Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

Relative Analysis of Codon Usage and Nucleotide Bias between Anthrax Toxin Genes Subsist Inpxo1 Plasmid of *Bacillus Anthracis*

Sushma Bylaiah[1], Seema Shedole[1], Kuralayanapalya Puttahonnappa Suresh[2], Leena Gowda[3], Sharanagouda S Patil[4,*], Uma Bharathi Indrabalan[2] and Chandan Shivamallu[5,*]

[1]Department of Computer Science & Engineering, M S Ramaiah Institute of Technology, Matthikere, Bengaluru, Karnataka, INDIA.

[2]Spatial Epidemiology Laboratory, Indian Council for Agriculture Research - National Institute of Veterinary Epidemiology and Disease Informatics, Yelahanka, Bengaluru, Karnataka, INDIA.

[3]Department of Veterinary Public Health and Epidemiology, VeterinaryCollege,Hebbal, Bengaluru, Karnataka, INDIA.

[4]Virology Laboratory, Indian Council for Agriculture Research - National Institute of Veterinary Epidemiology and Disease Informatics, Yelahanka, Bengaluru, Karnataka, INDIA.

[5,*]Department of Biotechnology and Bioinformatics, Faculty of Life Sciences, JSS Academy of Higher Education & Research, Mysuru- 570 015, Karnataka, INDIA.

*Correspondence and requests for materials should be addressed to

Name: Dr. Chandan Shivamallu

Mobile: +91-9538500423

Email: chandans@jssuni.edu.in

ABSTRACT

Anthrax is an ancient and acuteillness that affects alarge quantity of animal species and is caused by a bacterium *Bacillus anthracis*, which is a rod-shaped, gram-positive and spore-forming bacterium. Virulent forms of *B.anthracis*has two large pathogenicity related plasmids pXO1 and pXO2. pXO1 has the different anthrax toxin genes *cya*, *lef*, and *pagA* where as pXO2 has the genes accountable for capsule synthesis and degradation, *capA*, *capB*, *capC*, and *capD*. *B. anthracis* express its pathogenic activity mostly over the capsule and the manufacture of a toxic compound involving three proteins known as edema factor (EF), lethal factor (LF) and protective antigen (PA). These two enormous plasmids of *B.anthracis*are crucial for full pathogenicity, exclusion of either of the plasmids extremely weakens the malignity of *B. anthracis.* In the current study we conducted the relative analysis of the codon usage and nucleotide bias of virulent genes subsist in pXO1 plasmid of *B.anthracis*. Codon usage bias not only plays a substantial role at the extent of gene expression, but also supports to improve the efficacy and accurateness of translation. Codon usage pattern analysis of *B.anthracis*genome is essential for understanding the evolutionary characteristicsin the different species. To examine the codon usage arrangement of the*B.anthracis*genome, Nucleotide sequences of the virulent genes *viz cya, lef and pag* were collected from National Center for Biotechnology Information (NCBI). The correlations between GC3s, whole GC content, Effective No. of Codons (ENC), Codon Adaptation Index (CAI), Codon Bias Index (CBI), Frequency of Optimal Codons (FOP), General average hydropathicity (Gravy) and Aromaticity (Aroma), of the selected genes were determined. The ENC-plot i.e., ENc values vs GC3s, Pr2 plot i.e., relationship between A3 / (A3 +T3) and G3 / (G3 +C3), Neutrality plot i.e., GC12 versus GC3s, and the RSCU of the genes, all shows codon usage bias existence in all the virulent genes subsists in pXO1 plasmid of *B.anthracis* genome. These results expresses the codon usage bias existing in the pXO1 plasmid's virulent genes of *B.anthracis* genome could be utilized for further exploration on their evolutionary analysis as in design of primers, design of transgenes, determine of origin of species as well as prediction of gene expression level and gene function.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

Keywords
Anthrax; Bacillus anthracis; Codon usage bias; Nucleotide; pXO1 plasmid.

Introduction

Anthrax is a Zoonotic disease, affects animals and humans. It is caused by a bacterium called *Bacillus anthracis*(Alassane S. Barro et al., 2016).*B.anthracis*is a gram-positive, non-motile, rod-shaped, spore-forming bacterium, which form spores that can continue to persist in the environment for several years. *B. anthracis* spends most of its life time in the earth as a spore, until the suitable environmental conditions are produced and allowing it to instigate a reproductive cycle (Fasanella, 2013). Malignity of utmost *B. anthracis* strains is accompanied by two bulky plasmids pXO1 and pXO2, strains that are deficient of either plasmid are will become virulent or significantly weakened. *B.anthracis*conceals three monomeric, plasmid-encoded proteins collectively called as anthrax toxin (John A T Young et al., 2007). Plasmid pXO1 is crucial for production of the anthrax-toxin proteins essentially edema- factor, lethal-factor, and protective-antigen. These proteins work in assembly of two to yield the two anthrax toxins, edema toxin (assembly of PA and EF) and lethal toxin (assembly of PA and LF). Plasmid pXO2 carries genes necessary for the production of Poly-D-glutamic acid capsule. Irrespective of the important roles of pXO1 and pXO2 in *B. anthracis* virulence, few plasmid-encoded genes have been identified and described (Thomas S. Bragg et al., 1989)(Welkos, 1988). Plasmid pXO2 holds three different genes significant for capsule production (*capA*, *capB*, and *capC*), a gene correlated to capsule degradation (*dep*) and a *trans*acting controlling gene (*acpA*)(R. T. Okinaka et al., 1999). Plasmid pXO1 offers space for the structural genes of the toxin proteins in anthrax like *cya*(edema factor), *lef*(lethal factor), and *pagA*(protective antigen) leads to acute edema and cell decease(Nicholas A. Be et al., Detection of Bacillus anthracis DNA in Complex Soil and Air Samples Using Next-Generation Sequencing, 2013).

The genetic-code is the collection of rules that define the resemblance between nucleotide triplets known as codons in DNA and amino acids in proteins. One of the main characteristics of the code is to degenerate, meaning that multiple synonymous codons stipulate the similar amino acid (Hannah M. W. Salim et al., 2008). The distinctions in the incidence of occurrence of synonymous codons in coding DNA is called as codon usage bias (Pan Tao et al., 2009). Codon usage bias diverges from genome to genome, within genomes, and from gene to gene. The bias is utmost eminent in precise types of genes, mostly in highly articulated genes (Gerrit Brandis et al., 2016). The genetic code in messenger RNA is repetitive, with 61 codons converted into 20 different amino acids. Separate amino acids are coded by up till six dissimilar codons nevertheless internal to the codon families few are used more repeatedly than others. The unicellular animals such as bacteria have an utmost codon usage bias and the amount of bias diverges between genes within the same genome (Kurnald, 1991). Codon usage bias is caused by variation in two basic factors (1) mutation pressure, which causes codon diversity, and (2) natural selection against suboptimal codons, which decreases codon diversity(Kliman, 1994). Numerous measures are developed to quantify the codon usage bias *viz* ENC, CAI, CBI and FOP. Like other methods of codon bias, CAI and ENC does not depend on organism definite data andis certainly useful to the study of novel organisms (Cavalcanti et al., 2008). Variance of codon optimization amongst genes offers distinction efficiency along with accurateness in the transformation of genes. The study of codon bias is gaining attention with the initiation of whole genome sequencing of various organisms. Molecular evolutionary investigations recommend that codon usage bias diverges both inside and amongst genomes and possibly will have substantial importance to understand the genome progression amongst interrelated species (Susanta K. Behura et al., Comparative Analysis of Codon Usage Bias and Codon Context Patterns between Dipteran and Hymenopteran Sequenced Genomes, 2012). Further, exploration of codon usage bias is important in understanding the molecular biology, genetics and genome evolution, likewise helps in new gene discovery, design of primers, design of transgenes, determine of origin of species and prediction of gene expression level and gene function. Thus, the analysis on codon usage bias assists in obtaining an in-depth knowledge of mutations that leads to evolutionary changes and understand the changes in the viral adaptations.

The aim of the study was to describe our findings on the relative contributions of mutation and natural selection to the variation in codon bias among the virulent genes (*cya, lef and pag*) subsists in pXO1 plasmid of *B.anthracis* genome.

Materials and methods

Nucleotide sequence data

The complete nucleotide sequences of cya (43 sequences), lef(93 sequences) and pag(66 sequences) genes of *B.anthracis* genome were retrieved from the National Center for Biotechnology Information (NCBI) nucleotide

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

database (http://www.ncbi.nlm.nih.gov) in FASTA format. All the sequences were aligned using MEGA X (Sudhir Kumar et al., 2018) (version 10.2.2).ATG and TGG are unique codons for methionine and tryptophan, which cannot be biased are excluded. Three stop codons TAG, TAA, TGA are excluded either from the study.

