ijlpr 2021; doi 10.22376/ijpbs/lpr.2021.11.5.L35-60

International Journal of Life science and Pharma Research

Research Article

Bioinformatics for health care

Analysis of Codon Usage Pattern and Predicted Gene Expression in Neurospora Crassa: A Novel in Silico Approach

Satyabrata Sahoo

Department of Physics, Dhruba Chand Halder College, Dakshin Barasat, South 24 Parganas,W.B.,INDIA

Abstract: The codon usage pattern of genes has a key role in the gene expression and adaptive evolution of an organism. It is very significant in understanding the role of complex genomic structure in defining cell fates and regulating diverse biological functions. In this paper, we discussed that the codon usage index (CAI_g) based on all protein-coding genes is a promising alternative to the Codon Adaptation Index (CAI). CAI_g which measures the extent that a gene uses a subset of preferred codons relies exclusively on sequence features and is used as a good indicator of the strength of codon bias. A critical analysis of predicted highly expressed (PHE) genes in *Neurospora crassa* has been performed using codon usage index (CAI_g) as a numerical estimator of gene expression level. Analyzing compositional properties and codon usage pattern of genes in *Neurospora crassa*, our study indicates that codon composition plays an important role in the regulation of gene expression. We found a systematic strong correlation between CAI_g and CBI (codon bias index) or other expression-measures. Here, we show that codon usage index CAI_g correlates well with both protein and mRNA levels; suggesting that codon usage is an important determinant of gene expression. Our study highlights the relationship between gene expression and compositional signature in relation to codon usage bias in *Neurospora crassa* and sets the ground for future investigation in eukaryotic biology.

Keywords: Codon usage bias• gene expression • GC content • *Neurospora crassa* • PHE genes • CAI.

***Corresponding Author**

Satyabrata Sahoo , Department of Physics, Dhruba Chand Halder College, Dakshin Barasat, South 24 Parganas,W.B.,INDIA

Received On 29 May, 2021
Revised On 6 August, 2021
Accepted On 11 August, 2021
Published On 3 September, 2021

Funding This work is supported by Engineering Research Board, DST, Govt. of India Grant no. MATRICS [File No: MTR/2019/000274].

Citation Satyabrata Sahoo , Analysis of Codon Usage Pattern and Predicted Gene Expression in Neurospora Crassa: A Novel in Silico Approach.(2021).Int. J. Life Sci. Pharma Res.11(5), 35-60 http://dx.doi.org/10.22376/ijpbs/lpr.2021.11.5.L35-60

1. INTRODUCTION

Since the twentieth century *Neurospora crassa*, a multicellular filamentous fungus, had developed a model experimental organism for contributing to the fundamental understanding of modern genetics and molecular biology [1,2]. Undoubtedly, any useful insight in understanding the expression of functional proteins of *Neurospora crassa* will contribute to the development of eukaryotic biology as well as in the field of modern biotechnology. It is well discussed in the previous studies that the arrangement of genetic codes in a genomic DNA sequence, as well as the choices of synonymous codons, may affect the efficiency and accuracy of mRNA biosynthesis, translational rate, and other biological functions of an organism[3,4]. The codon usage pattern varies significantly between different organisms[5], and also between genes that are expressed at different levels in the same organism[6]. Several hypotheses prevail regarding the factors [7] which influence the codon usage pattern. Codon biases are mainly influenced by mutational pressure [8,9] and natural selection[10-12] due to translation. The other factors known to influence codon biases are protein secondary structures, gene lengths, gene expression levels, hydrophobicity, and aromaticity of encoded proteins, etc[13-15]. The objective of this work is to perform an analysis of codon usage patterns using various codon bias indices and identify highly expressed genes. The previous analyses have shown that codon biases in microbes are generally driven by mutation pressure, whereas selection constraints are the major influencing factors among invertebrate animals[6,16]. Information on the codon usage pattern can provide significant insights into the prediction and design of highly expressed genes. It is generally thought that a balance between mutation and natural selection on translational efficiency is expected to yield a correlation between codon bias and the rate of gene expression. Based on the hypothesis that highly expressed genes are often characterized by strong compositional bias in terms of codon usage, a number of varieties of computational tools [17-25] have been developed so far to provide numerical indices to predict the expression level of genes. Some indices are the effective number of codons (N_c)[18][codon bias index(CBI)[19], and relative synonymous codon usage(RSCU) that measures the deviation from uniform codon usage. The expression measure (E_g)[20] or codon adaptation index(CAI)[17] are based on the knowledge of codon bias of a reference set of highly expressed genes[17-21] while the score of relative codon bias (RCBS)[22,23,25] or the score of modified relative codon bias (MRCBS)[24] has been devised to predict gene expression level from their codon compositions in such a way that the score of the expression indicator may be calculated without any knowledge of previously set selective highly expressed genes as a reference set. Here, we have modified the expression measure by codon adaptation index and investigated the codon usage pattern and gene expression profile of *Neurospora crassa* from a whole-genome perspective without any dependence on the choice of the reference set. The small size of its genome with approximately 43 megabases long makes it a useful model for the computational biologist. The small genome size and the availability of the complete DNA sequence of *Neurospora crassa* have attracted the attention of a wide range of scientists, including evolutionary biologists and biotechnology companies.

