Codon Usage Bias Prefers AT Bases in Coding Sequences Among the Essential Genes of Haemophilus influenzae

Chakraborty SUPRIYO*, Paul PROSENJIT, Tarikul Huda MAZUMDER

Abstract

The base composition at three different codon positions in relation to codon usage bias and gene expressivity was studied in a sample of twenty five essential genes from Haemophilus influenzae. ENC, CBI and Fop were used to quantify the variation in codon usage bias for the cds. CAI is used to estimate the level of gene expression of the cds selected in the present study. To find out the relationship between the extent of codon bias and nucleotide composition the values of A, T, G, C and GC they were compared with the A3, T3, G3, C3 and GC3 values, respectively. The results showed relatively weak codon usage bias among the coding sequences (cds) of Haemophilus influenzae. This in turn, implies that the essential genes prefer to use a set of restricted codons. However, the base compositional analysis of essential genes in Haemophilus influenzae revealed preference of AT to GC bases within their coding sequences and this preference might affect gene expression as indicated by the relatively high CAI values of the coding sequences.

Keywords: codon usage pattern, gene expression, synonymous codons

Introduction

The genetic information from mRNA is passed to protein with the help of translational machinery, a fundamental process occurring within all living cells. The genetic code uses sixty-four codons to encode the proteins. Many codons are redundant i.e. two or more codons code for a single amino acid, with the exception of methionine (AUG) and tryptophan (UGG). Such codons are described as being synonymous, and mostly differ by one nucleotide at the third codon position. Codon usage bias (CUB) refers to the non-random usage of synonymous codons for encoding the same amino acid in a protein (Lu et al., 2005). In general, synonymous codons are used in unequal frequencies between genomes, genes from the same genome and within a single gene (Hooper and Berg, 2000; Lavner and Kotler, 2005; Supak and Vlahoviczek, 2005). The most frequently used codons are termed as optimal or major codons, recognized by the most abundant tRNA species (Butt et al., 2014). The least frequently used codons are non optimal or minor codons and lead to the premature termination during elongation stage of translation. Thus, CUB is accounted for by the positive selection for the optimal codons and the negative selection against the non optimal codons (Qin et al., 2004). CUB is a unique property, that shows species-specific deviation (Grantham et al., 1981). Thus the understanding of the extent of codon bias with compositional dynamics provides an insight into the prediction of the level of gene expression and genome characterization.

Advances in sequencing technology have provided an abundance of genomic data from different organisms. The study of CUB is gaining rehabilitated attention with the advent of whole genome sequencing of numerous organisms. Within a genome some genes are conserved and provide critical support to the organism, and these are termed as essential genes. The functions encoded by these genes are considered as a foundation of life itself. Haemophilus influenzae is a gram negative bacterium, highly adapted to its human host. The entire genome sequence of Haemophilus influenzae was completed and published in 1995. In the present work, the purpose was to study the CUB by analyzing the codon adaptation index (CAI), codon bias index (CBI), frequency of optimal codon (Fop), effective number of codons (ENC) and compositional dynamics for the essential genes of Haemophilus influenzae.

Materials and methods

Sequence data

For the present study there were selected genes that are essential for the growth or survival of Haemophilus influenzae. For CUB and compositional analyses were selected only those coding sequences (cds) having perfect initiator codon, terminator codon, devoid of any unknown bases (N) and are perfect multiple of three bases. The final data set consisted of twenty five (25) cds from
**Haemophilus influenzae.** Complete nucleotide coding sequence for each of the concerned gene satisfying the aforementioned criteria was retrieved from NCBI nucleotide database.

**Models**

ENC, CBI and Fop were used to quantify the variation in codon usage bias for the cds. ENC is generally used to measure the codon usage bias of a gene that is independent of the gene length and the number of amino acids (Wright, 1990). The ENC value ranges from 20-61. Fop is used to measure the codon usage bias in a gene (Zhou et al., 2005). The Fop value ranges from 0.36 for a gene in which codon usage pattern is uniform to 1 for a gene in which codon usage is highly biased (Zhou et al., 2005). Similar to Fop, CBI also measures the extent to which preferred codons are used in a gene (Bennetzen and Hall, 1982). CBI value is normalized between -1 and 1. Thus, CBI value of 1 means only preferred codons are used, zero means random choice and less than zero implies greater use of non preferred codons (Bennetzen and Hall, 1982).