Nucleotide content analysis

The occurrences of mononucleotides (A,C,T, and G), GC, GC1s (GC contented at the first codon positions), GC2s (GC contented at the second codon positions), GC12s (the mean value of GC1s and GC2s), and GC3s (GC contented at the third codon positions) were calculated using seqinr package (version 3.6-1) (Delphine Charif et al., 2007) of R (version 3.6.2) (RCoreTeam, 2020). The frequencies of mononucleotides at the third correspondent codon position (A3s, C3s, U3s, and G3s) were intended using CodonW software (version 1.4.2) developed by J. Peden (http://codonw.sourceforge.net/).

Relative dinucleotide abundances

Disparity in the incidence of dinucleotide sets may influence the codon usage. Dinucleotide frequency is generally used to find whether some dinucleotide pairs are gratified by an organism or not. A maximum of 16 dinucleotide groupings are possible. The patterns of dinucleotide frequency designate both selection and mutational pressures which was calculated using the following formula:

$$P_{xy} = f_{xy}/ (f_x * f_y)$$

where f_x and f_y are the frequency of single nucleotides (x & y, respectively), and f_{xy} is the frequency of dinucleotides (xy) in the matching sequence. The ratio of the identified dinucleotide frequency to the expected dinucleotide frequency is the odds ratio, if the odds ratio is greater than 1.25, the dinucleotide is inspected as overrepresented, however the values below 0.78 show a underrepresentation (Rekha Khandia et al., 2019). Dinucleotide content was calculated by R (version 3.6.2) (RCoreTeam, 2020)

Effective number of codons (ENC)

The Effective Number of Codons (ENC)analyses how distant the codon usage of a gene set out fromequivalent usage of synonymous codons via codon usagedata and is not dependent on the length of the gene and amino acidstructure. In the DNA coding sequence, maximum amino acids (except Methionine and Tryptophan), are encoded using two or more codons, they are called as synonymous codons. To detect the bias in the usage of synonymous codons, Wrightagreed the concept of the effective number of codons (ENc) (Wright, 1990). ENc values vary from 20 to 61. A value of 20 designates extreme bias, meaning regardless of the availableness of synonymous codons, the amino acid is encoded by only one codon. However, a value of 61 designates no bias in the codon usage and meaning that all the available codons are utilized correspondingly. Generally, if the observed ENc value is less than 40, the genome is considered to have highly biased codon usage (Snawar Hussain et al., 2020)(Fuglsang, 2006). The ENC values for *cya*, *lef* and *pag* genes of *B.anthracis* genome were determined using CodonW software (version 1.4.2) developed by J. Peden (http://codonw.sourceforge.net/).

Relative synonymous codon usage(RSCU) analysis

RSCU values for each codon were calculated to determine the patterns of synonymous codon usage. The RSCU is the ratio of the perceived frequency to the expected frequency for a specific amino acid. It is unaffected by the length of the DNA sequence or frequency of the amino acid(Susanta K. Behura et al., 2012). The codons with value of RSCU >1.0 indicate a positive codon usage bias, while codons with value of RSCU <1.0 indicates a negative codon usage bias (Snawar Hussain et al., 2020). Overrepresented codons possess RSCU values of more than 1.6, whereas underrepresented codons have values less than 0.6. Codons with values of RSCU ranging between0.6 and 1.6 are considered unbiased or randomly used (Xiaoting Yao et al., 2020). The RSCU values are useful in computing codon usage between genes that differ in size and the amino acid composition (S Aravind et al., 2014)To determine the codon usage bias pattern of *cya*, *lef* and *pag*genes, the relativesynonymous codon usage (RSCU)of the *B.anthracis* genome coding region was calculated using R (version 3.6.2) (RCoreTeam, 2020).

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

Codon adaptation index (CAI)

CAI value for a gene is determined from the frequency ofusage of all the codons in that gene. This can be used toequate codon usage in dissimilar genes and in diverseorganisms. It is a numerical value that indicates how repeatedly a preferential codon is used among the greatly expressed genes (Li Gun et al., 2018). The CAI value is sequence length independent, this value depends only on the amino acid frequency (Xia, 2007). CAI provides values vary from0 (meaning no bias) to 1.0 (meaning complete bias). Codon Adaptation Index (CAI) is used in recognition with the role of natural selection in producing high levels of codon bias. Codon adaptation index (CAI) of genes were determined by CodonW software (version 1.4.2) developed by J. Peden (http://codonw.sourceforge.net/).

Codon bias index (CBI)

The Codon Bias Index is a fraction whose numerator is the overall number of times that the preferred codons are used in the protein minus the number of such usages expected if the code were read randomly. The denominator is the overall amount of amino acid remains in the protein (excluding methionine, tryptophan, and aspartic acid residues) minus the random expectation for usage of the preferred codons (Bennetzen, 1982). The codon bias index reveals the existence of components by means of highcodon usage in a particular gene. The value of CBI more evidently designates the foreign gene expression in the host animal. The CBI index can beestimated by the following formula:

$$CBI = (N_{optimal} - N_{random}) / (N_{total} - N_{random})$$

where the $N_{optimal}$ is the total number of incidences of the superior codon in the gene, in the current work, the superior codons are considered as the codons whose value of RSCU is greater than 1.6. N_{random} is the sum of the number of incidences of the superior codon when entire synonymous codons are random in a particular protein, N_{total} is the existent amount of the amino acid matching to the superior codon in the gene (Li Gun et al., 2018). In the present work, the CBI of genes for *B.anthracis*genome is determined by CodonW software (version 1.4.2) developed by J. Peden (http://codonw.sourceforge.net/).

Frequency of optimal codons (FOP)

Frequency of Optimal Codons (FOP) is the weighted average of the relative synonymous codon usage of superior codons (RSCU value > 1.6) (Li Gun et al., 2018).Codon bias can be calculated as the frequency of optimal codons (Fop). Gene with extreme codon bias will have, Fop value equals to 1, while for a gene with random codon usage, Fop value equals to 0 (Hui Song et al., 2017). Fop of the genes for *B.anthracis* is determined by CodonW software (version 1.4.2) developed by J. Peden (http://codonw.sourceforge.net/).

General average hydropathicity (Gravy) and aromaticity (Aromo) indices

In examining the natural selection for influencing the codon usage bias of the *cya*, *lef* and *pag*genes of *B.anthracis*, the indices, comprising Gravy and Aromo values, were involved in the current study. These values were determined by CodonWsoftware (version 1.4.2) developed by J. Peden (http://codonw.sourceforge.net/) and evaluated the frequencies of hydrophobic and aromatic amino acids, correspondingly. Accordingly, the variation of the two indices reveals the usage of the amino acid (Naveen kumar et al., 2016). A larger Gravy or Aromo value advocates a more hydrophobic or aromatic amino acid content (Ye chen et al., 2017).

ENc-GC3s plot analysis

The ENc-plot analysis is the plotting of ENc-values versus GC3s is normally used to discover factors inducing the codon usage patterns (Huiguang Wu et al., 2020). The ENc-GC3s plot is used to find whether the codon usage of given genes is specially due to mutational pressure or selection pressure. If the data points fall on the expected curve, it specifies mutational pressure was the major force acting on third-position bases of codons, while if the points fall beneath the expected curve, the codon usage is considered to be affected by selection pressure (Snawar Hussain et al., 2020). ENc-GC3s plots generated using R (version 3.6.2) (RCoreTeam, 2020).

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

Neutrality plot analysis

A neutrality plot plotted over GC12 versus GC3 was used to describe the degree of mutational pressure and natural selection on the codon usage. In the synonymous codons, only the last nucleotide is dissimilar, and the amino acid will stay unchanged. As nucleotide commute at the third position of the codon it does not contribute to variations in the amino acid, it is the indication of a mutational force. When there is a nucleotide variation that brings about the changes in the altered amino acid, it reveals the indication of selection force. When there exists a correlation amongst the GC12 and GC3, it is probably due to mutational forces and the force impelling codon bias is existent at all the codon positions (Gareth M Jenkins et al., 2003). The GC3 and GC12 values (mean of GC1 and GC2) of the synonymous codons generated were plotted on the abscissa and ordinate, respectively, to yield a scatter map for the neutrality plot. The regression line was drawn between the GC3 and the GC12 values respectively. The slope (regression coefficient) of the regression line is observed as the mutation-selection equipoise coefficient. If all of the points are scattered along the diagonal (slope = 1) and the correlation between GC3 index and GC12 index is statistically significant, this indicates that mutation is the main force shaping the codon usage. Otherwise, if the regression curve is parallel or tilted toward the abscissa (near to zero slope), selection pressure is reflected as the dominant factor (Zhiwen Chena et al., 2020). The regression analysis to estimate the linear relationship among GC3 value and GC12 value, was accomplished by using R programming (version 3.6.2) (RCoreTeam, 2020).