2. MATERIALS AND METHODS

The whole-genome sequence of *Neurospora crassa* along with the gene annotations taken from NCBI GenBank have been considered in our study. All gene sequences under study along with those annotated as hypothetical have been extracted from the GenBank Accession Nos: NC_026501.1 (Chromosome 1), NC_026502.1 (Chromosome 2), NC_026503.1 (Chromosome 3), NC_026504.1 (Chromosome 4), NC_026505.1 (Chromosome 5), NC_026506.1(Chromosome 6), NC_026507.1 (Chromosome 7) and NC_026614.1 (Mitochondrion MT). Out of 10784 coding sequences for *Neurospora crassa* , genes with less than 100 codons, internal stop codons, not-translatable codons, and incomplete start and stop codons were excluded from the analysis. Therefore, for the present study, finally, 9700 genes were considered for analysis.

2.1 Analysis of codons usage pattern

Using an in-house Fortran program, we have estimated the overall frequencies of occurrence of the four nucleotides (A, G, C, and T), the occurrence of nucleotides (A, G, C, and T) at the first, second, and third position of all codons, the occurrence of GC at the first (GC1), second (GC2) or third position (GC3), the overall GC contents, AT1, AT2, and AT3, and the frequency of occurrence of each nucleotide at the third position of synonymous codons($A3_s$, $T3_s$, $G3_s$, and $C3_s$) for the analysis of codon usage pattern of genes in *Neurospora crassa* genome. The mutations that usually take place in the third position are mostly synonymous, whereas the mutations occurring in the first or the second positions are known as non-synonymous mutations. When there is no external pressure, mutations should occur in a random rather than in a specific direction. This will result in uniform base composition at three positions of codons. However, in the presence of selective pressure, preference for a particular base would occur in three different positions. GC content at the synonymous third synonymous codon position ($GC3_s$) and GC contents at the first and second synonymous codon positions (GC12) are important determinants to indicate the role of mutation or selective pressure in shaping the codon usage pattern of an organism. $GC3_s$ measures the frequency of G or C at the third position of synonymous codons and can be used as an index of mutation bias on codon usage. It is measured by,

$$GC3_s = \frac{\sum_{(NNS)\in C} f_{NNS}}{\sum_{(NNN)\in C} f_{NNN}}$$

where N= any base, S=G or C, and f_{xyz} is the observed frequency of codon xyz.

The neutrality plot is an analytical method to analyze the influence of mutation bias and natural selection on codon usage. In the neutrality plot, a regression line was plotted between GC12 and GC3. A slope of the regression line is indicative of the mutational force. A regression plot with a slope of zero indicates no effect of directional mutation

pressure, while a slope of 1 indicates complete neutrality[26]. The dinucleotide composition also plays an important role in setting up the codon usage pattern of the genes. Hence, the frequencies of 16 dinucleotides (GpA, GpC, GpG, GpT, CpA, CpC, CpG, CpT, TpA, TpC, TpG, TpT, ApA, ApC, ApG, and ApT) along with their expected frequencies were also calculated for the analysis of compositional bias in genes. The identification of favored dinucleotides and the patterns of dinucleotide usage may affect the selection of codons in genes. The ratio of observed and expected frequencies may be used for the identification of over-or under-represented dinucleotides[27] and is given by,

$$R_{XY} = \frac{f_{XY}}{f_X f_Y}$$

where, fx and f_Y are the frequency of individual nucleotides (x and y, respectively), and fxy is the frequency of dinucleotides (xy) in the same sequence. If the ratio of the observed to expected dinucleotide frequency is more than 1.2, the dinucleotide is considered overrepresented, whereas values below 0.8 indicate an underrepresentation.

In the present study codon usage of the genes has been measured by the following metrics:

a)Relative Synonymous Codons Usage (RSCU) of all genes in *Neurospora crassa* genome have been calculated to describe the synonymous codon usage pattern. RSCU was calculated by determining the ratio of observed usage frequency of a codon to the predicted frequency, given that all codons for a specific amino acid are used equally. Codons showing an RSCU value of 1 means no bias or the codon usage frequency is similar to the expected value, while codons with RSCU values >1 or <1 are showing positive or negative codon bias, respectively. RSCU has been calculated by using the following equation:

$$RSCU = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}}$$

where Xij is the observed number of the i^{th} codon for j^{th} amino acid which has a n_i number of synonymous codons for the amino acid.

b)The Relative Strength of Codons Bias (RCBS) has been proposed to describe the codon usage pattern under the assumption of random codon usage in genes under study. RCBS was calculated by determining the ratio of the observed frequency of a codon to the expected frequency, given that base composition is biased at three sites of all codons in the gene under study. Codons showing an RCBS value of 1 means no codon bias or the codon usage frequency is similar to the expected value, while codons with RCBS values >1 or <1 are showing overrepresented or underrepresented codons (with respect to a randomized sequence) respectively in respect of compositional bias of nucleotides. RCBS has been calculated by using the following equation:

$$RCBS_{xyz} = \frac{f_{xyz}}{f(x)_1 f(y)_2 f(z)_3}$$

where f_{xyz} is the normalized codon frequency of a codon xyz and $f_n(m)$ is the normalized frequency of base m at codon position n in a gene. The ratio of RSCU to RCBS indicates the influence of mutational bias over natural selection in the choice of codons in a gene. The optimal codons are identified as codons with RSCU>1 and RCBS>1, whereas for rare codons both RSCU<0.5 and RCBS<0.5.

