for 56.9%, 41.2%, 0, and 1.91% respectively. This shows that *C. perfringens* prefer codons ending in A/U.

Table 2 Codon usage of *C. perfringens* genes; AA: amino acid; N: the number of codons. The preferentially used codons are displayed in bold.

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Ala	**GCU**	13868	1.97	Leu	UUG	19459	1.00
	GCC	2413	0.34		CUU	17733	0.91
	GCA	10439	1.48		CUC	4415	0.23
	GCG	1414	0.20		CUA	17931	0.92
Arg	CGU	792	0.15		CUG	10253	0.52
	CGC	243	0.04	Lys	**AAA**	56565	1.20
	CGA	435	0.08		AAG	37365	0.80
	CGG	587	0.11	Phe	**UUU**	30526	1.53
	AGA	19860	3.66		UUC	9395	0.47
	AGG	10625	1.96	Pro	**CCU**	7415	1.84
Asn	**AAU**	37434	1.58		CCC	778	0.19
	AAC	10068	0.42		**CCA**	7392	1.84
Asp	**GAU**	27603	1.66		CCG	491	0.12
	GAC	5675	0.34	Ser	**UCU**	10354	1.43
Cys	**UGU**	6538	1.35		UCC	2283	0.32
	UGC	3120	0.65		**UCA**	10894	1.51
Gln	**CAA**	20147	1.17		UCG	1131	0.16
	CAG	14324	0.83		**AGU**	13140	1.82
Glu	**GAA**	39982	1.36		AGC	5539	0.77
	GAG	18994	0.64	Thr	**ACU**	15210	1.79
Gly	**GGU**	9025	1.02		ACC	2809	0.33
	GGC	2340	0.26		**ACA**	13559	1.60
	GGA	18688	2.11		ACG	2338	0.28
	GGG	5317	0.60	Tyr	**UAU**	29089	1.57
His	**CAU**	10635	1.47		UAC	7861	0.43
	CAC	3822	0.53	Val	**GUU**	18159	1.47
Ile	AUU	28698	0.98		GUC	2437	0.20
	AUC	7171	0.25		**GUA**	19173	1.55
	AUA	51808	1.77		GUG	9762	0.79
Leu	**UUA**	47527	2.43				

Association between Effective Number of Codons (ENC) and GC3s

Plotting ENC and GC3s is an effective way to explore the main features of codon usage among genes[11]. The relationship between codon usage parameters was also analyzed and the relationship between ENC and GC3s is shown (Figure 1). The solid line represents the relationship between ENC and GC3s without any selection pressure. Notably, most genes with lower ENC values than expected were lying well below the curve, suggesting there were other factors, particularly selective expression level affecting the codon usage combined with composition mutation.

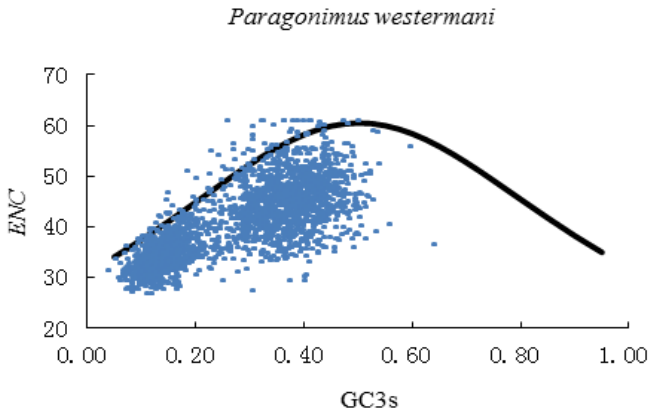

Figure 1 Nc-plot of C. perfringens genes Strain Del1

Neutrality Plot Analysis

In the neutrality plot of all the genes generated to evaluate the relationships among the three positions in *C. perfringens* codons (Figure. 2), most did not lie on or along the diagonal line. In addition, the ranges of GC12 and GC3 were narrow (16.75%–53.40% and 6.30%–31.70%, respectively). These data suggested that *C perfringens* codon usage is affected by natural selection. Moreover, linear regression of the entire coding sequence data yielded a slope of 0.3324, revealing that directional mutation pressure accounts for only 33.24% of the effect, while other factors (e.g. natural selection) account for 66.76% of the influence.

