International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 **Volume 3 Number 12 (2014) pp. 8-15**
http://www.ijcmas.com

IJCMAS
INTERNATIONAL JOURNAL OF CURRENT MICROBIOLOGY AND APPLIED SCIENCES
2014 NOVEMBER
www.ijcmas.com

Original Research Article

In silico comparison of nucleotide composition and codon usage bias between the essential and non- essential genes of *Staphylococcus aureus* NCTC 8325

Prosenjit Paul, Tarikul Huda Mazumder and Supriyo Chakraborty*

Department of Biotechnology, Assam University, Silchar-788011, Assam, India
**Corresponding author*

ABSTRACT

Keywords

Codon usage bias, Synonymous codon, *Staphylococcus aureus*, **S**election, Essential gene

It is well known that the synonymous codons are not used uniformly; certain synonymous codons are used preferentially, a phenomenon called codon usage bias (CUB). We estimated the possible codon usage variation in the essential and non-essential genes of *Staphylococcus aureus* NCTC 8325 by analyzing % GC at third codon position and effective number of codons. We further analyzed the pattern of amino acid usage. Our result indicates that there is almost no variation in codon usage with a little variation of amino acid usage between essential and non essential genes of *Staphylococcus aureus* NCTC 8325. We also investigated the relationship between the differences in nucleotide composition and the level of gene expression. Our results showed that though the base usage in all three codon positions reveals a selection-mutation balance, the codons starting with A and ending with T are used optimally by the essential genes and the codons starting with T and ending with A are used optimally by the non essential genes of *Staphylococcus aureus* NCTC 8325.

Introduction

Gene expression is a fundamental cellular process by which proteins are synthesized in a cell. Codon bias is the probability that a given codon will be used to code for an amino acid over a different codon which codes for the same amino acid. In unicellular organisms, a strong correlation was observed between gene expressivity and the extent of codon bias from studies, mostly on *Escherichia coli* and *Saccharomyces cerevisiae*, on codon usages in genes expressed at different levels (Bennetzen and Hall, 1982; Gouy and Gautier, 1982; Sharp and Li, 1986). Molecular evolutionary investigations suggest that codon usage varies at three levels: between genomes, between genes in the same genome, and within a single gene (Hooper and Berg, 2000). The codon bias is most prominent in highly expressed genes, exhibits a strong preference for a subset of synonymous codons (Sharp and Li, 1987). Such codons are often referred to as optimized codons or preferred codons. Strength of codon usage bias in a gene can be used to make predictions about its expression level.

Fundamental processes that affect codon usage include translational selection related to gene expression levels (Sharp and Li, 1987; Akashi, 1994), directional mutational biases (Bernardi, 1985; D'Onofrio *et al.*, 1991) and distinct mutational spectra on the leading or lagging strands of replication (McInerney, 1998).

With the increasing number of fully sequenced genomes, it is now possible to study the distinct genome patterns that reflect these varying evolutionary pressures. Synonymous codon usage biases had been determined in numerous living organisms and in viruses (Hershberg and Petrov, 2008; Sharp *et al.*, 2010; Anhlan *et al.*, 2011; Liu *et al.*, 2012). Codon usage analysis has enriched both the basic and applied biological sciences in a number of ways; supported in understanding the expression levels of the genes, provided clues about the evolution of the genes and genomes, maximized protein expression in the heterologous system and enhanced immunogenicity of vaccines, etc (Crameri *et al.*, 1996; Deml *et al.*, 2001; Geddie and Matsumura 2004). *Staphylococcus aureus*, a gram-positive bacterium, causes various mild as well as life-threatening diseases to human and other animals. The complete genome sequences of the fifty one staphylococcal phages are available, out of which forty phages were investigated. These investigations identified the virulent phages having prospects in therapy (Bishal *et al.*, 2012). Codon usage bias (CUB) is one of the useful analytical methods for comparative genomics, provides the gateway to understand the evolutionary and functional relations between different species also between the genes from the same species.