c)The Codon Adaptation Index, CAI, a measure of codon bias[17] based on RSCU values of the codons in reference to a set of highly expressed genes is given by,

$$CAI = \left(\prod_{1}^{N} w_i \right)^{\frac{1}{N}}$$

where, N is the number of codons in the gene and relative adaptiveness of a i^{th} codon, w_i is defined as

$$w_i = \frac{(RSCU)_i}{(RSCU)_{i,max}}$$

$RSCU_i$ is the RSCU value of the i^{th} codon for j^{th} amino acid and $RSCU_{i,max}$ is the RSCU value of the most frequent codon used for encoding j^{th} amino acid. The score measured by CAI ranges from 0 to 1 indicating that the higher the CAI values, the genes are more likely to be highly expressed. CAI was proposed as a measure of codon bias of a gene relative to a highly expressed reference set of genes. Although this method has been applied successfully for the prediction of highly expressed genes in a query genome sequence, it relies on the prior definition of the reference set of highly expressed genes.

d)In the present work, we propose to introduce an alternative methodology to calculate the codon usage index (CAI_g) of a gene from a whole-genome perspective to study the codon usage pattern of an organism.

$$CAI_g = \prod_{i=1}^{N} (S_i)^{\frac{1}{N}}$$

where S_i is impact score of i^{th} codon defined as

$$S_i = \frac{F_{ij}}{F_{max,j}}$$

where Fij is the number of occurrences of i^{th} codon for a j^{th} amino acid which has n_i number of synonymous codons for the whole set of coding sequences in a genome and $F_{max,j}$ is the observed number of the most frequent codon used for encoding j^{th} amino acid. N is the codon length of a gene.

The numerical value computed by this method may be used to rank the set of genes with respect to codon bias towards gene expression. Finally, statistical analyses were performed to identify the relationship among the overall nucleotide compositions , codon usage bias, and the expression levels of genes. A diagrammatic representation of the methodology to predict the highly expressed (PHE) genes in an organism has been depicted in Fig.1.

3. RESULTS AND DISCUSSION

In the present study, we have analyzed the codon usage pattern of the *Neurospora crassa* genome with respect to nucleotide compositions at synonymous and nonsynonymous sites of the codons, dinucleotide composition, and codon usage of the genes. By calculating the RCBS[22] and RSCU[17] score of 61 codons, we have measured the codon usage bias of all protein-coding genes in the genome under study. The codon bias index and expression level of all protein-coding genes was calculated by CAI_g and compared with other codon usage models like CAI[17], CBI[19], and N_C[18]. In reference to a set of highly expressed genes of *Escherichia coli*, and *Saccharomyces cerevisiae* CAI has been calculated by using CodonW (available at http://sourceforge.net/projects/codonw). In order to study the factors influencing the codon usage pattern of the genes in an organism, it is essential to study the overall nucleotide composition and the compositional feature at different nucleotide positions in genes. For the genome under study, the average GC value of the genes lies between 0.708 to 0.204 [Fig. 2a] and the $GC3_s$ score varies between 0.968 to 0.019 [Fig. 2b]. Many researchers have argued that GC content or $GC3_s$ may be viewed as the primary influence on the codon usage pattern [28-30] and thus on the expression profile [31-33]. Moreover, the preference for one type of codon over another can be greatly influenced by the nucleotide composition of the genome. We first analyzed nucleotide composition and observed that the nucleotides C and G were higher and followed by A and T [Table 1]. The *Neurospora crassa* genome is rich with C content having a mean value of 29.10. For a better understanding, we analyzed nucleotide composition at synonymous and non-synonymous sites of codon and observed the dominance of G1, A2, and $C3_s$ nucleotides with a mean value of 41.54,42.39, and 46.02 respectively. Although the percentage of $GC3_s$ is higher compared to GC content, $AT3_s$ is less compared to AT (respective mean values for GC, AT, $GC3_s$ and $AT3_s$ being 56.17, 43.83 64.40, and 35.60). Dinucleotide composition may have consequences on the intrinsic characteristics of the codon usage pattern[27]. In the case of the *Neurospora crassa* genome, the calculated frequency ratios did not deviate much from 1 for most dinucleotides, but there are some exceptions [Fig. 3]. The dinucleotides TpC and GpA showed slight over-representation in *Neurospora crassa*. The presence of relatively abundant TpC reflects a high abundance of C in the genome, whereas, TpA was, the least abundant dinucleotide pair with the lowest odds ratio. Among other dinucleotides, mild suppression of GpC and CpG has been observed. The ratio CpG/GpC may be important in estimating the role of the evolutionary process and mutational pressure acting upon constituent nucleotides[34]. Codon usage profile of the *Neurospora crassa*_genome has been described in terms of relative synonymous codon usage, RSCU, and RCBS of 9730 complete protein-coding sequences (length>150) of the genome[Fig. 4]. Although most of the amino acids can be specified by more than one codon, it is hypothesized that only a subset of potential codons is used in highly expressed genes. RCBS and RSCU of 61 codons have been displayed in Table 2. The codons having RCBS greater than 1.0 are preferred codons for increasing the translational efficiency of the protein-coding genes, whereas codons having RCBS greater than 1.0 are overrepresented codons for the organism under study. The preferred codons in *Neurospora crassa* are found to be used in coding Asn (AAC), Lys (AAG),Thr (ACC), Arg (AGG,CGC), Iln (ATC), His (CAC), Gln (CAG), Pro (CCC), Leu (CTC,CTG,CTT,TTG), Asp (GAC), Glu (GAG), Ala (GCC), Gly (GGC),Val (GTC,GTG),Tyr (TAC), Cys (TGC), Phe (TTC,TTT). Importantly, these codons reflect a simple compositional bias. Most of the preferred codons have C or G at the 3rd codon position. Out of 22 preferred codons, 14 are C-ending codons. Whereas, Lys (AAA,AAG), Thr (ACA), Arg (AGA,CGA,CGC), Ser (AGC,TCA,TCC,TCG, TCT), Iln (ATC,ATT), Met (ATG), Gln (CAA,CAG), Pro (CCA,CCT), Leu (CTC,CTG,CTT,TTG), Asp (GAT), Glu (GAA,GAG), Ala (GCC,GCT), Gly (GGC,GGT), Trp (TGG), Phe (TTC,TTT) are the overrepresented codons. Although different synonymous codons favoured by an organism for translational efficiency in different genes are identified by RSCU, the set of optimal codons used in a gene effectively measures its expressivity. The optimal codons enhance the rate of elongation while non-optimal codons slow it down[22].In the present study, we observed that AAG(Lys), ATC(Iln), CAG (Gln), CGC (Arg), CTC, CTG,CTT,TTG (Leu), GAG (Glu), GCC (Ala), GGC (Gly), TTC,TTT (Phe), are optimal (RSCU>1 and RCBS>1) whereas, rare codons (RSCU<0.5 and RCBS<0.5) are ATA(Iln),GTA (Val), TTA