Parity Rule 2 (PR2) plot analysis

Parity rule 2 (PR2) plot analysis is another technique used to examine the impact of mutation pressure and selection pressure on codon usage. In the PR2 plot, codon usage bias was determined based on the GC bias [G3/ (G3 + C3)] and AT bias [A3/ (A3 + T3)] and was plotted as the abscissa and ordinate, respectively. The middle of the plot, i.e., G = C and A = T (PR2), described as coordinates of the origin (0.5, 0.5), designates no bias between the influences of mutation pressure and natural selection (Wu, 2020). The vector from the midpoint shows the extent and direction ofbiases from PR2. PR2 bias plots are predominantly informative whenPR2 biases at the third position of the codon in four codon sequences of individual genes are plotted (Sueoka, 1999). Chargaff's second parity rule (PR2) says, the number of residues A will be equal to T and residues C will be equal to G in a DNA sequences (Rapoporta, 2013). The PR2-bias plots were plotted using R programming (version 3.6.2) (RCoreTeam, 2020).

Resultsand discussion

Nucleotide contents analysis

Codon usage pattern belonging to compositional characteristics such as the T3, C3, A3, G3, GC12, GC3, overall GC content, ENc, FOP, CAI and CBI of virulent genes subsists in pXO1 plasmid of *B.anthracis*genome are calculated (Supplementary file S1). Table1 shows the average values of basic parameters namely,T3s, $A_{3s,}$ G_{3s}, C_{3s}, overall GC content, GC3s, GC_{12}, ENc, FOP, CBI, and CAI of the genes *cya, lef* and *pag* of *B.anthracis* genome.

RSCU pattern

RSCU value is a vital factor for assessing the bias of the synonymous codon. It represents the ratio between the frequency of incidence of one codon to the expected usage frequency in a gene sample. If RSCU values of codons more than 1.0 then it is considered as having positive codon usage bias. On the other side, if RSCU values of codons are less than 1.0 then it is considered as the less abundant codons. The overall RSCU value can intuitively shows the preference for codon usage bias in a certain genome. The overall RSCU values of *B.anthracis*genome are calculated and the results are shown in Figure 1. RSCU values of the *cya, lef* and *pag*genes is provided in supplementary file (Supplementary file S2).

The codon usage bias indicates that the bars in those RSCU values are greater than 1.6, which could be considered as the abundant codons, these codons are ACA, AGA, AGT, CAT, GCT, TAT, TTA and TTT in *cya* gene, AAT, ACA, AGA, AGT, ATT, CCA, GAT, GCA, GTA, TAT, TC, TGT and TTA in *lef*gene, AGA, AGT, CAA, CAT, GAT, GCA, GGA, TCT and TTA in *pag*gene. Among them, TTA has the highest RSCU value of 3.73, 3.22 and 4.06 in *cya, lef* and *pag* genes respectively. The bars depicting the RSCU values less than 0.6 are considered to be less abundant codons. In general, the codons ending with the A or T has the RSCU values smaller compared to the codons ending with G or C RSCU values. Three stop codons namely TGA, TAA, and TAG, and there is no corresponding amino acid related to stop codons and therefore removed from the data. The value of RSCU is usually utilized as a dimension of the codon usage bias. The overall RSCU values (Figure 1), to a certain degree can show the features of codon usage, but if we see the codon usage of individual gene sequences for a particular genome, it can be observed that there exists a difference of RSCU values between the different genes even though they belong one genome.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

Relative dinucleotide frequency abundances influence the codon usage bias

We accomplished a dinucleotide analysis on the three different virulent genes (*cya, lef* and *pag*) subsists in pXO1 plasmid of *B.anthracis* to understand the possible influence of dinucleotide frequencies on the codon usage. Dinucleotides CC and TT were overrepresented (P_{xy}>1.25) in *cya*, CC was overrepresented in *lef* and GG was overrepresented in *pag*. Whereas dinucleotides CG was underrepresented (Pxy<0.78) in *cya* gene, dinucleotides AC, CG and GT were underrepresented in *lef*and none of the dinucleotides are underrepresented in *pag*gene (Table 2 and Figure 2). These results showed that significant biases of the dinucleotide content variation were observed in the three virulent genes (*cya, lef* and*pag*) subsists in pXO1 plasmid of *B.anthracis*.

Identification of the factors influencing codon usage patterns

To assess the forces influencing the codon usage patterns in the three virulent genes viz *cya,lef* and *pag* genes of *B.anthracis* genome, ENC plots, Neutrality plots and PR2 bias plots analyses techniques were used. ENC is a significant index to measure the codon usage bias in a genome and play a major role in their codon usage pattern. In order to examine the synonymous codon usage pattern of the three virulent genes *cya, lef* and *pag* genes subsists in pXO1 plasmid of *B.anthracis*genome, the ENC versus GC3 is plotted and the result is depicted in Figure 3. Each point denotes coding sequence in the respective gene. The ENC value is lies in the range of 21 to 61, but when the f_k is calculated via the equation $(n\Sigma(n_i/n)^2 - 1)/ (n - 1)$, all ENC values lies in the range 20 to infinite, the value 20 represents the codon usage bias using only one possible synonymous codon corresponding to an amino acid, and the larger value represents that there is less bias of using all possible synonymous codons, and the incidence of using all possible synonymous codons may incline to be equal. In the ENC plot, *cya* and *lef* genes ENC values fell closer as well as just below the expected ENC curve where as in *pag* gene the ENC values fell relatively far from the expected ENC curve (Figure 3). Additionally, in *pag* gene sequences were clustered separately, whereas sequences of *cya* and *lef* were scattered in the ENC plots. These results indicate that mutation pressure and natural selection led to the codon usage bias of the genes. In *cya* and *lef* genes mutation pressure has more influence on codon usage bias than natural selection where in *pag* gene natural selection has more influence on codon usage bias than mutation pressure.

The neutrality analysis performed between the GC3s and GC12s values was used to find out the degree of the two evolutionary forces *viz* mutation pressure and natural selection on the codon usage pattern of *B.anthracis* genome. Generally, the neutrality plot is used to evaluate the directional mutation pressure against natural selection of a certain genome, which could also shows the relationship between the GC12 and the GC3 with the GC12 as the vertical axis and the GC3 as the horizontal axis. The neutrality plots of *cya, lef* and *pag* genes of *B.anthracis* genome is depicted in Figure 4. Each point in Figure 4 denotes the sequences of the individual gene of the *B.anthracis*. A substantial correlation between GC3s and GC12s was observed in the*cya* and *lef* genes (y = 0.27 + 0.0487 x, R^2= 0.024 and y = 0.304 - 0.016 x, R^2 = 0.016 respectively) (Figure 4). Therefore, the percentage of constraints of natural selection was found to be less for the *cya* and *lef* genes. No substantial correlation between GC3s and GC12s was observed in the *pag* gene (y = 0.501 – 0.498 x, R^2 = 0.87) (Figure 4). Hence, natural selection plays a dominant role in driving codon usage bias for *pag* gene of *B.anthracis* genome.

The parity rule says that if there is no mutation in genes, or no bias on the codon selection effect, the base content must follow the rules A = T and G=C. This method is generally used to analyze the PR2 bias of the third position codon by comparing the values A3/ (A3 + T3) and G3/ (G3 + C3). The distance between dot and the center denotes the amount and direction of the PR2 bias. The PR2 bias plots of *cya, lef*and *pag* genes of *B.anthracis* genome is depicted in Figure 5. In the PR2 bias analysis, substantial deviations from the parity rules were noticed (A, T, C, G) (Figure 5), showing that the amount of the evolutionary forces influencing the codon usage patterns of the three genes were different.

Overall, the above results indicate that the consequence of mutation pressure has more influence in the codon usage of *cya*and *lef* genes, but natural selection dominates the evolution of codon usage of the *pag* gene of *B.anthracis* genome. Meanwhile, the correlation analysis of *cya, lef* and *pag* genes shown in Table 3, Table 4 and Table 5 respectively. The result shows there is a strong correlation with the value of 0.16 (GC3s and GC12) in *cya* gene. Whereas, there is weak correlation with values -0.13 and 0.0 in *lef*and *pag* genes respectively. The significant correlation with p-value less than 0.01 between the parameters is highlighted using red colored cell in the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

table.Gravy and aromo is non significantly correlated with all the parameters except CAI and Fop, with P<0.01, in *cya* gene. while in *lef*and *pag* genes gravy and aromo is significantly correlated with only GC12. From these data performance, it can be comprehended that the mutation pressure has greater impact in *cya* gene. The *lef* and *pag* genes may be has greater impact of natural selection on codon preference in *B.anthracis* genome.

