Journal of Theoretical and Applied Information Technology
28th February 2013. Vol. 48 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COMPUTATIONAL ANALYSIS OF THE SYNONYMOUS CODON USAGE AND MUTATION DISTRIBUTION IN CARD15

Sim-Hui Tee
Multimedia University, Cyberjaya, Malaysia
E-mail: shtee@mmu.edu.my

ABSTRACT

Synonymous codon usage is an extensive phenomenon found across species, including human. Understanding of this phenomenon may aid in better understanding of gene expression mechanism, which has a practical utility in biomedicine research such as combating diseases. This research employed a computational approach to study the synonymous codon usage and mutation distribution in Card15 gene. The obtained results characterize the main factor that drives the codon usage in Card15 gene and the extensiveness of mutation. This computational insight into Card15 gene may shed light on the immunologic mechanisms of human against pathogen such as viruses.

Keywords: *Bioinformatics, Database, Scientific Computing, Synonymous Codon, Gene, Card15*

1. INTRODUCTION

Because of the degeneracy of genetic code, most amino acids (except Met and Trp) are encoded by more than one codon, which is known as synonymous codon [1-2]. Synonymous codons are used preferentially in different organisms, with some codons used more often than others [3-4]. This phenomenon is known as codon usage bias [5], which could be influenced by various factors such as mutational bias [6-8], translational selection [8-10], tRNA abundance [11,13], G+C content [12], temperature [12], genome segmentation [14], and CpG islands [15]. Despite these factors that shape codon usage, it is widely agreed that the codon usage in unicellular organisms is the result of the balance between mutational biases (either GC- or AT-inclined) and translational section [15]. In general, unicellular organisms such as E. coli display a pattern of high codon usage bias when the genes are highly expressed. The converse is true when the gene expression level is low. Mutational bias plays a role as a major factor when the genomic composition is biased (e.g., high G+C or A+U content). Different patterns of synonymous codon usage were observed in vertebrate. However, it was reported that some species exhibit the identical factor of codon usage as the unicellular organisms [15]. It was also observed that, for cold-blooded organisms, the G+C content at third codon position appears to be the main factor that shapes codon usage [15]. However, the determining factors of the synonymous codon usage in most organisms remained elusive [16].

Because of the wide disparity in codon usage pattern of multicellular organisms, it is important to study the synonymous codon usage in order to understand the evolution and gene expression of an organism. This study delimits to study the synonymous codon usage in Card15 gene, which is a gene that codes for NOD2 protein in the human immune system. NOD2 protein is a crucial pathogen sensor that detects the microbial molecular patterns and activates the downstream signaling cascade for immunologic defense [18-20]. Point mutation in the NOD2-encoding Card15 gene will result in the attenuation of NF-κB activation [21], which is an essential transcription factor that regulates the expression level of proinflammatory cytokines and other major cytokines of immunologic sentinels [22-26]. Therefore, understanding of the relation between mutation and synonymous codon usage pattern may shed light into the gene expression mechanism.

Although most of the synonymous codon usage investigations involve the whole genome in the study, it was reported that the synonymous codon usage can vary to certain extents across genes within an organism [17]. Therefore, we focus on the identification and analysis of the synonymous codon usage in a single gene, which is Card15, in human. In addition, the mutation distribution of

Card15 is also studied using a computational approach.

2. METHODS

The nucleotide sequence of Card15 gene was retrieved from GenBank of National Center for Biotechnology Information (NCBI). The open reading frame of Card15 was identified and the internal stop codons were removed. To analyze the synonymous codon usage in Card15, we used the relative synonymous codon usage (RSCU), a metric formulated by Sharp et al. [27]:

$$RSCU_{ij} = \frac{X_{ij}}{(1/n_i)\sum_{j=1}^{ni} X_{ij}} \qquad (1)$$

where X_{ij} is the number of the j^{th} codon for the i^{th} amino acid encoded by n_i synonymous codons. RSCU captures the ratio of observed number of occurrence of a codon to the expected random (non-bias) synonymous codon usage. Amino acid Trp and Met always yield 1.0, because they do not have alternative synonymous codon. RSCU value of a codon that is higher than 1.0 implies that there is a higher preferential usage than the expected random usage. Three stop codons were excluded from analysis because they do not code for amino acids.