(Leu). These rare codons have A at the 3rd codon position. The low relative abundance of TpA in rare codons, and the relatively high abundance of TpC, TpT, and TpG in the choice of preferred codons indicate the influence of selection pressure in codon usage bias in the *Neurospora crassa* genes. The low relative abundance of TpA is reflected in the set of rare codons which are associated with a generally slower rate of protein synthesis. The codon optimality has been shown to affect mRNA stability due to its role in affecting translation elongation[35]. To explore the amino acid usage trend in *Neurospora crassa* genes, we calculated the number of each amino acid for all ORFs across the genome. A wide difference in amino acid usage was observed among genes. The mean amino acid usages of alanine, leucine, serine, and glycine were high for the novel virus, while amino acids such as cysteine, histidine, tyrosine, methionine, and tryptophan were low [Fig. 5]. Expression profiles of the genes are determined by calculating CAI_g for each gene and their distributions are shown in Figure 6. The majority of genes (90%) have CAI_g lying between 0.645 and 0.845 with a mean and standard deviation of 0.725 and 0.046 respectively. The threshold score for identifying highly expressed genes has been determined by the z score of CAI_g values of the gene under study. The corresponding genes having a z score greater than 2.00 are assumed to be PHE genes and the threshold score of CAI_g has been calculated to be 0.848. Only 5% of genes of the *Neurospora crassa* genome have CAI_g values greater than 0.848. The overall variation of GC or GC3 content of the genes is depicted in Figure 1 and Figure 2 respectively. It indicates that the majority of genes (93%) have a GC3 score lying between 0.5 to 0.8 and 90% of genes have GC content lying between 0.50 to 0.62. Table 3 displays the statistics of PHE genes and the top 20 PHE genes of the *Neurospora crassa* genome along with their functions and gene expression level(CAI_g). N_C is a measure of bias from equal codon usage in a gene. N_C values were calculated to determine the inter-genic codon bias. In our study, we observed that N_C of *Neurospora crassa* genes ranged from 33.57 to 47.05 with a mean value of 40.89±7.02. The more biased a gene is, the smaller is the N_C value. Only 5% of Neurospora genes have N_c<35. To clarify the effects of mutation pressure and natural selection, N_C–GC3 plot is constructed for all protein-coding sequences of the genome. The clustering of points below the expected curve[Fig. 7] indicates that natural selection plays a dominant role in defining the codon usage variation among those genes. We also observe that some of the data points lie around the expected curve indicating that not only the natural selection but also other factors, such as mutation, are likely to be involved in determining the codon usage in *Neurospora crassa* . The magnitude of both forces has been investigated by constructing a neutrality plot(GC12 vs $GC3_s$) [Fig. 8]. The weak correlation (r=-0.357) between $GC3_s$ and GC12 suggests that codon usage is influenced by natural selection. The mutations are independent single-site events. However, the base frequencies of the third codon position in *Neurospora crassa* genes are influenced by bases at the first and second codon positions. In the neutrality plot [Fig. 8], most of the genes are far from the regression line. The slope (-0.158) of the regression line indicates the relative neutrality (mutation pressure) was 15.8%, and the relative constraint on $GC3_s$ (natural selection) was 84.2%, indicating the dominant influence of natural selection on the codon usage patterns of *Neurospora crassa.* To find out the putative relationship of base composition at different synonymous codon positions, when the values were compared at third synonymous codon positions, the significant correlations were observed only between $A3_s$ and $C3_s$ (r =-0.800), $C3_s$ and $GC3_s$(r=0.854),and $A3_s$ and $AT3_s$(r=0.853). Moreover, a significant positive correlation between GC_l and GC2(r=0.877), GC1 and GC3 (r = 0.871), GC1 and GC(r=0.975) suggested that selection force along with mutational pressure both have significantly influenced the codon usage pattern in *Neurospora* genes. The large data set analyzed here revealed that selective constraints play a major influencing factor towards a strong codon usage bias of different sets of preferred codons in genes with high cytoplasmic mRNA levels. In contrast, genes with low mRNA levels showed very little synonymous codon usage bias. Codon usage bias was proposed as a result of translational selection, since using a codon that is translated via an abundant tRNA species was hypothesized to boost translational efficiency. Codon frequencies are found to vary between genes in the same genome. The standard version of the genetic code includes 61 sense codons and three stop codons. Although almost all organisms have made the same codon assignments for each amino acid, the preferred use of individual codons varies greatly among genes. The overall nucleotide composition of the genome which influences the codon usage pattern introduces selective forces acting on highly expressed genes to improve the efficiency of translation. It is now widely accepted that synonymous codon preferences in a unicellular organism are affected by the cellular amount of isoacceptor tRNA species. However, many tRNAs can translate more than one codon, but with the variable ability and it is suggested that impact codons have favored translational efficiency and the highly expressed genes use a preferred set of optimal codons. We observed that only 5% of genes in *Neurospora crassa* belong to PHE genes. The preferred set of codons used in these genes are C3 rich and the codons with A3 are rarely used in highly expressed genes. In fact, the correlation between CAI_g and GC content is not significant(r=0.539). However, a strong negative correlation between CAI_g and N_c (r=-0.933)[Fig. 9] suggests that highly expressed genes display more biased codon usage than the lowly expressed genes. We observed that PHE genes of *Neurospora crassa* mostly include ribosomal protein(RP) genes, translation initiation factors, translation elongation factors, transcription factors, chaperon, heat shock protein, histone, and many binding protein genes. The top 20 genes with the highest predicted expression levels for *Neurospora crassa* genomes are displayed in Table 2. Our analysis predicted 473 highly expressed genes in *Neurospora crassa*. A list of well-characterized PHE genes has been displayed in Table-4. It is worth noticing that these genes are separated into different functional categories. Table-4 displays a set of well-characterized PHE genes segregated into different functional categories. It has been observed that PHE genes belonged to various functional classes and variably represented in the genome. These include oxidase, reductase, peroxidase, hydrolase, dehydratase, dehydrogenase, oxidoreductase, protease, metalloprotease, transaminase, aminoacyl- tRNA synthetase, ligase, transferase, mutase, scaffold/adaptor protein, hydrophobin, clock-controlled protein, DNA directed DNA/RNA polymerase, desaturase, cell wall protein, membrane protein, mitochondrial protein, ATP-dependent RNA helicase, ATP synthetase, transporter, and transfer/ carrier protein. Besides PHE genes also include transaldolase, transketolase, RNA processing factor, carbohydrate kinase, nucleotide kinase, protease inhibitor, glycosidase, Iron-sulfur cluster assembly protein, neuronal calcium sensor, intermembrane space