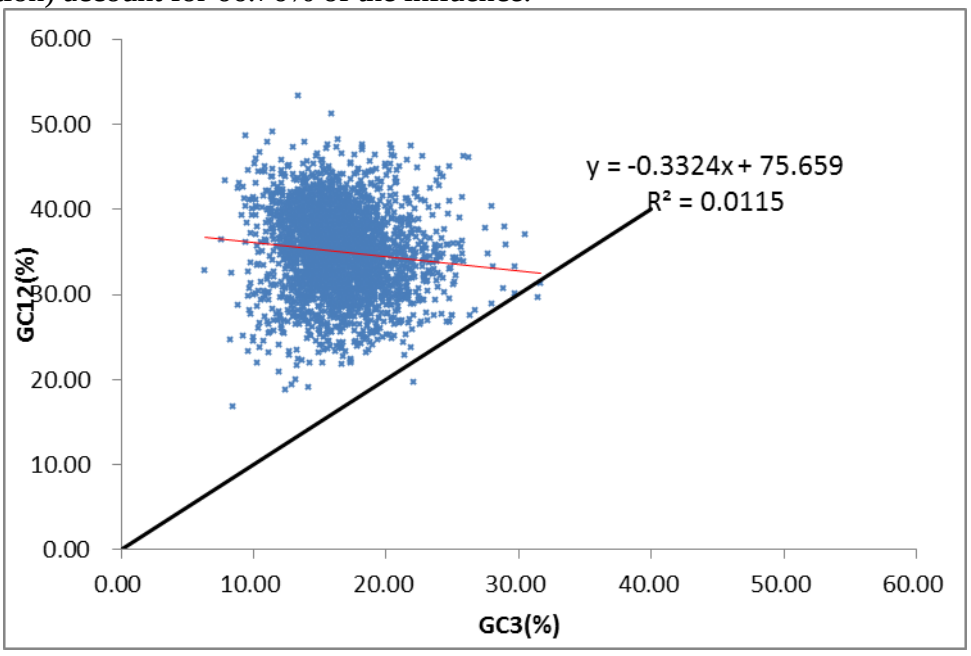

Figure 2 Neutrality plot analysis of the GC12 and GC3 values of the C. perfringens. GC12 represents for the average value of GC-content in the first and second position of the codons (GC1 and GC2), while GC3 represents the GC-content in the third position. The red line shows the linear regression of GC12 against GC3, R2=-0.0115.

Optimal Translational Codons

The average RSCU values of high/low expressed gene sample group are listed in Table 3. Twenty-seven codons, including UUU, UUA, AUA and AUU, were identified as optimal translational codons based on the average RSCU values of the

high and low datasets. The optimal codons all ended with A or U (Table 3).