Currently, very little is known about the codon usage bias and pattern of usage between the essential and non essential genes of *Staphylococcus aureus NCTC 8325*. In the present study, we performed codon usage analysis in the essential and non essential genes of *Staphylococcus aureus* NCTC 8325 by analyzing the codon adaptation index (CAI), relative codon usage bias (RCBS), effective number of codons (ENc), codon deviation coefficient (CDC), synonymous codon usage order (SCUO), GC content and also the skewness value for GC, AT, purine, pyrimidine, keto and amino contents. Our results revealed that both the essential and non essential genes maintains the same pattern of synonymous codon usage, but differs in GC percent, skewness value for AT, GC, pyrimidine and amino group content when compared with the level of gene expression, indicating that these aforementioned parameters could be used to identify the essential genes for *Staphylococcal* phages.

Materials and Methods

The coding sequences (cds) for the essential and non essential genes of *Staphylococcus aureus* NCTC 8325 were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/) and DEG (http://tubic.tju.edu.cn/deg/). To minimize sampling errors we have taken only those cds which are greater than or equal to 300bp and have the correct initial and termination codons, devoid of N (any unknown base) and intercalary stop codons. Finally, thirty four (34) cds sequences from each group (essential and non essential gene) were selected for the present study.

RCBS is the overall score of a gene indicating the influence of RCB of each codon in a gene. RCB reflects the level of gene expression. RCBS was calculated as per Roymondal *et al.*, (2009) by using the formula

$$RCBS = (\prod_{i=1}^{L}(1 + d^{i}_{xyz})^{1/L} - 1$$

Where, d^{i}_{xyz} is the codon usage difference of i^{th} codon of a gene. L is the number of codons in the gene. Gene expressivity was again measured by calculating the CAI as per Sharp and Li (1986, 1987). CAI was calculated as

$$CAI = \exp \frac{1}{L}\sum_{k=1}^{L} \ln w_{c(k)}$$

Where, L is the number of codons in the gene and $w_{c(k)}$ is the ω value for the kth codon in the gene. ENc is the total number of different codons used in a sequence. The values of ENc for standard genetic code range from 20 (where only one codon is used per amino acid) to 61(where all possible synonymous codons are used with equal frequency). ENc measures bias toward the use of a smaller subset of codons, away from equal use of synonymous codons. For example, as mentioned above, highly expressed genes tend to use fewer codons due to selection. ENc value for the cds was calculated as per Wright (1990) as given below

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

Where, F_k (k = 2, 3, 4 or 6) is the average of the F_k values for k-fold degenerate amino acids. The F value denotes the probability that two randomly chosen codons for an amino acid with two codons are identical.

The measure of codon usage, SCUO of genes was computed as per Wan *et al.* (2004). Then the average SCUO for each sequence was calculated by using the formula:

$$O = \sum_{i=1}^{n_i} F_i O_i$$

Where, F_i is the composition ratio of the ith amino acid in each sequence and O_i is the synonymous codon usage order (SCUO) for the ith amino acid in each sequence.

CDC considers both GC and purine contents as background nucleotide composition (BNC) and derives expected codon usage from observed positional GC and purine contents. CDC adopts the cosine distance metric to quantify CUB and employs the bootstrapping to assess its statistical significance, requiring no prior knowledge of reference gene sets. The CDC value for each of the cds sequence was calculated using the algorithm given by Zhang *et al.* (2012). GC3s is the frequency of (G+C) and A3s, T3s, G3s, and C3s are the distributions of A, T, G and C bases at the third codon position (Chaudhuri, Allen *et al.* 2009). We have measured the correlations between all the above mentioned parameters and codon position. Skewness values for the AT, GC, purine, pyrimidine, keto and amino content was analyzed by perl script (program) developed by SC (corresponding author).