The effective number of codon (ENC) [28] was computed to measure the general non-uniformity of synonymous codon usage in Card15. The values of ENC are in the range between 20 (only one codon is used among the synonymous codon for each amino acid) to 61 (all synonymous codons are equally used for each amino acid). The lower the ENC value, the more bias the codon is used in gene expression.

Codon bias index (CBI) [29] was computed to measure the directional codon bias, which is the extent to which a ribosome uses a subset of optimal codons in translation. CBI value for extreme codon usage bias is always 1.0, while a gene which exhibits a random codon usage pattern will yield 0 for its CBI value. There are cases where the CBI value being negative, implying that the number of optimal codon usage is less than the expectation.

COSMIC database [30] was used to mine the mutated nucleotides in Card15 gene. The extensiveness of the somatic mutation in the form of insertion/deletion (Indel) and substitution was studied.

3. RESULTS AND DISCUSSION

RSCU values were calculated for all codons in Card15. RSCU values greater than 1.0 implies that the investigated codon is used more frequently than expected; the reverse is true when RSCU value is less than 1.0. A list of RSCU values and the number of occurrence of each sense codon in Card15 gene are enumerated in Table 1. The preferentially used codons for each amino acid are displayed in bold font style.

Table 1. RSCU Values For Codons Of Card15 Gene

Amino acid	Codon	No. of occurrence	RSCU value
Phe	**UUU**	22	1.07
	UUC	19	0.93
Leu	UUA	9	0.42
	UUG	25	1.17
	CUU	25	1.17
	CUC	24	1.13
	CUA	4	0.19
	CUG	41	1.92
Ile	AUU	11	1.14
	AUC	6	0.62
	AUA	12	1.24
Met	AUG	34	1.00
Val	GUU	15	0.85
	GUC	14	0.79
	GUA	7	0.39
	GUG	35	1.97
Ser	UCU	36	1.13
	UCC	54	1.69
	UCA	39	1.22
	UCG	10	0.31
Pro	CCU	30	0.90
	CCC	50	1.50
	CCA	37	1.11
	CCG	16	0.48
Thr	ACU	27	1.04
	ACC	31	1.19
	ACA	34	1.31
	ACG	12	0.46
Ala	GCU	25	1.00
	GCC	37	1.48
	GCA	33	1.32
	GCG	5	0.20
Tyr	**UAU**	11	1.29
	UAC	6	0.71
His	CAU	17	0.94
	CAC	19	1.06
Gln	CAA	18	0.60
	CAG	42	1.40
Asn	**AAU**	18	1.38
	AAC	8	0.62
Lys	AAA	22	0.96
	AAG	24	1.04
Asp	GAU	7	0.64
	GAC	15	1.36
Glu	GAA	10	0.83
	GAG	14	1.17
Cys	UGU	33	0.73
	UGC	57	1.27
Trp	UGG	66	1.00
Arg	CGU	4	0.20

	CGC	8	0.41
	CGA	4	0.20
	CGG	16	0.81
Ser	AGU	19	0.59
	AGC	34	1.06
Arg	AGA	32	1.63
	AGG	54	2.75
Gly	GGU	12	0.50
	GGC	27	1.13
	GGA	28	1.17
	GGG	29	1.21