CBI is used to estimate the level of gene expression of the cds selected in the present study. CBI measures the degree of bias towards the codons in highly expressed genes and thus assesses the effective selection which helps in shaping the codon usage pattern (Naya et al., 2001; Gupta et al., 2004). The value of CBI ranges from 0 to 1. For a gene in which all synonymous codons are used equally, the CBI would be 0 indicating no bias and if optimal codons are used, the value will be 1 for the strongest codon bias (Stenico et al., 1994).

GC content at three codon positions i.e. GC1, GC2, GC3 was calculated to quantify the relationship between codon usage variations. GC content at the three codon position is presumed to be a good indicator of base composition bias (Zhou et al., 2005).

### Analysis tools

All the above mentioned parameters except CAI for the CUB and compositional analysis were carried out by using the online tool CodonW. The CAI value for each of these cds was computed by using the software acua available for non-commercial purposes.

### Results

The coding sequence of 25 essential genes was retrieved and their nucleotide composition bias and codon usage bias were analyzed. The results of the compositional analysis with the accession numbers for each cds are given in Tab. 1. To find out the relationship between the extent of codon bias and nucleotide composition the values of A, T, G, C and GC were compared with the A3, T3, G3, C3 and GC3 values, respectively. Highly significant positive correlations were observed between GC3 and A (r=0.926, p<0.01), GC3 and T (r=0.933, p<0.01), GC3 and G (r=0.973, p<0.01), GC3 and C (r=0.966, p<0.01), GC3 and GC (r=0.984, p<0.01) contents. Simultaneously, GC1, GC2 and GC3 values were calculated for each gene to investigate the relationship between codon usage variation and compositional constraints. The GC3% of the cds was in the range between 13.7 to 21 with a mean of 16.86 and standard deviation of 1.78. Moreover, when comparing the GC content at first codon position (GC1) and second codon position (GC2) with that of the third codon position (GC3) (Fig. 1) a striking positive correlation (r=0.968, p<0.01) was observed, indicating that the patterns of base compositions are most likely the result of mutation pressure rather than that of natural selection, since at all codon positions its effects are present.

**Tab. 1. Compositional analysis of Haemophilus influenzae genes with their accession numbers**

| Sl. No | Genes                                    | Accession Numbers | Nucleotide contents |
|-------|------------------------------------------|-------------------|--------------------|
| 1     | Glyceraldehyde-3-phosphate dehydrogenase | NP_438174.1       | A T G C A3 T3 G3 C3|
| 2     | Metalloprotease                          | NP_438177.1       |                   |
| 3     | Formate dehydrogenase accessory protein  | NP_438182.1       |                   |
| 4     | Ribosomnal-protein-alanine N-acetyltransferase | NP_438183.1 |                   |
| 5     | GTP-binding protein Era                  | NP_438186.1       |                   |
| 6     | Ribonuclease III                         | NP_438187.1       |                   |
| 7     | Signal peptidase I                       | NP_438188.1       |                   |
| 8     | Uracil-DNA glycosylase                   | NP_438191.1       |                   |
| 9     | Hypothetical protein HI0020              | NP_438193.1       |                   |
| 10    | Citrate lyase subunit alpha              | NP_438195.1       |                   |
| 11    | Citrate lyase subunit beta               | NP_438196.2       |                   |
| 12    | Lipoate-protein ligase B                 | NP_438200.1       |                   |
| 13    | Lipoprotein                              | NP_438203.1       |                   |
| 14    | Penicillin-binding protein 2             | NP_438205.1       |                   |
| 15    | Hypothetical protein HI0035              | NP_438208.1       |                   |
| 16    | Rod shape determining protein MscA       | NP_438211.1       |                   |
| 17    | Rod shape determining protein MscD       | NP_438212.1       |                   |
| 18    | Hypothetical protein HI0104              | NP_438213.1       |                   |
| 19    | Exonuclease III complete cds             | NP_438214.1       |                   |
| 20    | Pseudoottinin synthase-like protein      | NP_438215.1       |                   |
| 21    | Hypothetical protein HI0043              | NP_438216.1       |                   |
| 22    | Hypothetical protein HI0045              | NP_438218.1       |                   |
| 23    | D-mannonate oxidoreductase               | NP_438221.1       |                   |
| 24    | 2-dehydro-3-deoxyglucokinase             | NP_438222.1       |                   |
| 25    | Hypothetical protein HI0056              | NP_438229.1       |                   |
expression increases with the rise of Fop value (Fig. 5).