Results and Discussion

Codon usage bias analysis

For the present study we have selected a total of sixty eight genes from *Staphylococcus aureus* NCTC 8325 out of which thirty four genes were found to be essential and the remaining genes are non essential for the bacteria (Chaudhuri *et al.*, 2009). Selected genes with their accession number along with the overall RCBS, CAI, SCUO, CDC, ENc, percentage of AT, GC, GC1, GC2, GC3 and also the skewness values for the AT, GC, purine, pyrimidine, amino and keto content are given in the supplementary file. It was found that the cds of *Staphylococcus aureus* NCTC 8325 are rich in A and/or T. However, the overall codon usage values may obscure some heterogeneity of codon usage bias among the genes that might be superimposed on the extreme genomic composition of this organism.

The variation of codon usage biases was measured by calculating the mean and standard deviation of ENc for the two groups of genes. The mean and standard deviation values were found to be 45.77, 2.88 for essential genes and 46.05, 2.94 for the non essential genes, respectively. This result indicates that there is almost no variation of codon usage bias among the genes of *Staphylococcus aureus* NCTC 8325. We compared the usage pattern of all the sixty one synonymous codons for the two groups of genes shown in Figure 1. It was found that both the group maintains a specific pattern of codon usage. Moreover, we analyzed the pattern of amino acid usage; our analysis revealed that the usage of amino is same for the two groups except aspartic acid, glutamic acid, arginine and valine, which are used more frequently among the essential genes shown in Figure 2. Our result indicates that there is almost no variation in codon usage with a little variation of amino acid usage between essential and non essential genes of *Staphylococcus aureus* NCTC 8325.

We further analyzed synonymous codon usage orders of the genes. SCUO is a relatively easier approach compared to RSCU and is considered as more robust for comparative analysis of codon usage. It was found that 70.58% of essential and 52.94% of non essential genes from *Staphylococcus aureus* NCTC 8325 have SCUO values greater than 0.3, suggesting that the majority of the genes selected for our present study are associated with high codon usage bias. CDC values analyzed for the genes showed that the biasness of codon usage is greater for the non essential (standard deviation 0.008) genes than that of essential genes (standard deviation 0.014).

Codon composition and its relationship with gene expressivity

The codon composition of coding sequences plays an important role in the regulation of gene expression. We analyzed the codon composition for the cds sequence of both the groups. A survey of the literature indicated that the value of RCBS depends on cds length; CAI was used as a prime measure for expressivity analysis. The correlation coefficients of the frequencies of four bases in three codon positions against the CAI values for the two groups of genes have been estimated. In *Staphylococcus aureus* NCTC 8325, the frequency of G at the second codon positions and that of C at the third codon position showed highest correlation with CAI for both the groups. But the frequencies of A and T showed the differences in correlation analysis with CAI. The frequency of A at first and T at the third codon position, respectively showed highest correlation with CAI ($r=0.834$, $p<0.01$; $r=0.828$, $p<0.01$) for the essential genes. In case of non-essential genes of *Staphylococcus aureus* NCTC 8325 the frequency of A at third and T at the first codon position, respectively showed highest correlation with CAI ($r=0.7$, $p<0.01$; $r=0.79$, $p<0.01$). We again measured the relationship between the aforementioned genetic indices with gene expressivity for the two groups of gene shown in Fig. III.

The present investigation highlights the codon usage and nucleotide composition variation between the essential and non essential genes of *Staphylococcus aureus* NCTC 8325. To study the codon usage variation the nucleotide composition, CAI, RCBS, CDC, ENc, SCUO values for the sixty eight protein-coding genes of the phage was determined. With reference to the other *Staphylococcus* phages which are AT-rich, A and T ending codons also appeared to be predominant for *Staphylococcus aureus* NCTC 8325 (Bishal *et al.,* 2012). We analyzed the possible codon usage variation in the two group of genes, GC3s was determined and found to vary from 13.93 to 20.59 (with a mean of 17.08 and standard deviation 1.64) for essential gene and from 11.38 to 18.14 (with a mean of 16.10 and standard deviation 1.45) for non essential genes, respectively. We also predicted the heterogeneity of codon usage by analyzing ENc. The standard deviation and standard error value of ENc were found to be 2.88, 0.49 for essential genes and 2.94, 0.5 for the non essential genes, respectively.