Virulent structure of *B. anthracis* are the two large pathogenicity related plasmids: 1) pXO1, which encodes the anthrax toxin genes *cya*, *lef*, and *pagA* and 2) pXO2, that carries the genes *capA*, *capB*, *capC*, and *capD* accountable for capsule synthesis and degradation. *B.anthracis*shows its pathogenic action mostly through the anti-phagocytic activity and the manufacture of a toxic complex comprising of three proteins known as protective antigen (PA), lethal factor (LF), and edema factor (EF). The two bulky plasmids of *B. anthracis* are vital for full pathogenicity, exclusion of either intensely weakens the virulence effect of *B. anthracis* (Fasanella, 2013). Toxins and the factors essential for bacterial encapsulation are encoded on two bulky plasmids, pXO1 and pXO2. Out of three of the anthrax toxin proteins two factors, edema factor (EF) and lethal factor (LF), each form a binary combination with protective antigen (PA), causing acute edema and cell death (Nicholas A. Be et al., Detection of Bacillus anthracis DNA in Complex Soil and Air Samples Using Next-Generation Sequencing, 2013)The *B. anthracis* genome is three-way and contained a two circular virulence plasmids and a single circular chromosome. The genome nucleotide composition is extremely biased towards adenine and thymine, with only ~35% of the bases from guanine and cytosine. The occurrence of A+T means the DNA has a greater resistant density and lesser melting temperatures than many others. The plasmids are comparatively large and code for numerous different genes, but essentially they carry the capsule and toxin factors. The pXO1 plasmid has *pag*A (protective antigen and a intra membrane toxin carrier), *lef*(lethal factor, is a Zn2+-dependent end protease) and *cyaA*(edema factor, is a calmodulin sensitive adenylate cyclase). The pXO2 plasmid carries the capsule biosynthesis genes found in a group and is essential for complete anthrax disease (Paul Keim et al., 2009)(Nicholas A. Be et al., 2013). Virulence of many *B. anthracis* strains is connected with two large plasmids, and strains deficient either plasmid will become either a virulent or significantly weakened. Plasmid pXO2 carries genes essential for the synthesis of an anti-phagocytic poly-D-glutamic acid capsule. The 110-MDa plasmid pXO1 is essential for synthesis of the anthrax toxin proteins such as edema factor, lethal factor, and protective antigen. These proteins work in combinations of two to produce the two anthrax toxins namely edema toxin (protective antigen and edema factor) and lethal toxin (protective antigen and lethal factor). Plasmid pXO2 carries three genes essential for capsule synthesis (*capA*, *capB*, and *capC*), a gene related to capsule degradation (*dep*), and a trans acting controlling gene (*acpA*). Plasmid pXO1 has the structural genes for the anthrax toxin proteins such as *cya* (edema factor), *lef* (lethal factor), and *pagA* (protective antigen)], as well as two trans acting controlling genes (*atxA* and *pagR*), a gene encrypting a type I topoisomerase (*topA*), and a newly categorized operon comprising of three genes whose functions seem to affect germination(R. T. Okinaka et al., 1999). The virulence of anthrax bacilli is due to the manufacture of protein exotoxins and a poly D glutamicacid capsule. Three different toxin proteins have been known, and they are Edema Factor (EF), Lethal Factor (LF) and Protective Antigen (PA) (Donald L robertson et al., 1988). From the literature studies we found that the virulent genes of *B.anthracis*are*cya, lef and pag (pXO1 plasmid)* genes *and cap* genes *(pXO2 plasmid).* Hence selected the pXO1 plasmid genes *cya, lef* and *pag* for our codon usage bias analysis study.

Codon bias has a very extensive importance for exploring a genome. In bacteria percentage of synonymous codon usage differ significantly among the genes and genes having a high codon bias develop more slowly (Paul M sharp et al., 1987). Codon bias has an extensive importance for exploring a genome. Codon bias can also describe the basic methodology of evolutionary process in biology (Tai-Chun Wang et al., 2012). The early literature studies revealed that the codon usage of the EF gene showed the high value of A + T (71%) base composition for its DNA (Donald L robertson et al., 1988). The codon usage of the*lef* gene showed its high A + T (70%) base composition (Thomas S. Bragg et al., 1989). The codon usage for the *pag*gene showed its high A + T (69%) content (Welkos, 1988).In this study along with the standard methods, correlation between the parameters CAI, CBI, FOP, Gravy and Aromo were also analyzed to examine the codon usage bias of *B.anthracis* genome. Overall RSCU values of the virulent genes *cya*, *lef* and *pag* of *B.anthracis*genome (Figure 1) shows strong codon usage bias. Relationship among GC3s and GC12 showed that the natural selection pressure is a slightly more important than mutation pressure in *pag* gene compared to *cya* and *lef* genes.

Conclusion

In this paper, the codon usage patterns of virulent genes *vizcya*, *lef* and *pag* subsists in pXO1 plasmid of *B.anthracis*genome were analyzed. The ENC-plot, theA3/ (A3 + T3) versus G3/ (G3 + C3) plot, the relationship GC12 versus GC3s, the overall RSCU and the dinucleotides are all analyzed. The codon usage pattern and its influencingfactors, were identified for *B.anthracis*genome. As observed, codon usage patternsin *B.anthracis*genome are influencedvia GC3s bias. Correlation among codon bias index andGC3s in *cya, lef* and *pag*(0.79, 0.99 and 1.0

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

respectively) indicates that the GC3 bias of *B.anthracis*genome may also reveal its codon bias index.Negative correlation between overall GC contentand CAI in *cya* gene may indicate the important role of mutation pressure in modelling thecodon usage bias where as Positive correlation between overall GC contentand CAI in *lef* and *pag* genes indicate the important role of natural selection in modelling thecodon usage bias in *B.anthracis*genome. Weperformed the comparative analysis of codon usage bias in three different virulent genes subsists in pXO1 plasmid of *B.anthracis*genome. Theresults would help further expose the underlying dynamicsof genetic evolution in *B.anthracis*genome.These results showed that the codon usage bias exists in the*B.anthracis*genome. All these informationis important for explaining the function of *B.anthracis*and also helps in understanding the evolutionaryprocess of the *B.anthracis* genome.

Limitations and Future studies

Only three genes which are virulent and subsists in plasmid pXO1 of *B.anthracis* were analyzed in this work. it is necessary to expand the scope of the genome in the further study.

Acknowledgement

We would like to thank the Spatial Epidemiology lab, Indian Council for Agriculture Research (ICAR) - National Institute of Veterinary Epidemiology and Disease Informatics, Department of Veterinary Public Health and Epidemiology, VeterinaryCollege and Outreach project on Zoonotic diseases, ICAR for providing necessary support to carry out this research work.

References

[1] Alassane S. Barro et al. (2016). Redefining the Australian Anthrax Belt: Modeling the Ecological Niche and Predicting the Geographic Distribution of Bacillus anthracis. *PLOS One*.

[2] Bennetzen, J. L. (1982). Codon Selection in Yeast. *The Journal of Biological Chemistry*, 3026-3031.

[3] Cavalcanti et al. (2008). Factors Influencing Codon Usage Bias in Genomes. *J. Braz. Chem. Soc*, *19*, 257-262.

[4] Delphine Charif et al. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. *In Structural Approaches to Sequence Evolution; Springer: Berlin/Heidelberg, Germany*, 207–232.

[5] Donald L robertson et al. (1988). Nucleotide sequence of the Bacillus anthracis edema factor gene (cya): a calmodulin-dependent adenylate cyclase. *Elsevier*.

[6] Fasanella, A. (2013). Bacillus anthracis, virulence factors, PCR, and interpretation of results. *Landes Bioscience, 4:8*, 659–660.

[7] Fuglsang, A. (2006). Estimating the 'effective number of codons': The Wright way of Determining Codon Homozygosity Leads to Superior Estimates. *Genetics Society of America*, : 1301–1307.

[8] Gareth M Jenkins et al. (2003). The extent of codon usage bias in human RNA viruses and its evolutionary origin . *Elsevier Virus research*, 1-7.

[9] Gerrit Brandis et al. (2016). The Selective Advantage of Synonymous Codon Usage Bias in Salmonella. *PLOS Genetics*.

[10] Hannah M. W. Salim et al. (2008). Factors Influencing Codon Usage Bias in Genomes. *J. Braz. Chem. Soc*, 257-262.

[11] Hui Song et al. (2017). Comprehensive analysis of correlations among codon usage bias, gene expression, and substitution rate in Arachis duranensis and Arachis ipaensis orthologs . *scientific reports*, 7: 14853.

[12] Huiguang Wu et al. (2020). Comprehensive Analysis of Codon Usage on Porcine Astrovirus. *MDPI Viruses*, 991.

[13] John A T Young et al. (2007). Anthrax Toxin: Receptor Binding, Internalization, Pore Formation, and Translocation. *Annu. Rev. Biochem, 76*, 243–65.