Wright (1990) suggested that a plot of ENC against GC3s could be effectively used to explore the codon usage variation among the genes (Wright, 1990). We computed the ENC values and compared with the frequency of GC3s (Fig. 2). A wide variation of CUB among the genes was observed indicating that the distribution of GC3 is not the only determinant; apart from GC3s, compositional constraints and other trends might influence the overall codon usage variation among the genes of Haemophilus influenzae. Moreover, we compared the ENC value and GC3 value with gene length to measure the relationship between codon usage bias and gene length. We have plotted ENC value against gene length and observed that the shorter genes had a much wider variation in ENC values and vice versa for longer genes (Fig. 3). The analysis showed a significant negative correlation (Pearson one tail, \( r = -0.368 \), \( p < 0.05 \)) between ENC and gene length, but significant positive correlation (Pearson one tail, \( r = 0.956 \), \( p < 0.01 \), \( G (r = 0.884, p < 0.01) \)), \( C (r = 0.863, p < 0.01) \) and GC (\( r = 0.888, p < 0.01 \)). Similarly, C3 had also significant positive correlation with C (\( r = 0.862, p < 0.01 \)), T (\( r = 0.816, p < 0.01 \)), A (\( r = 0.871, p < 0.01 \)), G (\( r = 0.841, p < 0.01 \)) and GC (\( r = 0.863, p < 0.01 \)). A3 had significant positive correlation with A (\( r = 0.957, p < 0.01 \)), T (\( r = 0.951, p < 0.01 \)), G (\( r = 0.906, p < 0.01 \)), C (\( r = 0.920, p < 0.01 \)) and GC (\( r = 0.926, p < 0.01 \)). Similarly, G3 had also significant positive correlation with G (\( r = 0.971, p < 0.01 \)), C (\( r = 0.951, p < 0.01 \)), T (\( r = 0.927, p < 0.01 \)), A (\( r = 0.890, p < 0.01 \)) and GC (\( r = 0.977, p < 0.01 \)).

Moreover, the indices GC1, GC2, GC3 and GC12 (average of GC1 and GC2) were computed for each gene to establish the relationship among three codon positions. Our results showed that the coding sequences of Haemophilus influenzae have a wide range of GC3 and this difference usually influences the neutral mutation bias, leading to different codon choice in each gene (Liu et al., 2012). Concurrently, a significant positive correlation of GC12 with GC3 (\( r = 0.968, P < 0.01 \)) and GC1 with GC3 (\( r = 0.955, p < 0.01 \)) were observed; this may be due to the intragenomic GC mutational bias affecting the GC contents at all codon positions.

Several studies revealed that both natural selection and mutation pressure account for codon usage variation among different organisms. If synonymous codon usage bias is affected only by mutational pressure, then the frequency of nucleotides A and T should be equal to that of C and G at the synonymous third codon position (Zhang et al., 2013). However, the current study revealed wide variations in the nucleotide A and T with C and G base compositions, signifying that other factors, such as natural selection, might also be the determining factor in shaping the synonymous codon usage pattern. In general, there is an inverse relationship between ENC and gene expression i.e. a lower ENC value indicates a higher codon usage preference and higher gene expression and vice versa (Wright, 1990). Our results revealed that the overall codon usage bias among different genes of Haemophilus influenzae is low, slightly biased and mainly affected by base composition.