import and assembly protein, tryptophan synthetase, tubulin, zinc finger protein, calmodulin, cytochrome c1, cytochrome b-c1 complex, cell division control protein, and several uncharacterized genes. However, a fraction of poorly characterized hypothetical genes were also found among the PHE genes. Genes of unknown function with high predicted expression levels may be attractive candidates for experimental characterizations. The characteristic codon distribution of these genes indicates that they may have important functions in these organisms. A variety of PHE genes encoding proteins of unknown function may provide targets for the identification of additional key features of the organism.

3.1 Correlations among different codon bias indices

In this study, we compared the performances of several commonly used computation tools for predicting gene expression levels[17-25]. The expression profiles of the *Neurospora crassa* genome were analyzed in terms of CAI, N_c, MRCBS, CBI, and CAI_g. The CAI scores have been calculated in reference to *e.coli(CAI-1)* and *S. cerevisiae(CAI-2)*, and CBI scores have been calculated in reference to the genome under study. The results indicate that CAI scores depend on the reference set of highly expressed genes[25]. The correlation of CBI with CAI-1 is 0.759 [Fig.10]and that with CAI-2 is 0.700 [Fig.11], whereas the correlation between CAI_g and CAI is 0.741 [Fig.12], and that with CAI-2 is 0.640 [Fig.13]. Compared to CBI and CAI_g, better correlations have been observed between MRCBS and CAI-1(r=0.767) and that between MRCBS and CAI-2(r=0.832). The correlation between MRCBS and CAI_g is 0.830, whereas that with CBI is 0.845. The novel method of quantitatively predicting gene expressivity CAI_g is then compared with CBI and the correlation between them is found to be surprisingly good (r=0.969)[Fig. 14]. The correlation of the codon usage index with Nc is very much significant. The correlation of N_C with CAI-1 is -0.657, with CAI-2 is -0.570, and with MRCBS is -0.728. The strong negative correlation between CAI_g and Nc (r=-0.933) [Fig.9] compared to CBI and N_c (r=-0.923) [Fig.15] indicates that synonymous codon preferences have been taken into account in CAI_g. These correlation coefficients can be used to express the strength of the existing prediction methods[36]. It can be seen that CAI_g consistently yields a better correlation than others. We also observe that there are clear correlations between CAI_g with GC3(r=0.853), C3(r=0.941) and A3(r=-0.901) [Fig. 16-18], where correlation with GC is not much significant (r=0.539). So, GC3, C3 or A3, not GC content may be the accurate representation of the trend in codon usage bias. Similarly, no correlation between the length of the gene CAI_g has been observed in our study.