[14] Kliman, R. M. (1994). The Effects of Mutation and Natural Selection on Codon Bias in the Genes of Drosophila . *Genetics Society of America*, 1049-1056.

[15] Kurnald, C. G. (1991). Codon bias and gene expression. *Elsevier, 285*, 165 - 169.

[16] Li Gun et al. (2018). Comprehensive Analysis and Comparison on the Codon Usage Pattern of Whole Mycobacterium tuberculosis Coding Genome from Different Area. *Hindawi BioMed Research International, 2018*.

[17] Li, P. M.-H. (1986). An Evolutionary Perspective on Synonymous Codon Usage in Unicellular Organisms. *Journal of Molecular Evolution*, 28-38.

[18] Naveen kumar et al. (2016). Revelation of Influencing Factors in Overall Codon Usage Bias of Equine Influenza Viruses. *PLOS one*.

[19] Nicholas A. Be et al. (2013). Detection of Bacillus anthracis DNA in Complex Soil and Air Samples Using Next-Generation Sequencing. *PLOS One, 8*(9).

[20] Nicholas A. Be et al. (2013). Detection of Bacillus anthracis DNA in Complex Soil and Air Samples Using Next-Generation Sequencing. *PLOS one*.

[21] Pan Tao et al. (2009). Analysis of synonymous codon usage in classical swine fever virus. *Springer*, 104–112.

[22] Paul Keim et al. (2009). The genome and variation of Bacillus anthracis. *Mol Aspects Med, 30*(6), 397–405.

[23] Paul M sharp et al. (1987). The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. © *IRL Press Limited*, 15/3.

[24] R. T. Okinaka et al. (1999). Sequence and Organization of pXO1, the Large Bacillus anthracis Plasmid Harboring the Anthrax Toxin Genes. *JOURNAL OF BACTERIOLOGY, 181*(20), 6509–6515.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

[25] Rambaut. (n.d.). *A. 2012. FigTree v1. 4.* Retrieved from http://tree.bio.ed.ac.uk/software/figtree/

[26] Rapoporta, A. E. (2013). Compensatory nature of Chargaff's second parity rule. *Journal of Biomolecular Structure and Dynamics*, 1324–1336.

[27] RCoreTeam. (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. Retrieved from http://www.r-project.org/

[28] Rekha Khandia et al. (2019). Analysis of Nipah Virus Codon Usage and adaptation to hosts. *Frontiers in microbiology*.

[29] S Aravind et al. (2014). Bioinformatics study involving characterization of synonymous codon usage bias in the duck enteritis virus glycoprotein D (gD) gene. *Asian journal of animal and veterinary advances* , 229-242.

[30] Snawar Hussain et al. (2020). Analysis of Codon Usage and Nucleotide Bias in Middle East Respiratory Syndrome Coronavirus Genes. *Evolutionary Bioinformatics*, 1–13.

[31] Sudhir Kumar et al. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol*, 1547–1549.

[32] Sueoka, N. (1999). Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. *Elsecier Gene*, 53–58.

[33] Susanta K. Behura et al. (2012). Comparative Analysis of Codon Usage Bias and Codon Context Patterns between Dipteran and Hymenopteran Sequenced Genomes. *PLOS ONE*.

[34] Susanta K. Behura et al. (2012). Comparative Analysis of Codon Usage Bias and Codon Context Patterns between Dipteran and Hymenopteran Sequenced Genomes. *PLOS one*.

[35] Tai-Chun Wang et al. (2012). The evolutionary landscape of the Mycobacterium tuberculosis genome. *Elsevier*, 7.

[36] Thomas S. Bragg et al. (1989). Nucleotide sequence and analysis of the lethal factor gene (Zef) from Bacillus anthracis. *Elsevier*, 45-54.

[37] Welkos, S. (1988). Sequence and analysis of the DNA encoding protective antigen of Bacillus anthracis. *Elsevier*, 287-30.

[38] Wright, F. (1990). The 'effective number of codons' used in a gene. *Elsevier*, 23-29.

[39] Wu, H. (2020). Comprehensive Analysis of Codon Usage on Porcine Astrovirus. *MPDI* , 991.

[40] Xia, X. (2007). An Improved Implementation of Codon Adaptation Index. *Evolutionary Bioinformatics*, 53–58.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

[41] Xiaoting Yao et al. (2020). Codon usage bias analysis of Bluetongue virus causing livestock infection. *Frontiers in microbiology*.

[42] Ye chen et al. (2017). Comprehensive analysis of the codon usage patterns in the envelope glycoprotein E2 gene of the classical swine fever virus. *PLOS one*.

[43] Zhiwen Chena et al. (2020). Comparative analysis of codon usage between Gossypium hirsutum and G. barbadense mitochondrial genomes. *Taylor & Francis*, 2500–2506.

FIGURES

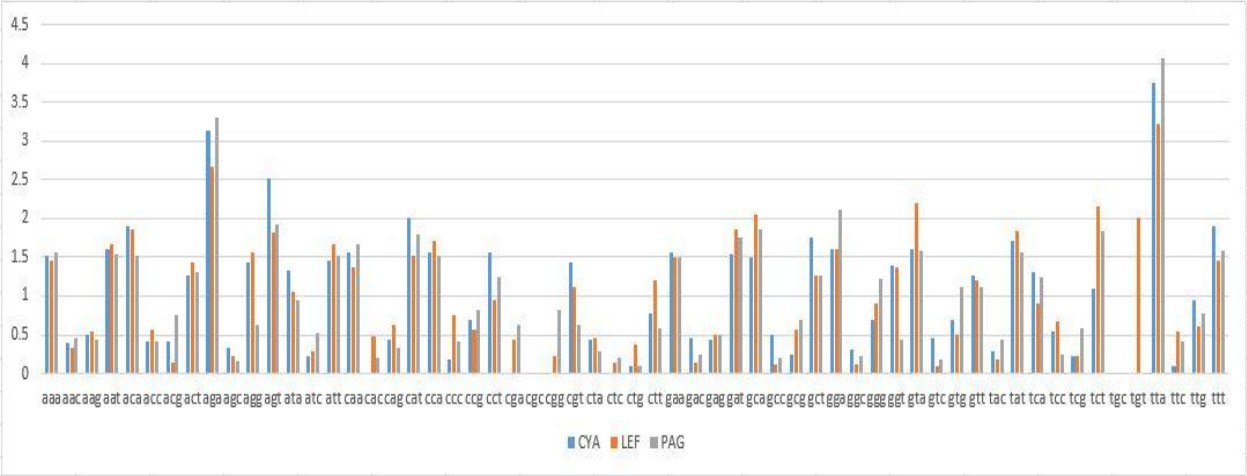
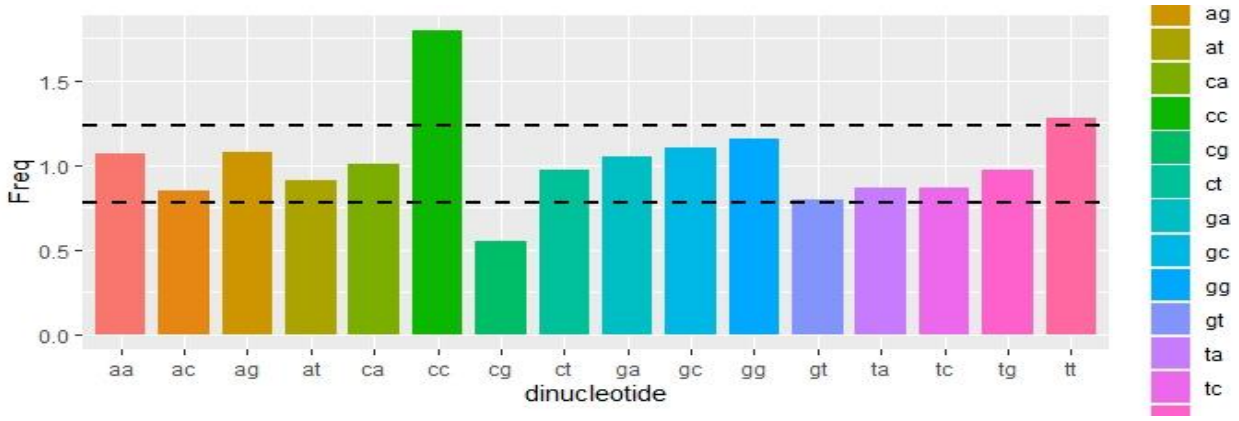


Figure 1: Overall RSCU of *cya* gene,*lef* gene and *pag* gene of *B.anthracis* genome

a)



Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

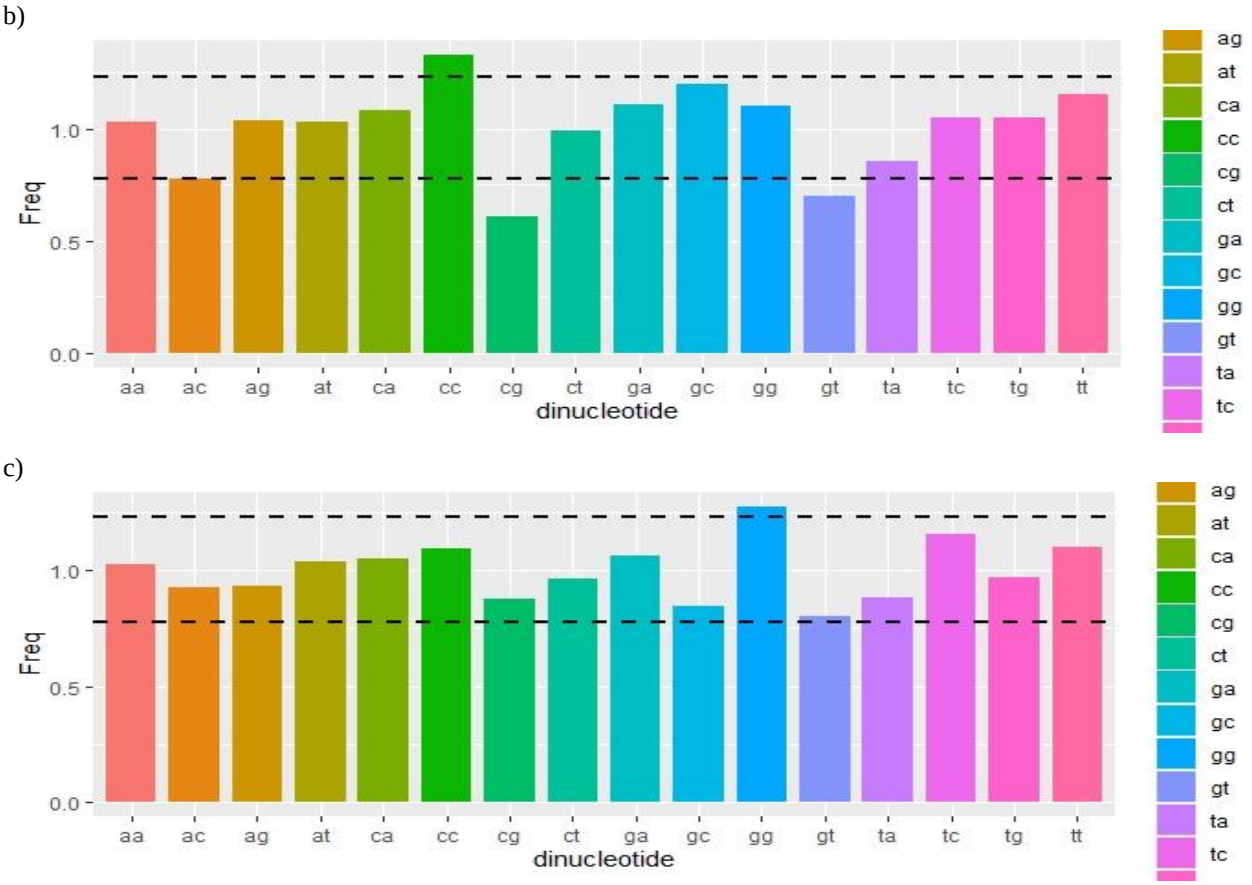
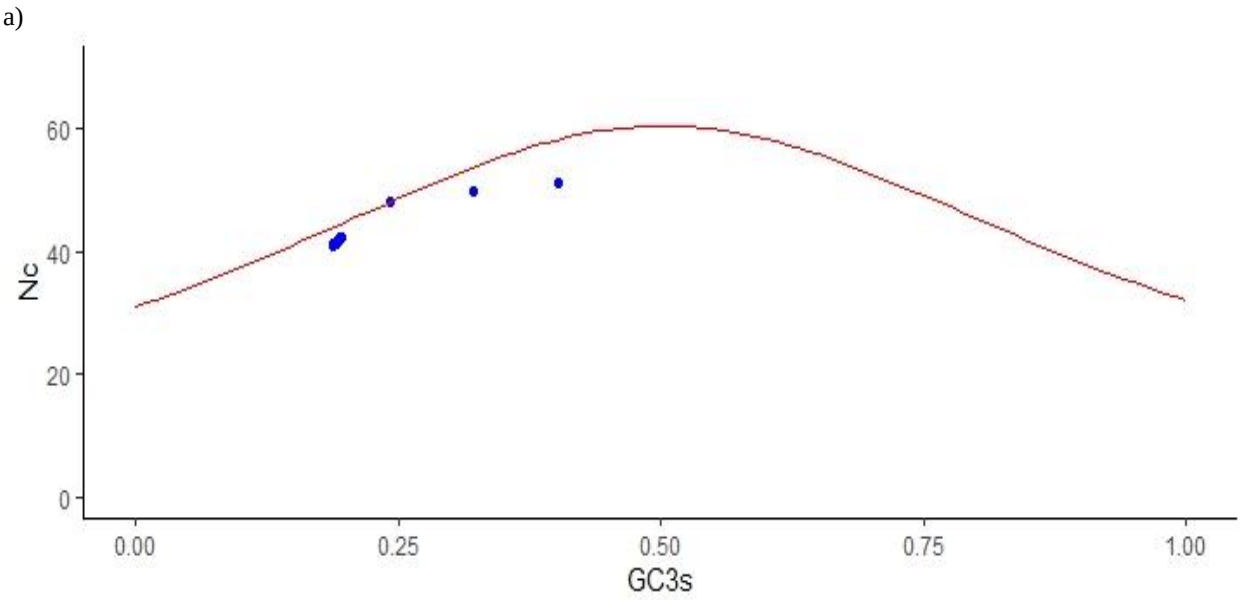
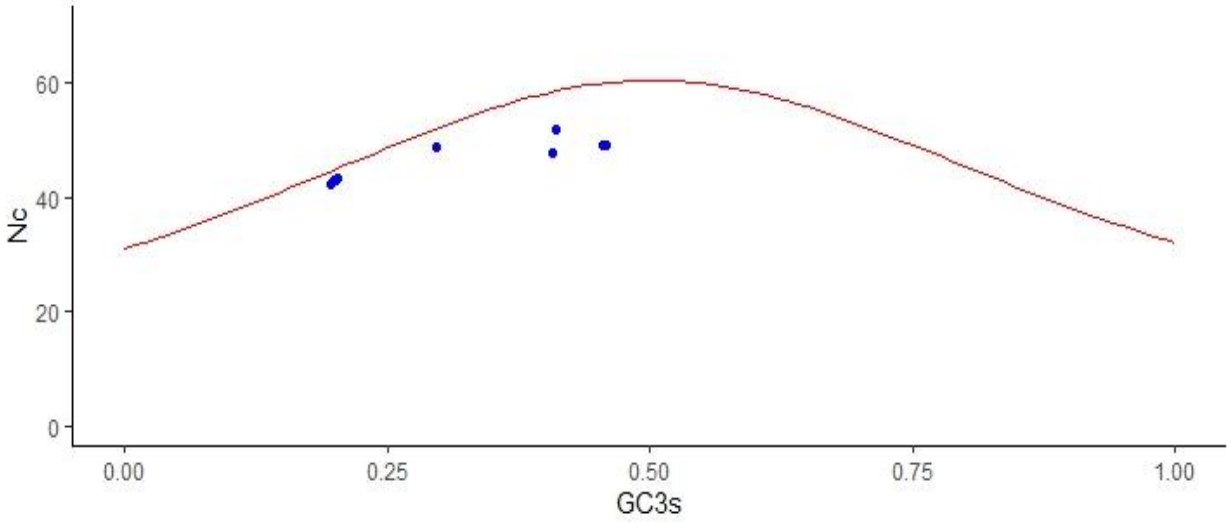


Figure 2: Dinucleotide abundancy of the a) *cya*, b) *lef* and c) *pag* genes of *B.anthracis*. The various colors represent the various dinucleotides. Dinucleotides are considered as under-represented or over-represented if the relative abundance values are less than 0.78 or greater than 1.25 (dashed lines), respectively.



Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

b)



c)

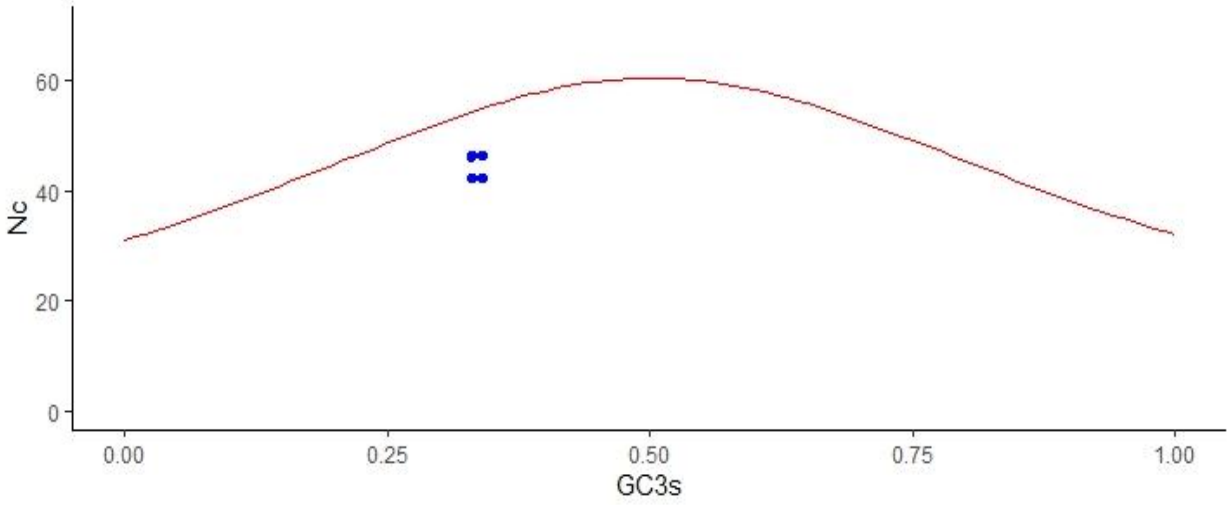
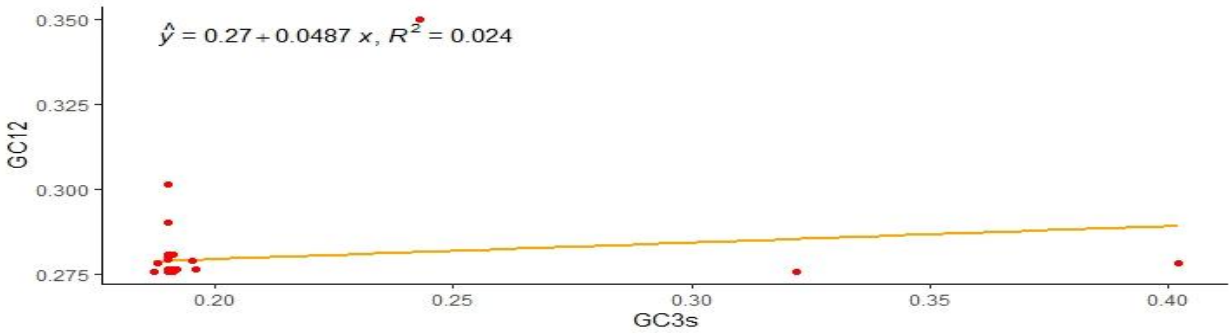


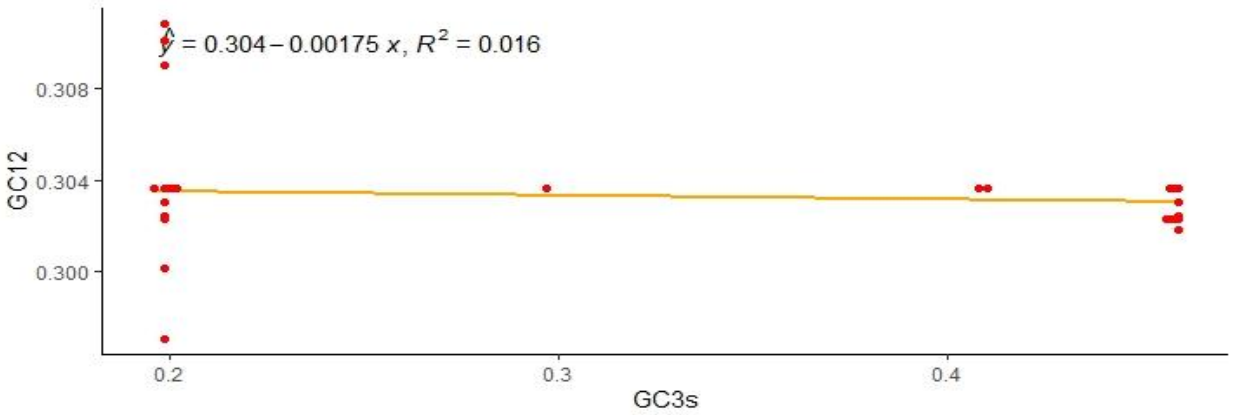
Figure 3: ENC plot analysis of the complete coding sequences of a) *cya*, b)*lef* and c) *pag*genes of *B.anthracis*. The ENC plot displayed the relationships between ENc and GC content at the third codon position (GC3s) of protein-coding sequences. The curve represents the expected ENC values for all GC3s compositions.

a)



Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

b)



c)

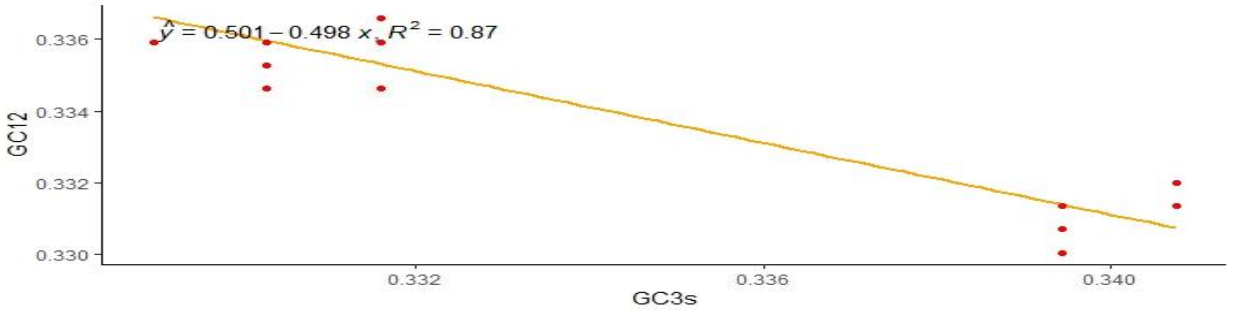
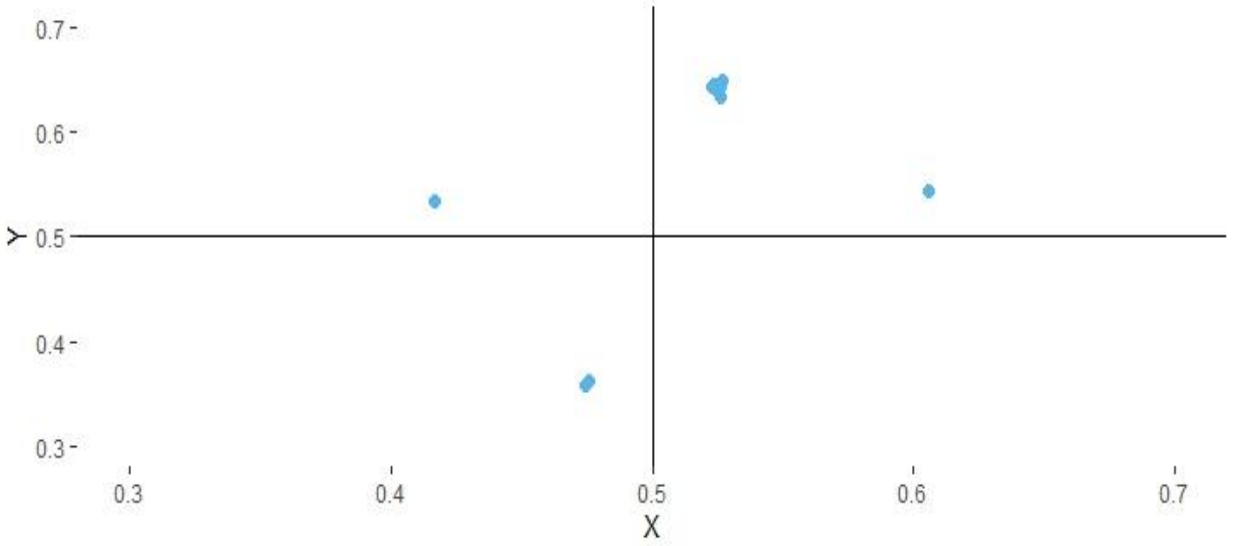