The CAI value was found in the range 0.64 to 0.76 for the 25 cds, further indicating low codon usage bias and relatively high gene expression level, similar to the ENC analysis. More interestingly, a significant correlation was observed between ENC and CAI (\( r = -0.72, p < 0.01 \)) which indicates that the essential genes have a strong preference for a subset of codons, as shown by the positive correlation between optimal codons and tRNA abundance (Ikemura, 1981). Selection for translational accuracy is predicted to have a positive correlation between codon bias and gene length (Eyre-Walker, 1996). From the plot drawn with gene length against ENC (Fig. 3), it is

**Discussion**

The present study was taken up to analyze several widely used parameters of codon usage bias namely CAI, CBI, Fop and ENC along with the base composition of the coding sequences of some essential genes in Haemophilus influenzae. The accurate coding sequences were retrieved using a program in perl, developed by the researchers involved in the current study. After preliminary analysis of base composition it was found that the cds of Haemophilus influenzae are rich in AT.

A complex correlation was observed while investigating the relationship between different nucleotide constraints. T3 had a significant positive correlation with A (\( r = 0.831, p < 0.01 \)), T (\( r = 0.956, p < 0.01 \)), G (\( r = 0.884, p < 0.01 \)), C (\( r = 0.863, p < 0.01 \)) and GC (\( r = 0.888, p < 0.01 \)). Similarly, C3 had also significant positive correlation with C (\( r = 0.862, p < 0.01 \)), T (\( r = 0.816, p < 0.01 \)), A (\( r = 0.871, p < 0.01 \)), G (\( r = 0.841, p < 0.01 \)) and GC (\( r = 0.863, p < 0.01 \)). A3 had significant positive correlation with A (\( r = 0.957, p < 0.01 \)), T (\( r = 0.951, p < 0.01 \)), G (\( r = 0.906, p < 0.01 \)), C (\( r = 0.920, p < 0.01 \)) and GC (\( r = 0.926, p < 0.01 \)). Similarly, G3 had also significant positive correlation with G (\( r = 0.971, p < 0.01 \)), C (\( r = 0.951, p < 0.01 \)), T (\( r = 0.927, p < 0.01 \)), A (\( r = 0.890, p < 0.01 \)) and GC (\( r = 0.977, p < 0.01 \)).

Moreover, the indices GC1, GC2, GC3 and GC12 (average of GC1 and GC2) were computed for each gene to establish the relationship among three codon positions. Our results showed that the coding sequences of Haemophilus influenzae have a wide range of GC3 and this difference usually influences the neutral mutation bias, leading to different codon choice in each gene (Liu et al., 2012). Concurrently, a significant positive correlation of GC12 with GC3 (\( r = 0.968, P < 0.01 \)) and GC1 with GC3 (\( r = 0.955, p < 0.01 \)) were observed; this may be due to the intragenomic GC mutational bias affecting the GC contents at all codon positions.

Several studies revealed that both natural selection and mutation pressure account for codon usage variation among different organisms. If synonymous codon usage bias is affected only by mutational pressure, then the frequency of nucleotides A and T should be equal to that of C and G at the synonymous third codon position (Zhang et al., 2013). However, the current study revealed wide variations in the nucleotide A and T with C and G base compositions, signifying that other factors, such as natural selection, might also be the determining factor in shaping the synonymous codon usage pattern. In general, there is an inverse relationship between ENC and gene expression i.e. a lower ENC value indicates a higher codon usage preference and higher gene expression and vice versa (Wright, 1990). Our results revealed that the overall codon usage bias among different genes of Haemophilus influenzae is low, slightly biased and mainly affected by base composition.

The CAI value was found in the range 0.64 to 0.76 for the 25 cds, further indicating low codon usage bias and relatively high gene expression level, similar to the ENC analysis. More interestingly, a significant correlation was observed between ENC and CAI (\( r = -0.72, p < 0.01 \)) which indicates that the essential genes have a strong preference for a subset of codons, as shown by the positive correlation between optimal codons and tRNA abundance (Ikemura, 1981). Selection for translational accuracy is predicted to have a positive correlation between codon bias and gene length (Eyre-Walker, 1996). From the plot drawn with gene length against ENC (Fig. 3), it is

![Fig. 1. Correlation between GC content at first and second codon positions (GC1 & GC2) with that at third codon position (GC3). GC12 is the average of GC content at first and second codon positions](image_url)
Fig. 2. Frequency of GC3s and the distribution of ENC values of 25 selected coding sequences