These results are indicative of the presence of unique pattern of synonymous codon usage among the genes of *Staphylococcus aureus* NCTC 8325. We analyzed the frequency of occurrence of all the synonymous codons for the two groups of genes shown in the Fig. I. It was found that the pattern of codon usage for all the synonymous codons is relatively similar between the essential and non essential genes of *Staphylococcus aureus* NCTC 8325. Moreover, the amino acid frequency which is defined as the total frequency of their respective codons was also analyzed as shown in the Fig. II. Our data showed that the frequency is almost the same in both the groups but that of Asparagine (N), Glutamic acid (E), Arginine (R) and Valine (V) are varied.

The frequency of amino acids N, E, R & V have dissimilar pattern of distribution in essential and non-essential genes. These amino acids are used in higher frequency for the essential genes. Further analysis has to be carried out to find out whether these amino acids play crucial role in *Staphylococcus aureus* or these variations are due to random probability, by taking all Staphylococcal phages into consideration. To examine the possible contribution of translational selection in addition to mutational bias in dictating synonymous codon usage, we analyzed the codon usage patterns, individual base frequencies at three codon positions and their correlations with expression levels for both group of genes. It is evident from the present analysis that both the essential and non essential genes exhibit similar codon usage profiles in the second and third codon positions for G and C, respectively when compared with the expression level.

Our analysis revealed that the overall AT percentage, GC percentage, GC1, GC3, skewness value for GC, pyrimidine and amino content showed different pattern for the two groups when allied with the expressivity values. It has been found that CAI increases with the increase of AT content, and the skewness values of pyrimidine and amino content within the cds sequence for the essential genes. But in case of the non-essential gene of *Staphylococcus aureus* NCTC 8325 CAI increases with the increase of GC percentage, GC1, GC3, skewness value for GC content. This implies that though the base usage in all three codon positions reveals a selection- mutation balance, the codons starting with A and ending with T are used optimally by the essential genes but the codons starting with T and ending with A are used optimally by the non essential genes of *Staphylococcus aureus* NCTC 8325.

Fig.1 Comparison of synonymous codon usage between essential and non essential genes of *Staphylococcus aureus* NCTC 8325