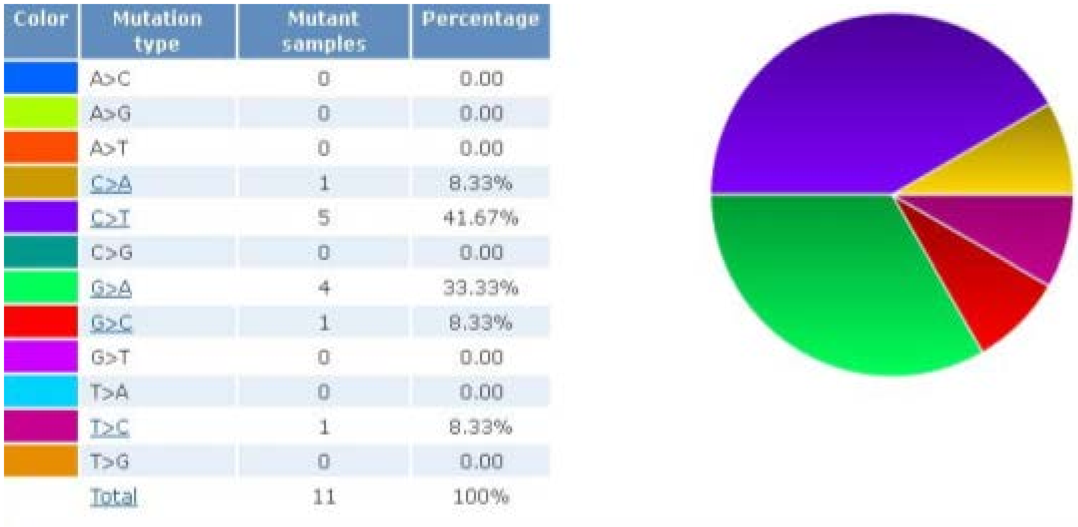
We observe that the preferentially used codons tend to be G+C at the third synonymous position (GC3s), which yields a value of 54.90%. Besides, the GC content is 55.8%, which is relatively higher than AU content. Taken GC3s and GC content together, it appears that mutational bias is the factor that drives the synonymous codon usage bias in Card15 gene. From Table 1, we observe that codon AGG of Arg is the most preferentially used synonymous codon (RSCU=2.75), while codon CUA of Leu is the least preferentially used synonymous codon (RSCU=0.19).

It is interesting to compare GC3s of Card15 gene with other genes. A study carried out by Sau et al. [31] showed that the mean GC3s value of 16 *Staphylococcus aureus* phages is as low as 23%, which does not play a role in shaping the synonymous codon usage of these phages. Gu et al. [32] reported a low GC3s value for transmissible gastroenteritis virus (27.02%) and avian infectious bronchitis virus (26.09%), while a GC3s value for porcine reproductive and respiratory syndrome virus (53.76%) is almost at the same level as our observation of Card15 gene. The low GC content (37.52%) observed in severe acute respiratory Coronavirus (SARSCoV) genes led Gu et al. [32] to conclude that A+U codons are preferentially used. Romero et al. [33] reported a correlation between GC3s and the synonymous codon usage in three species of fishes from the family Cyprinidae, with high GC3s value for *B. rerio* (57%), *C. carpio* (58%), and *C. auratus* (57%).

To measure the extent of mutational bias, we computed ENC value. High ENC value, which is 52.28, was observed in Card15 gene. This implies that though GC3s drives the preferential usage of synonymous codon in Card15 gene, the usage bias is quite low. This could be due to a lesser mutational pressure on the gene. Romero et al. [33] have reported an ENC value of 33, 29, 31 for three species of fishes *B. rerio*, *C. carpio*, and *C. auratus*, respectively, demonstrating a very biased pattern of synonymous codon usage in these fishes. Zhao et al. [34], on the other hand, reported a moderately biased pattern of synonymous codon usage in 11 human bocavirus (HBoV) isolates, with WLL-3-VP2 gene (ENC=41.27), WLL-2-NP1 gene (ENC=47.96), and BJ3722-VP1 gene (ENC=41.97), among others. Das et al. [35] have observed a less biased synonymous codon usage pattern in several adenoviruses, including canine adenovirus (ENC=54.67), fowl adenovirus A (ENC=52.36), human adenovirus A (ENC=54.15), and human adenovirus B (ENC=51.88). Interestingly, despite belong to the same *Adenoviridae* family, different adenoviral species display different extent of codon usage bias. Das et al. [35] also reported moderately biased codon usage in bovine adenovirus D (ENC=44.46), human adenovirus C (ENC=47.21), and ovine adenovirus D (ENC=42.56); besides, highly biased codon usage was found in porcine adenovirus A (ENC=38.97). The variation of ENC value in different species of the same family suggests that the synonymous codon usage bias varies across organisms.