3.2 Correlation of protein and transcript levels with codon usage bias

In this study, we compared our results with the experimental datasets[37]. The value of codon-based expression indicators can perhaps be appreciated by comparing results with the experimental gene expression data in general. The expression data that we have used in this study stems from mass spectrometry experiments[37] which led to the identification and quantification of about 3200 proteins based on their emPAI(exponentially modified protein abundance index) values. These emPAI values are proportional to their relative abundances in a protein mixture[38]. In addition, we have compared our results with mRNA levels obtained by RNA-sequencing(seq) analyses of *Neurospora* mRNA to determine correlations between mRNA levels with codon usage biases. We have collected a set of about 3200 selected genes for which the experimental data set of relative protein abundance and mRNA levels can be generated along with the codon-based expression indicator. To assess CAI_g for predicting protein expression levels, we plotted the two experimental sets of data versus CAI_g. The distribution patterns for both the protein expression data with respect to these expression indicators are highly similar to CBI[37]. For these datasets, the predicted gene expression level using CAI_g value is found to correlate well with emPAI values (r= 0.593)[Fig. 19]. The correlation is better than the quantitative measure of CAI-1 (r =0.497), CAI-2 (r=0.452), N_c(r=-0.585), GC3(r=0.469) and GC(r=0.324), whereas the correlation with CBI (r=0.595)[Fig.20] is comparable to CAI_g(r=0.593). It suggests that a quantitative estimate of the expression level by CAI_g values performs better than other indices of expression-measure. The novel method of quantitatively predicting gene expressivity is then compared with mRNA levels[37]. We observe that the correlation coefficient of mRNA levels with CAI_g (r = 0.560) is good[Fig. 21] which is consistent with protein abundance data. In fact, the pairwise correlation coefficient among the gene expression levels from two experimental datasets (r=0.704) is good and it can be clearly seen that the agreement of predicted and actual protein expression levels quantified by mRNA levels varied greatly between all examined combinations of prediction method and data set ($r_{CAI\text{-}1}$=0.487,$r_{CAI\text{-}2}$=0.490, r_{GC3}=0.395, r_{GC}=0.248 and rN_c=-0.538[Fig. 22], r_{CBI}=0.579) [Fig. 23]. Comparing the performance of CAI_g, CAI, CBI, GC3, and N_c as numerical indices of the gene expression level in terms of the Pearson correlation coefficient with the expression data, we observed that CAI_g may be a potential tool in estimating gene expression level. Our study demonstrates that CAI_g may be a useful tool for predicting highly expressed genes. The idea of developing our method is based on the hypothesis that the codon usage pattern is largely responsible for the regulation of gene expression which can occur during transcription[39,40] or at the level of protein translation. Although the concept of predicting gene expression level from the codon usage pattern was proposed a decade ago, only recently these methods have been successfully applied to the identification of highly expressed genes in various bacteria and eukaryotic genomes. There has not been an adequate study of codon usage patterns in the *Neurospora crassa* genome. The codon usage bias of Neurospora genes, the codon bias index (CBI) for every protein-coding gene in the genome was calculated by Zhou et al. and they reported significant positive correlations of relative protein abundances and mRNA levels with CBI. However, CBI measures need a reference set of identified 'optimal' or 'preferred' codons, which are dominantly used in previously identified highly expressed genes and the scores depend on the reference set, whereas CAI_g has been calculated in respect of all genes in a genome. CAI_g was introduced as an alternative method of calculating the codon adaptation index for correcting the codon bias of background nucleotides of other genes in a genome. The improved reliability of CAI_g for estimating expression levels in the *Neurospora crassa* genome thus makes this index a superior choice for undertaking and benchmarking predictions of gene expression. In this study, various approaches to estimating gene expression levels based on codon usage have been applied to the *Neurospora crassa* genome with the objectives

of testing the present alternative method of studying whole-genome gene expression. Our results demonstrate significant heterogeneity in codon usage among genes in the *Neurospora crassa* genome. Furthermore, the predicted gene expression level using the quantitative measure by CAI_g was found to correlate well with CBI and N_c. The strong negative correlation between CAI_g and N_c supports the hypothesis that highly expressed genes are strongly biased. In addition, since the expression levels measured by relative protein abundances and mRNA levels detected by mass spectrometry represent the accumulated results of expression and degradation, the results from this computational approach could be used as reference data for calibrating and better interpreting experimental data. For example, observation of a low level of expression from proteomic data for a gene with a high PHE index might suggest the possible involvement of degradation in regulating the expression levels of that gene. Although most of the PHE genes are essential genes responsible for the habitat, energy sources, and lifestyle of an organism, the study also identified several functionally unknown genes as PHE genes based on their codon usage profile. Further investigation of these genes by an integrated computational and experimental approach will enhance our knowledge of metabolism. Given that a large volume of experimental data is available, such a novel method may help extract meaningful information for understanding the details of functional genomics.