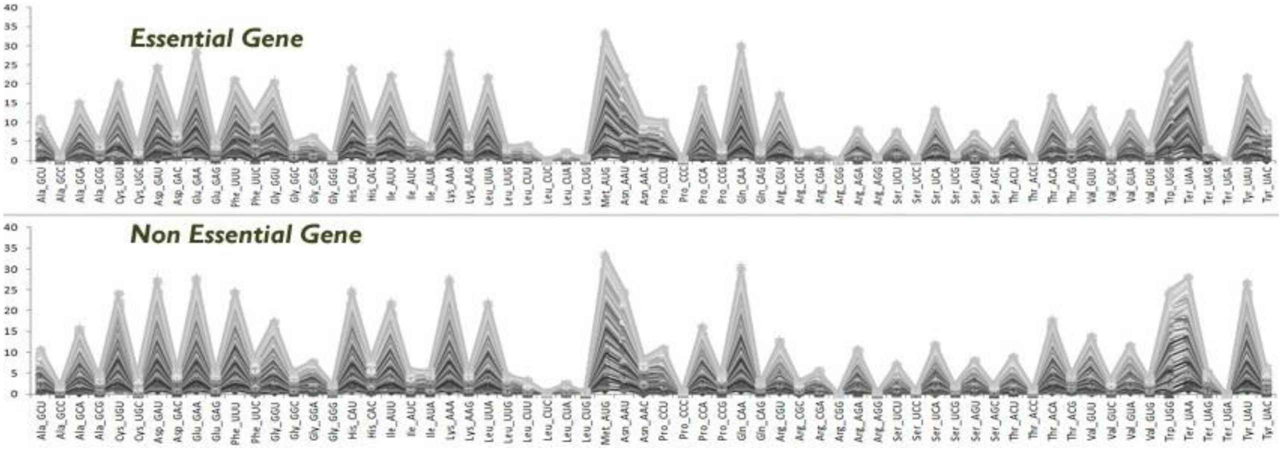


Fig.2 Comparison of amino acid codon usage between essential and non essential genes of *Staphylococcus aureus* NCTC 8325

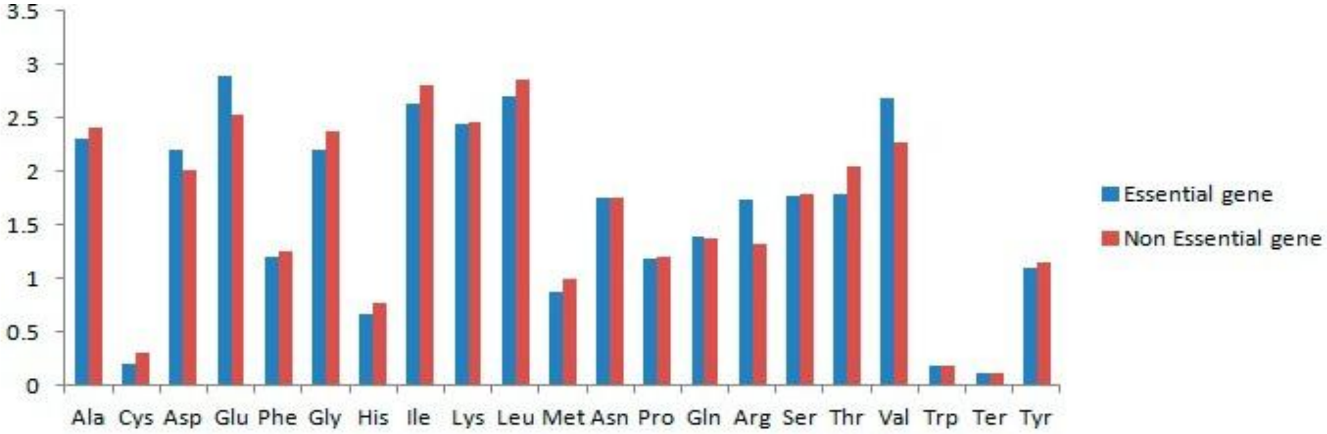
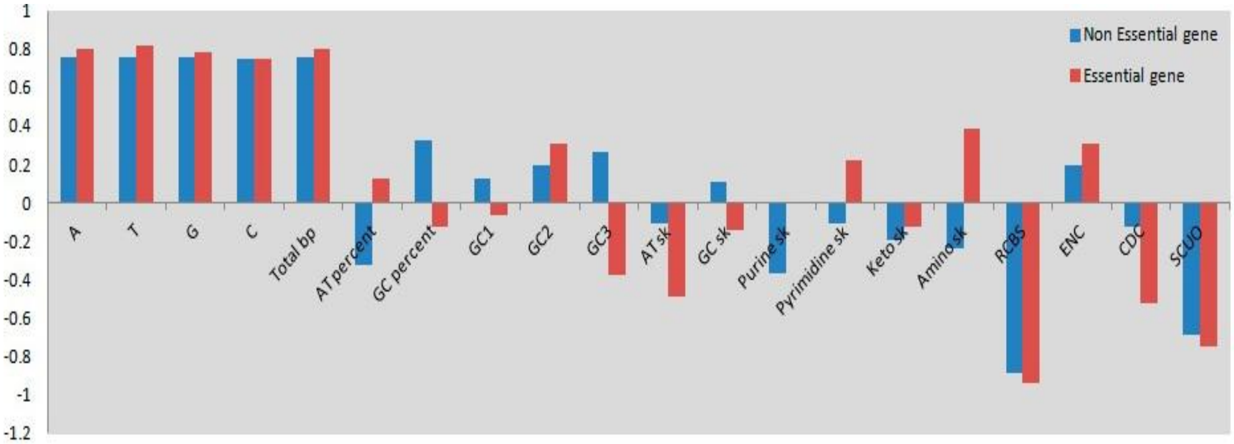