Fig. 3. Plot of ENC versus gene length for the selected cds of *Haemophilus influenzae*

Fig. 4. Correlation between (a) ENC values and gene length (bp); (b) GC3 values and gene length (bp)
understood that shorter genes have a much wider variance in ENC values, but vice versa for longer genes. Furthermore, a significant negative correlation between ENC and gene length revealed that gene length is one of the important factors which influence codon usage in the cDNA of *Haemophilus influenzae*. Lower ENC values in longer genes may be due to the direct effect of translation time on fitness or the extra energy cost of proofreading associated with longer translation time (Hassan et al., 2009; Liu et al., 2003).

**Conclusion**

In brief, our analysis showed that overall codon usage bias in the essential genes of *Haemophilus influenzae* is slightly biased and mutation pressure played a major role to shape the codon usage pattern in these genes. In addition, the contributions of other factors such as compositional bias, translational forces and natural selection are also evident in shaping the codon usage pattern of *Haemophilus influenzae*. From the present analysis it can be concluded that the essential genes in *Haemophilus influenzae* preferred AT to GC bases (compositional analysis) within their coding sequences and this preference might affect gene expression as indicated by the relatively high CAI values of the cDNA.

**Acknowledgment**

We are thankful to Assam University, Silchar, Assam, India for providing the necessary facilities in carrying out this research work. No funding was received from any external funding agency like DBT or DST, Government of India.

**References**

Behura SK, Severson DW (2012). Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. PLoS One 10 1371 7:0043111.

Bennetzen JL, Hall BD (1982). Codon selection in yeast. J Biol Chem 257(6):3026-3031.

Butt AM, Nasrullah I, Tong Y (2014). Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. PLoS One 9:90905.

Eyre-Walker A (1996). Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol Biol Evol 13(6):864-872.

Granthon R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:43-74.

Gupta SK, Bhattacharyya TK, Ghosh TC (2004). Synonymous codon usage in Lactococcus lactis: mutational bias versus translational selection. J Biomol Struct Dyn 21(1):527-536.

Hassan S, Mahalingam V, Kumar V (2009). Synonymous codon usage analysis of thirty two mycobacteriophage genomes. Adv Bioinform 316936.

Hooper SD, Berg OG (2000). Gradients in nucleotide and codon usage along Escherichia coli genes. Nucleic Acids Res 28(18):3517-3523.

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(3):389-409.

Lavner Y, Koifman D (2005). Codon bias as a factor in regulating expression via translation rate in the human genome. Gene 345(1):127-38.

Liu H, Zhao WM, Zheng Y, Wang H, Qi M, Yu XP (2005). Analysis of synonymous codon usage bias in Chlamydia. Acta Biochim Biophys Sin (Shanghai) 37(1):1-10.

Liu H, Huang Y, Du X, Chen Z, Zeng X, Chen Y, Zhang H (2012). Patterns of synonymous codon usage bias in the model grass *Brachypodium distachyon*. Genet Mol Res 11(4):4695-4706.

Liu QP, Tan J, Xue QZ (2003). Synonymous codon usage bias in the rice cultivar 93-11 (*Oryza sativa* L. ssp. indica). Yi Chuan Xue Bao30(4):335-340.

Naya H, Romero H, Carles N, Zavala A, Musto H (2001). Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii*. FERS Lett 501(2-3):127-130.

Qin H, Wu WB, Comerón JM, Krügner M, Li WH (2004). Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics 168(4):2245-60.

Sternici M, Lloyd AT, Sharp PM (1994). Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res 22(13):2437-2446.

Supek F, Vlahovicek K (2005). Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformat 6:182.

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

Zhou T, Gu W, Ma J, Sun X, Lu Z (2005). Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. Bioinformatics 21:77-86.

Zhang Z, Dai W, Wang Y, Lu C, Fan H (2013). Analysis of synonymous codon usage patterns in torque teno sus virus 1 (TTSuV1). Arch Virol 158(1):145-154.

http://codonw.sourceforge.net

http://www.bioinsilico.com/acua