TABLE 1 - Characteristics of top 5% genes with the highest predicted expression levels and last 5% genes with the lowest predicted expression levels for *Neurospora crassa* genome.

	Top 5% genes	Lowest 5% genes	Total Number of Genes under study		Top 5% genes	Lowest 5% genes	Total Number of Genes under study
Average length	1010	853	1515	Average A3(%)	3.75	31.68	20
Average A(%)	20.71	25.54	23.53	Average C3(%)	53.97	32.74	42.16
Average C(%)	34.27	25.12	29.1	Average G3(%)	32.22	33.6	35.1
Average G(%)	26.08	26.7	27.07	Average T3(%)	29.74	35.74	32.91
Average T(%)	18.94	22.63	20.3	Average AT(%)	39.65	48.17	43.83
Average A1(%)	43.42	34.86	37.6	Average GC(%)	60.35	51.83	56.17
Average C1(%)	21.35	31.64	27.09	Average $AT3_s$(%)	19.61	50.4	35.6
Average G1(%)	46.54	38.16	41.54	Average $GC3_s$(%)	80.39	49.6	64.4
Average T1(%)	25.98	27.8	27.05	Average CAI_{ecoli}	0.321	0.182	0.236
Average A2(%)	52.82	33.45	42.39	Average CAI_{yeast}	0.154	0.08	0.097
Average C2(%)	24.68	35.63	30.75	Average CAI_g	0.875	0.626	0.728
Average G2(%)	21.23	28.24	23.36	Average MRCBS	0.909	0.833	0.861
Average T2(%)	44.28	36.46	40.04	Average N_c	31.38	56.83	50.26

TABLE-2 The RCBS and RSCU of 61 codons of *Neurospora crassa* genes understudy.

Codon	RCBS	RSCU	Codon	RCBS	RSCU	Codon	RCBS	RSCU
GCA	0.9596	0.60011	GGC	1.26039	1.59883	CCG	0.81956	0.91665
GCC	1.06474	1.62778	GGG	0.64477	0.64249	CCT	1.16409	0.94747
GCG	0.67104	0.80586	GGT	1.35588	0.98318	AGC	1.02138	0.85246
GCT	1.10565	0.96624	CAC	0.56172	1.20061	AGT	0.89463	0.42681
AGA	1.12415	0.78188	CAT	0.65428	0.79939	TCA	1.49075	0.4586
AGG	0.86415	1.15419	*ATA*	*0.4539*	*0.28812*	TCC	1.27584	0.95948
CGA	1.09707	0.70049	ATC	1.13876	1.76706	TCG	1.21156	0.71572
CGC	1.08516	1.69385	ATT	1.06516	0.94482	TCT	1.36531	0.58693
CGG	0.67859	0.83204	CTA	0.7043	0.43592	ACA	1.06638	0.7278
CGT	0.93866	0.83754	CTC	1.25163	1.89381	ACC	0.96922	1.6171
AAC	0.92111	1.42302	CTG	1.11049	1.31988	ACG	0.69318	0.90849
AAT	0.65335	0.57698	CTT	1.19563	1.03412	ACT	0.78284	0.74661
GAC	0.86704	1.14004	*TTA*	*0.49395*	*0.20476*	TGG	1.58244	1
GAT	1.14416	0.85996	TTG	1.39632	1.11151	TAC	0.9636	1.32585

TGC	0.70029	1.35878	AAA	1.02146	0.47173	TAT	0.85712	0.67415
TGT	0.57812	0.64122	AAG	1.72327	1.52827	*GTA*	*0.45733*	*0.37268*
CAA	1.60917	0.80855	ATG	1.19882	1	GTC	0.81463	1.62283
CAG	1.23482	1.19145	TTC	1.52597	2.54449	GTG	0.67784	1.06073
GAA	1.52356	0.71286	TTT	1.52702	1.45551	GTT	0.82877	0.94377
GAG	1.43256	1.28714	CCA	1.33633	0.77834			
GGA	0.97111	0.77549	CCC	0.95342	1.35753			

TABLE-3 Codon adaptation index(CAI),GC content, A,C,G,T, and GC3 at 3rd position of synonymous codons, effective number of codons (N_c) and length of top 20 genes with the highest predicted expression levels and last 20 genes with lowest predicted expression level for *Neurospora crassa* genome. CAI-1 is the Codon adaptation index using reference set of highly expressed genes of *Escherichia coli* and CAI-2 is the Codon adaptation index using reference set of highly expressed genes *Saccharomyces cerevisiae*.