Fig.3 Relationship between different parameters used for codon bias analysis and gene expressivity for the genes of *Staphylococcus aureus* NCTC 8325



References

Akashi, H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics,* 136(3): 927–935.

Anhlan, D., Grundmann, N., *et al.* 2011. Origin of the 1918 pandemic H1N1 influenza A virus as studied by codon usage patterns and phylogenetic analysis. *RNA,* 17(1): 64–73.

Bennetzen, J.L., Hall, B.D. 1982. Codon selection in yeast. *J. Biol. Chem.,* 257(6): 3026–3031.

Bernardi, G. 1985. Codon usage and genome composition. *J. Mol. Evol.,* 22(4): 363–365.

Bishal, A.K., Saha, S., *et al.* 2012. Synonymous codon usage in forty *Staphylococcal* phages identifies the factors controlling codon usage variation and the phages suitable for phage therapy. *Bioinformation,* 8(24): 1187–1194.

Chaudhuri, R.R., Allen, A.G., *et al.* 2009. Comprehensive identification of essential *Staphylococcus aureus* genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC Genomics,* 10: 291.

Crameri, A., Whitehorn, E.A., *et al.* 1996. Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat. Biotechnol.,* 14(3): 315–319.

Deml, L., Bojak, A., *et al.* 2001. Multiple effects of codon usage optimization on expression and immunogenicity of DNA candidate vaccines encoding the human immunodeficiency virus type 1 Gag protein. *J. Virol.,* 75(22): 10991–11001.

D'Onofrio, G., Mouchiroud, D., *et al.* 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.,* 32(6): 504–510.

Geddie, M.L., Matsumura, I. 2004. Rapid evolution of beta-glucuronidase specificity by saturation mutagenesis of an active site loop. *J. Biol. Chem.,* 279(25): 26462–26468.

Gouy, M., Gautier, C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.,* 10(22): 7055–7074.

Hershberg, R., Petrov, D.A. 2008. Selection on codon bias. *Ann. Rev. Genet.,* 42: 287–299.

Hooper, S.D., Berg, O.G. 2000. Gradients in nucleotide and codon usage along *Escherichia coli* genes. *Nucleic Acids Res.,* 28(18): 3517–3523.

Liu, H., Huang, Y., *et al.* 2012. Patterns of synonymous codon usage bias in the model grass Brachypodium distachyon. *Genet. Mol. Res.,* 11(4): 4695–4706.

McInerney, J.O. 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc *Natl. Acad. Sci. USA.,* 95(18): 10698–10703.

Roymondal, U., Das, S., *et al.* 2009. Predicting gene expression level from relative codon usage bias: an application to *Escherichia coli* genome. *DNA Res.,* 16(1): 13–30.

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

Sharp, P.M., Emery, L.R., *et al.* 2010. Forces that influence the evolution of codon bias. *Philos. Trans. R. Soc. Lond. B Biol. Sci.,* 365(1544): 1203–1212.

Sharp, P.M., Li, W.H. 1987. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications.

Nucleic Acids Res., 15(3): 1281–1295.

Wan, X.F., Xu, D., *et al.* 2004. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC. Evol. Biol.,* 4: 19.

Wright, F. 1990. The effective number of codons used in a gene. *Gene,* 87(1): 23–29.

Zhang, Z., Li, J., *et al.* 2012. Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. *BMC. Bioinformatics.,* 13: 43.