Codon bias index (CBI) [29] was computed to measure the directional codon bias in Card15 gene, which is the extent to which a ribosome uses a subset of optimal codons in translation. We have obtained -0.012 as CBI value, implies that the optimal codon usage in Card15 gene is less than the norm. This fact has further corroborated the conclusion of a less biased synonymous codon usage in Card15 gene, as indicated by high ENC value.

We used COSMIC database [30] to mine the mutated nucleotides in Card15 gene. Mutations in gene could lead to the change of phenotype and the pathogenesis of various intractable diseases such as cancers [36-39], neurological disorders [40], and cardiovascular diseases [41-42]. We retrieved the mutation data for substitution at the coding strand. There was no insertion/deletion form of mutation found in Card15 gene. Figure 1 illustrates the substitution that occurs at the coding strand.



Color	Mutation type	Mutant samples	Percentage
🟦	A>C	0	0.00
🟩	A>G	0	0.00
🟧	A>T	0	0.00
🟨	C>A	1	8.33%
🟪	C>T	5	41.67%
🟦	C>G	0	0.00
🟩	G>A	4	33.33%
🟥	G>C	1	8.33%
🟪	G>T	0	0.00
🟦	T>A	0	0.00
🟪	T>C	1	8.33%
🟧	T>G	0	0.00
	Total	11	100%

Figure 1. Distribution Of Substitution At The Coding Strand

Figure 1 demonstrates various mutation types in the form of substitution that occur at the coding strand. These mutations have significant impacts because the coding strand is used by the ribosomes for protein translation, mutation of which will result in the change of amino acid. From Figure 1, it is found that four out of five mutated nucleotides are either guanine or cytosine. There are five cytosines were found mutated to thymine, following with four guanine mutated to adenine. Among a total of 11 mutations, 5 were found at the functional domains. It was found that there is 1 mutation at the N-terminal Card domain (position 233; G>C), 3 mutations at the central Nacht domain (position 976 C>A; position 1121 T>C; position 1195 G>A), and 1 mutation at leucine-rich repeat domain (position 2776 C>T). Mutations at these functional domains may have impacts on the signaling pathway of Card15-coded NOD2 protein.

4. CONCLUSION

This study has investigated the synonymous codon usage patterns and the mutation distribution of Card15 gene using a computational approach. It was found that mutational bias is the factor that drives the synonymous codon usage bias in Card15 gene. However, high ENC value implies that though GC3s drives the preferential usage of synonymous codon in Card15 gene, the usage bias is quite low. The obtained negative CBI value further corroborates the fact that synonymous codon usage is less biased in Card15 gene. Our use of COSMIC database demonstrates that mutations occur at the functional domains of Card15 gene. Future work is required to examine the impact of these mutations on the NOD2 signaling pathways and the pathogen immuno-surveillance.

REFERENCES

[1] L.D' Andrea, R.M. Pintó, A. Bosch, H. Musto, and J. Cristina, "A detailed comparative analysis on the overall codon usage patterns in Hepatitis A virus," *Virus Research*, Vol. 157, 2011, pp. 19-24.

[2] T. Zhou, W. Gu, J. Ma, X. Sun, and Z. Lu, "Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses," *BioSystems*, Vol. 81, 2005, pp. 77-86.

[3] Y. Li, C. Wang, X. Cheng, T. Wu, and C. Zhang, "Synonymous codon usage of the VP2 gene of a very virulent infectious bursal disease virus isolate serial passaged in chicken embryos," *BioSystems*, Vol. 104, 2011, pp. 42-47.

[4] P. Jiang, X. Sun, and Z. Lu, "Analysis of synonymous codon usage in *Aeropyrum pernix* K1 and other *Crenarchaeota* microorganisms," *Journal of Genetics and Genomics*, Vol. 34 (3), 2007, pp. 275-284.

[5] R. Hershberg and D.A. Petrov, "Selection on codon bias," *Annual Review of Genetics*, Vol. 42, 2008, pp. 287-299.

[6] K-N. Zhao, W.J. Liu, and I.H. Frazer, "Codon usage bias and A + T content variation in human papillomavirus genomes," *Virus Research*, Vol. 98, 2003, pp. 95-104.