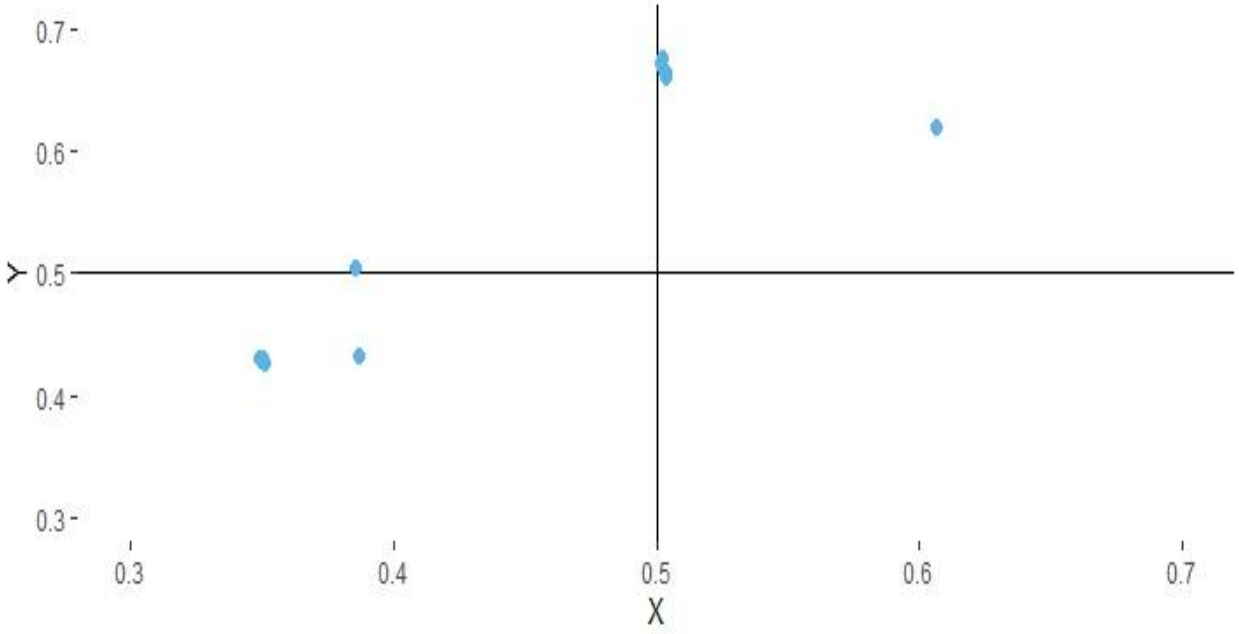
Figure 4: Neutrality analysis of the complete coding sequences of a) *cya*, b)*lef* and c) *pag*genes of *B.anthracis*.. The neutrality plot shows the correlation between GC content at synonymous positions (GC12s) and GC content at non-synonymous positions (GC3s).

a)



Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

b)



c)

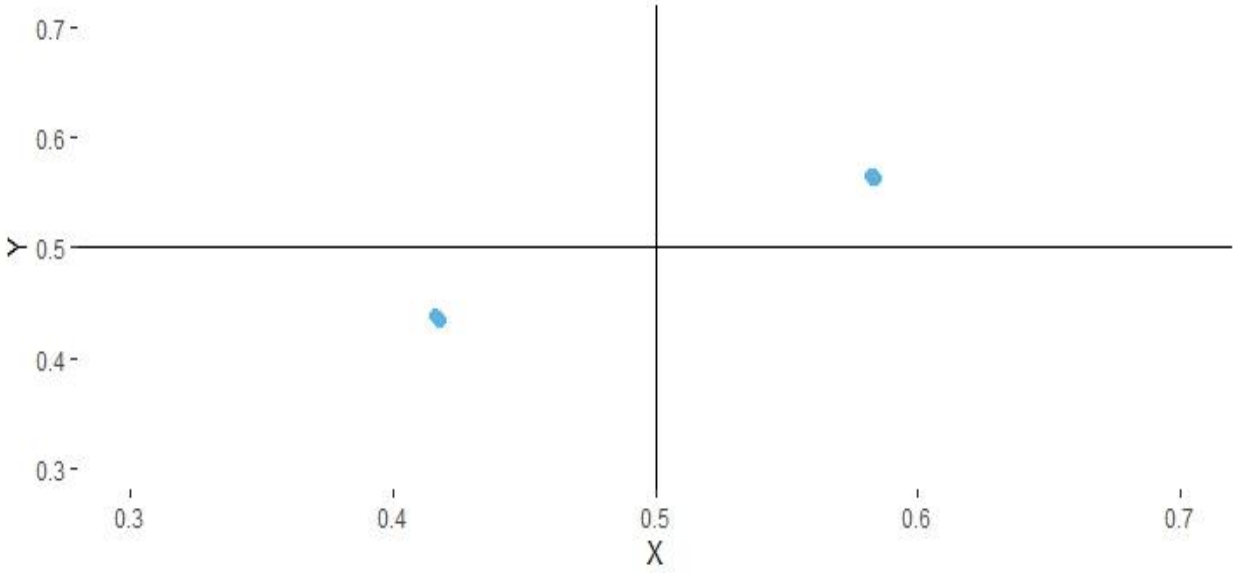


Figure 5.PR2 plot analysis of the complete coding sequences of a) *cya*, b)*lef* and c) *pag* genes of *B.anthracis* genome.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

TABLES

Table 1: Average values of basic parameters of the genes *cya*, *lef* and *pag* of *B.anthracis* genome.

GENE	T3s	C3s	A3s	G3s	GC	GC3s	GC12	ENC	FOP	CAI	CBI
CYA	**0.518135**	**0.110421**	**0.55137**	**0.175095**	**0.292395**	**0.199791**	**0.279251**	**42.00395**	**0.342628**	**0.183256**	**-0.15465**
LEF	**0.499873**	**0.179912**	**0.444628**	**0.219409**	**0.311269**	**0.279957**	**0.303351**	**44.99075**	**0.359344**	**0.191957**	**-0.10678**
PAG	**0.468574**	**0.157086**	**0.456695**	**0.199879**	**0.337864**	**0.264833**	**0.334657**	**43.19379**	**0.336864**	**0.173076**	**-0.15392**

Table 2: Relative dinucleotide frequencies of cya, lef and pag genes subsists in pXO1 plasmid of *B.anthracis*. The odds ratios of over-represented (>1.25) and the under-represented (<0.78) dinucleotides are highlighted using read and yellow, respectively.

	Freq		
Dinucleotide	*cya*	*lef*	*pag*
aa	1.067736	1.028519	1.028633
ac	0.847283	0.775875	0.928263
ag	1.076212	1.037724	0.936159
at	0.908234	1.028644	1.038485
ca	1.007821	1.084556	1.052031
cc	1.793618	1.328525	1.094518
cg	0.549322	0.607539	0.875528
ct	0.967602	0.993393	0.963846
ga	1.048897	1.109865	1.067341
gc	1.098643	1.196667	0.843691
gg	1.151781	1.102134	1.276149
gt	0.795658	0.699643	0.8014
ta	0.863679	0.852506	0.884335
tc	0.861563	1.051828	1.158781
tg	0.966157	1.049464	0.968358
tt	1.278482	1.15476	1.101767

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 5758 - 5774
Received 05 March 2021; Accepted 01 April 2021.

Table 3: Correlation of basic parameters of *cya* gene in *B.anthracis.*

	CAI	CBI	Fop	GC	GC3s	ENc	GC12	Gravy	Aromo
CAI	1								
CBI	-0.5	1							
Fop	0.54	0.43	1						
GC	-0.5	0.76	0.18	1					
GC3s	-0.69	0.79	0.07	0.9	1				
ENc	-0.82	0.62	-0.23	0.79	0.95	1			
GC12	-0.39	-0.38	-0.79	0.14	0.16	0.39	1		
Gravy	-0.83	0.19	-0.75	0.27	0.29	0.47	0.49	1	
Aromo	-0.67	0.04	-0.75	0.42	0.31	0.49	0.71	0.88	1

Table 4: Correlation of basic parameters of *lef* gene in *B.anthracis.*

	CAI	CBI	Fop	GC	GC3s	ENc	GC12	Gravy	Aromo
CAI	1								
CBI	0.42	1							
Fop	0.5	0.99	1						
GC	0.32	0.99	0.97	1					
GC3s	0.31	0.99	0.98	0.99	1				
ENc	0.15	0.96	0.92	0.98	0.98	1			
GC12	-0.08	-0.13	-0.13	-0.12	-0.13	-0.12	1		
Gravy	0.24	0.98	0.95	0.99	0.99	0.99	-0.12	1	
Aromo	0.24	0.98	0.95	0.99	0.99	0.99	-0.12	1	1

Table 5: Correlation of basic parameters of *pag*gene in *B.anthracis.*

	CAI	CBI	Fop	GC	GC3s	ENc	GC12	Gravy	Aromo
CAI	1								
CBI	1	1							
Fop	1	1	1						
GC	1	1	1	1					
GC3s	1	1	1	1	1				
ENc	1	1	1	1	1	1			
GC12	0	0	0	-0.01	0	0	1		
Gravy	1	1	1	1	1	1	0	1	
Aromo	1	1	1	1	1	1	0	1	1