Table 3 Translational optimal codons of C. perfringens

AA	Codon	High RSCU(N)	Low RSCU(N)	AA	Codon	High RSCU(N)	Low RSCU(N)
Phe	UUU*	1.40 (1216)	1.21 (808)	Ser	UCU*	1.34 (508)	1.05 (271)
	UUC	0.60 (523)	0.79 (531)		UCC	0.06 (22)	0.60 (155)
Leu	UUA*	4.46 (2642)	1.16 (1413)		UCA*	2.77 (1050)	1.13 (292)
	UUG	0.07 (43)	1.40 (1706)		UCG	0.01 (2)	1.06 (276)
	CUU*	1.11 (654)	0.67 (818)		AGU*	1.45 (547)	1.18 (307)
	CUC	0.00 (1)	0.40 (491)		AGC	0.38 (142)	0.98 (254)
	CUA	0.35 (206)	1.24 (1503)	Pro	CCU	1.26 (401)	1.54 (91)
	CUG	0.01 (5)	1.13 (1369)		CCC	0.01 (3)	0.63 (37)
Ile	AUU*	0.87 (950)	0.76 (738)		CCA*	2.71 (864)	0.71 (42)
	AUC	0.20 (214)	0.46 (443)		CCG	0.02 (7)	1.12 (66)
	AUA*	1.93 (2101)	1.78 (1730)	Thr	ACU*	2.11 (1065)	1.13 (482)
Met	AUG	1.00 (1162)	1.00 (1913)		ACC	0.05 (27)	0.49 (210)
Val	GUU*	2.27 (1582)	0.53 (281)		ACA*	1.82 (917)	1.29 (547)
	GUC	0.04 (25)	0.46 (245)		ACG	0.01 (7)	1.08 (460)
	GUA*	1.59 (1107)	1.28 (682)	Ala	GCU*	2.37 (1765)	0.89 (85)
	GUG	0.11 (77)	1.74 (929)		GCC	0.15 (110)	0.48 (46)
Tyr	UAU*	1.59 (1193)	1.37 (941)		GCA*	1.44 (1074)	0.77 (73)
	UAC	0.41 (305)	0.63 (436)		GCG	0.04 (29)	1.85 (176)
His	CAU*	1.53 (446)	1.32 (629)	Cys	UGU*	1.72 (386)	0.98 (179)
	CAC	0.47 (136)	0.68 (322)		UGC	0.28 (63)	1.02 (186)
Gln	CAA*	1.92 (859)	0.89 (1334)	Trp	UGG	1.00 (327)	1.00 (554)
	CAG	0.08 (35)	1.11 (1656)	Arg	CGU*	0.40 (84)	0.06 (12)
Asn	AAU*	1.53 (1714)	1.37 (1068)		CGC	0.00 (0)	0.07 (16)
	AAC	0.47 (530)	0.63 (492)		CGA	0.01 (2)	0.13 (28)
Lys	AAA*	1.50 (2650)	0.93 (2331)		CGG	0.00 (0)	0.30 (65)
	AAG	0.50 (880)	1.07 (2657)		AGA*	5.53 (1155)	2.27 (487)
Asp	GAU*	1.71 (1978)	1.27 (655)		AGG	0.06 (12)	3.17 (679)
	GAC	0.29 (341)	0.73 (373)	Gly	GGU*	1.24 (931)	0.44 (44)
Glu	GAA*	1.64 (2795)	1.00 (1289)		GGC	0.12 (88)	0.46 (46)
	GAG	0.36 (605)	1.00 (1280)		GGA*	2.47 (1856)	0.77 (78)
					GGG	0.18 (132)	2.34 (236)

Correlation Analysis of Codon Evaluation Parameters

The ENC values of each gene significantly positive correlated with its C3s, G3s, GC3s and GC content (r=0.674, 0.738, 0.780 and 0.368, respectively, $p<0.01$). However, ENC showed significant negative correlations with T3s, A3s and CAI (r=-0.668, -0.688 and -0.255, respectively, $p<0.01$, Table 4). These results indicated that gene expression levels were affected GC content, especially G/C content of the third position. To be more specific, genes with higher expression levels had a greater degree of FOP and GC-rich content. Furthermore, these genes exhibited preference for codons with C or G at the synonymous position.

Table 4 Correlation analysis of each related parameters

	T3s	C3s	A3s	G3s	CAI	Fop	ENC	GC3s	GC
T3s	1.000								
C3s	-0.684**	1.000							
A3s	0.578**	-0.639**	1.000						
G3s	-0.808**	0.602**	-0.838**	1.000					
CAI	0.469**	-0.044*	0.189**	-0.384**	1.000				
Fop	0.068**	0.289**	0.038*	-0.171**	0.742**	1.000			
ENC	-0.668**	0.674**	-0.688**	0.738**	-0.255**	-0.034	1.000		
GC_{3s}	-0.863**	0.774**	-0.864**	0.964**	-0.298**	-0.021	0.780**	1.000	
GC	-0.508**	0.454**	-0.537**	0.436**	0.185**	0.318**	0.368**	0.539**	1.000

Note:**: Correlation is significant at the 0.01 level; *: Correlation is significant at the 0.05 level

Discussion

Nucleotide composition is one of the most important factors that shapes codon usage with GC-content reflecting the overall trend of codon mutation[14]. The average GC-content of the total of 2,744 C. perfringens genes investigated was 29.9% (below the average AU content), while the average GC3s content was lower at 25.6%. These results are consistent with the GC and AU contents of C. acetobutylicum[15].