[7] I. Chanda, A. Pan, S.K. Saha, and C. Dutta, "Comparative codon and amino acid composition analysis of Tritryps-conspicuous

features of *Leishmania major*," *FEBS Letters*, Vol. 581, 2007, pp. 5751-5758.

[8] A. Fuglsang, "Patterns of context-dependent codon biases," *Biochemical and Biophysical Research Communications*, Vol. 304, 2003, pp. 86-90.

[9] Y. Huang, E.V. Koonin, D.J. Lipman, and T.M. Przytycka, "Selection for minimization of translational frameshifting errors as a factor in the evolution of codon usage," *Nucleic Acids Research*, Vol. 37, No. 20, 2009, pp. 6799-6810.

[10] P.M. Sharp, E. Bailes, R.J. Grocock, J.F. Peden, and R.E. Sockett, "Variation in the strength of selected codon usage bias among bacteria," *Nucleic Acids Research*, Vol. 33, 2005, pp. 1141-1153.

[11] E.P.C. Rocha, "Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization," *Genome Research*, Vol. 14, 2004, pp. 2279-2286.

[12] D.J. Lynn, G.A. Singer, and D.A. Hickey, "Synonymous codon usage is subject to selection in thermophilic bacteria," *Nucleic Acids Research*, Vol. 30, 2002, pp. 4272-4277.

[13] M. dos Reis, L. Wernisch, and R. Savva, "Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome," *Nucleic Acids Research*, Vol. 31, 2003, pp. 6976-6985.

[14] P. Tao, L. Dai, M. Luo, F. Tang, P. Tien, and Z. Pan, "Analysis of synonymous codon usage in classical swine fever virus," *Virus Genes*, Vol. 38, 2009, pp. 104-112.

[15] V. Scaiewicz, V. Sabbía, R. Piovani, and H. Musto, "CpG islands are the second main factor shaping codon usage in human genes," *Biochemical and Biophysical Research Communications*, Vol. 343, 2006, pp. 1257-1261.

[16] G.M. Jenkins, M. Pagel, E.A. Gould, P.A. Zanotto, E.C. Holmes, "Evolution of base composition and codon usage bias in the genus Flavivirus," *Journal of Molecular Evolution*, Vol. 52, 2001, pp. 383-390.

[17] M. Stenico, A.T. Lloyd, and P.M. Sharp, "Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases," Nucleic Acids Research, Vol. 22, 1994, pp. 2437-2446.

[18] P.J. Murray, "NOD proteins: an intracellular pathogen-recognition system or signal transduction modifiers?" *Current Opinion in Immunology*, Vol. 17, 2005, pp. 352-358.

[19] A. Williams, R.A. Flavell, and S.C. Eisenbarth, "The role of NOD-like receptors in shaping adaptive immunity," *Current Opinion in Immunology*, Vol. 22, 2010, pp. 34-40.

[20] M. Saleh, "The machinery of Nod-like receptors: refining the paths to immunity and cell death," *Immunological Reviews*, Vol. 243, 2011, pp. 235-246.

[21] M.H. Shaw, T. Reimer, Y-G. Kim, and G. Nuñez, "NOD-like receptors (NLRs): bona fide intracellular microbial sensors," *Current Opinion in Immunology*, Vol. 20, 2008, pp. 377-382.

[22] O. Takeuchi and S. Akira, "MDA5/RIG-I and virus recognition," *Current Opinion in Immunology*, Vol. 20, 2008, pp. 17-22.

[23] W. Ouyang, S. Rutz, N.K. Crellin, P.A. Valdez, and S.G. Hymowitz, "Regulation and functions of the IL-10 family of cytokines in inflammation and disease," *Annual Review of Immunology*, Vol. 29, 2011, pp. 71-109.

[24] S. Vallabhapurapu and M. Karin, "Regulation and function of NF-κB transcription factors in the immune system," *Annual Review of Immunology*, Vol. 27, 2009, pp. 693-733.

[25] Y. Qiao, P. Wang, J. Qi, L. Zhang, C. Gao, "TLR-induced NF-κB activation regulates NLRP3 expression in murine macrophages," *FEBS Letters*, Vol. 586, 2012, pp. 1022-1026.

[26] J. Hiscott, "Convergence of the NF-κB and IRF pathways in the regulation of the innate antiviral response," *Cytokine & Growth Factor Reviews*, Vol. 18, 2007, pp. 483-490.

[27] P.M. Sharp, T.M. Tuohy, and K.R. Mosurski, "Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes," *Nucleic Acids Research*, Vol. 14, No. 13, 1986, pp. 5125-5143.

[28] F. Wright, "The effective number of codons used in a gene," *Gene*, Vol. 87, 1990, pp. 23-29.

[29] J.L. Bennetzen and B.D. Hall, "Codon selection in yeast," *Journal of Biological Chemistry*, Vol. 257, 1982, pp. 3026-3031.

[30] S.A. Forbes, N. Bindal, S. Bamford, C. Cole, C.Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J.W. Teague, P.J. Campbell, M. Stratton, and P.A. Futreal, "COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer," *Nucleic Acids Research*, Vol. 39, 2011, pp. D945-D950.

[31] K. Sau, S.K. Gupta, S. Sau, and T.C. Ghosh, "Synonymous codon usage bias in 16 *Staphylococcus aureus* phages: Implication in phage therapy," *Virus Research*, Vol. 113, 2005, pp. 123-131.

[32] W. Gu, T. Zhou, J. Ma, X. Sun, Z. Lu, "Analysis of synonymous codon usage in SARS *Coronavirus* and other viruses in the *Nidovirales*," *Virus Research*, Vol. 101, 2004, pp. 155-161.

[33] H. Romero, A. Zavala, H. Musto, and G. Bernardi, "The influence of translational selection on codon usage in fishes from the family Cyprinidae," *Gene*, Vol. 317, 2003, pp. 141-147.

[34] S. Zhao, Q. Zhang, X. Liu, X. Wang, H. Zhang, Y. Wu, and F. Jiang, "Analysis of synonymous codon usage in 11 human bocavirus isolates," *BioSystems*, Vol. 92, 2008, pp. 207-214.

[35] S. Das, S. Paul, C. Dutta, "Synonymous codon usage in adenoviruses: influence of mutation, selection and protein hydropathy," *Virus Research*, Vol. 117, 2006, pp. 227-236.

[36] T. Soussi, "Advances in carcinogenesis: A historical perspective from observational studies to tumor genome sequencing and TP53 mutation spectrum analysis," *Biochimica et Biophysica Acta*, Vol. 1816, 2011, pp. 199-208.

[37] H. Nilsen, Q. An, and T. Lindahl, "Mutation frequencies and AID activation state in B-cell lymphomas from Ung-deficient mice," *Oncogene*, Vol. 24, 2005, pp. 3063-3066.

[38] R. Vinall, C.G. Tepper, X-B. Shi, L.A. Xue, R. Gandour-Edwards, and R.W. White, "The R273H p53 mutation can facilitate the androgen-independent growth of LNCaP by a mechanism that involves H2 relaxin and its cognate receptor LGR7," *Oncogene*, Vol. 25, 2006, pp. 2082-2093.

[39] M. Ferreira, H. Fujiwara, K. Morita, and F.M. Watt, "An activating β1 integrin mutation increases the conversion of benign to malignant skin tumors," *Cancer Research*, Vol. 69, 2009, pp. 1334-1342.

[40] C. Gundacker, M. Gencik, and M. Hengstschläger, "The relevance of the individual genetic background for the toxicokinetics of two significant neurodevelopmental toxicants: Mercury and lead," *Mutation Research*, Vol. 705, 2010, pp. 130-140.

[41] J. Guergnon and C. Combadière, "Role of chemokines polymorphisms in diseases," *Immunology Letters*, Vol. 145, 2012, pp. 15-22.

[42] W-S. Choe, H-L. Kim, J-K. Han, Y-E. Choi, B. Seo, H-J. Cho, H-K. Yang, K-J. Park, J-S. Park, H-J. Park, P-J. Kim, S-H. Baek, K-B. Seung, and H-S. Kim, "Association between OPG, RANK and RANKL gene polymorphisms and susceptibility to acute coronary syndrome in Korean population," *Journal of Genetics*, Vol. 91, 2012, pp. 87-89.