The uses of codons are different between species during gene expression. Different organisms have different preference of the synonymous codons. The base of the codon at three sites will mutate during evolution. But the meaning of the mutation is very different. The mutations in the first two positions are usually lead to changes in the edcoded amino-acid sequence, while the mutations of the third position rarely induces such sequence variation [16]. It is generally acknowledged that the third codon position is subject to lower selection pressure compared with that of the first and second codon positions. Thus, ENc-GC3s correlation analysis, and neutrality plot analysis based on GC3 or GC3s are important for elucidation of the codon usage patterns in many organisms.

In this study, we firstly analyzed the codon usage of *C. perfringens* using the the whole genome data. The result shown that all of the optimal codons are ended with A or U, which indicate that the *C. perfringens* genome prefers codon ending with A or U. Neutrality plot analysis shown that most genes with lower ENC values than expected ones were lying well below the curve. GC3s-ENC revealed that most genes were not lie on or along the diagonal line. These data suggested that there are other facters (e.g. natural selection) influence the codon usage as well as composition mutation.

Acknowledgments

This work was supported by the Applied and Basic Research Program of Sichuan Province (No. 2015JY0018), Scientific and Technological Support Program of Sichuan Province (2016NZ0003, 2016NZ0002) and Youth Foundation of Chengdu University(2017XJZ21)

References

[1] Grantham R, Gautier C, Gouy M, etal. Codon catalog usage and the genome hypothesis[J]. Nucleic Acids Res,1980,8(1):r49-r62.

[2] 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 Res,1992,20(20):5289-5295.

[3] Sharp P M, Li W H. An evolutionary perspective on synonymous codon usage in unicellular organisms[J]. J Mol Evol,1986,24(1-2):28-38.

[4] Powell J R, Moriyama E N. Evolution of codon usage bias in Drosophila[J]. Proc Natl Acad Sci U S A,1997,94(15):7784-7790.

[5] Zheng Y, Zhao W M, Wang H, etal. Codon usage bias in Chlamydia trachomatis and the effect of codon modification in the MOMP gene on immune responses to vaccination[J]. Biochem Cell Biol,2007,85(2):218-226.

[6] Zhou T, Gu W, Ma J, etal. Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses[J]. Biosystems,2005,81(1):77-86.

[7] Salamov A A, Solovyev V V. Ab initio gene finding in Drosophila genomic DNA[J]. Genome Res,2000,10(4):516-522.

[8] Lin K, Kuang Y, Joseph J S, etal. Conserved codon composition of ribosomal protein coding genes in Escherichia coli, Mycobacterium tuberculosis and Saccharomyces cerevisiae: lessons from supervised machine learning in functional genomics[J]. Nucleic Acids Res,2002,30(11):2599-2607.

[9] Fickett J W. Recognition of protein coding regions in DNA sequences[J]. Nucleic Acids Res,1982,10(17):5303-5318.

[10] McClane B A. Clostridium perfringens[J]. FOOD SCIENCE AND TECHNOLOGY-NEW YORK-MARCEL DEKKER-,2003,:91-104.

[11] Wright F. The 'effective number of codons' used in a gene[J]. Gene,1990,87(1):23-29.

[12] Sharp P M, Li W H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications[J]. Nucleic Acids Res,1987,15(3):1281-1295.

[13] Carbone A, Zinovyev A, Kepes F. Codon adaptation index as a measure of dominating codon bias[J]. Bioinformatics,2003,19(16):2005-2015.

[14] Shang M. Z., Liu F., P. H J. Analysis on codon usage of chloroplast genome of Gossypium hirsutum[J]. Scientia Agricultura Sinica,2011,44(2):245-253.

[15] Musto H, Romero H, Zavala A. Translational selection is operative for synonymous codon usage in Clostridium perfringens and Clostridium acetobutylicum[J]. Microbiology,2003,149(Pt 4):855-863.

[16] Hui F, Junjie Q, Jiaman S, etal. The Codon Usage Bias of NBS Disease Resistance Genes in Whole Genomeof Banana[J]. Molecular Plant Breeding,2017,15(3):883-889.