Top 20 Genes

Locus Tag/Gene Name	Function	NN	GC3s	GC	T3s	C3s	A3s	G3s	CAI-1	CAI-2	Nc	CAI g	CBI
NCU04114	Uncharacterized protein	216	0.889	0.601	0.130	0.852	0.000	0.233	0.366	0.155	26.22	0.941	0.787
NCU03565	Ribosomal protein L26	411	0.91	0.632	0.110	0.881	0.000	0.245	0.293	0.174	24.58	0.937	0.869
NCU06464	Translationally-controlled tumor protein	513	0.877	0.578	0.159	0.833	0.000	0.339	0.343	0.231	26.16	0.933	0.835
NCU00399	Cell wall protein PhiA	597	0.968	0.695	0.036	0.837	0.000	0.310	0.327	0.089	26.47	0.932	0.719
tca-15	Malate dehydrogenase	1011	0.873	0.642	0.134	0.848	0.018	0.210	0.271	0.166	24.37	0.931	0.86
qcr7	Cytochrome b-c1 complex subunit 7	372	0.906	0.645	0.118	0.828	0.000	0.315	0.274	0.103	26.41	0.931	0.838
NCU07829	60S ribosomal protein L7	747	0.872	0.582	0.160	0.812	0.010	0.366	0.338	0.158	26.18	0.927	0.857
NCU06431	40S ribosomal protein S22	393	0.864	0.592	0.168	0.822	0.000	0.281	0.365	0.167	28.6	0.927	0.857
eat-5	transcript-5 protein	411	0.884	0.65	0.122	0.841	0.021	0.255	0.298	0.102	28.93	0.926	0.822
rhd	Mitochondrial peroxiredoxin PRX1	678	0.881	0.619	0.128	0.840	0.012	0.238	0.311	0.167	25.6	0.926	0.798
NCU09707	Eukaryotic translation initiation factor 3 subunit K	714	0.9	0.634	0.098	0.821	0.028	0.335	0.295	0.095	27.61	0.924	0.737
cox-4	Cytochrome c oxidase subunit 4, mitochondrial	561	0.865	0.667	0.151	0.822	0.007	0.220	0.292	0.141	26	0.922	0.825
cpc-2	Guanine nucleotide-binding protein	951	0.87	0.614	0.142	0.824	0.009	0.215	0.36	0.181	26.24	0.920	0.82
NCU06661	60S ribosomal protein L22	381	0.846	0.566	0.211	0.778	0.000	0.370	0.367	0.215	26.85	0.919	0.824
xr	Xylose reductase	969	0.855	0.61	0.161	0.796	0.017	0.284	0.33	0.13	26.74	0.918	0.774
crp-3	40S ribosomal protein S17	441	0.875	0.619	0.159	0.805	0.000	0.327	0.342	0.144	26.78	0.916	0.851
NCU08951	H/ACA ribonucleoprotein complex subunit 2	729	0.898	0.61	0.127	0.777	0.020	0.468	0.258	0.155	29.14	0.915	0.774
NCU06279	Eukaryotic translation	1425	0.869	0.599	0.138	0.801	0.024	0.308	0.316	0.139	26.85	0.914	0.748

	initiation factor 3 subunit L												
NCU08894	Glutamyl-tRNA synthetase	1911	0.876	0.615	0.149	0.783	0.007	0.351	0.317	0.136	27.99	0.914	0.759
NCU06346	ACB domain-containing protein	306	0.835	0.587	0.229	0.743	0.000	0.387	0.335	0.221	30.53	0.914	0.725

TABLE-3 Codon adaptation index(CAI),GC content, A,C,G,T, and GC3 at 3rd position of synonymous codons, effective number of codons (N_c) and length of top 20 genes with the highest predicted expression levels and last 20 genes with lowest predicted expression level for *Neurospora crassa* genome. CAI-1 is the Codon adaptation index using reference set of highly expressed genes of *Escherichia coli* and CAI-2 is the Codon adaptation index using reference set of highly expressed genes *Saccharomyces cerevisiae*.

		Lowest 20 genes											
GENE	**Function**	**NN**	**GC3s**	**GC**	**T3s**	**C3s**	**A3s**	**G3s**	**CAI-1**	**CAI-2**	**Nc**	**CAI$_g$**	**CBI**
NCU16375	Uncharacterized protein	399	0.433	0.465	0.333	0.232	0.367	0.330	0.21	0.076	0.082	0.583	-0.132
NCU03707	Uncharacterized protein	312	0.49	0.495	0.296	0.261	0.281	0.324	0.136	0.09	0.076	0.583	-0.091
NCU16591	Uncharacterized protein	405	0.386	0.48	0.294	0.353	0.472	0.143	0.255	0.067	0.09	0.581	-0.069
NCU17014	Uncharacterized protein	327	0.452	0.491	0.266	0.255	0.410	0.303	0.18	0.029	0.067	0.580	-0.147
NCU17185	Uncharacterized protein	210	0.531	0.551	0.250	0.179	0.281	0.453	0.129	0.085	0.029	0.580	-0.18
NCU16634	Uncharacterized protein	474	0.458	0.45	0.252	0.304	0.446	0.297	0.174	0.099	0.085	0.580	-0.01
NCU16477	Uncharacterized protein	525	0.393	0.42	0.413	0.219	0.307	0.294	0.176	0.113	0.099	0.579	-0.108
NCU16352	Uncharacterized protein	354	0.33	0.422	0.321	0.214	0.512	0.215	0.18	0.106	0.113	0.576	-0.222
NCU01043	Uncharacterized protein	300	0.441	0.475	0.325	0.275	0.329	0.260	0.165	0.086	0.106	0.575	0.054
NCU16681	Uncharacterized protein	303	0.454	0.423	0.371	0.281	0.299	0.312	0.18	0.053	0.086	0.573	-0.134
NCU00516	Uncharacterized protein	474	0.427	0.522	0.269	0.231	0.395	0.279	0.182	0.083	0.053	0.571	-0.131
NCU16363	Uncharacterized protein	321	0.38	0.484	0.348	0.225	0.388	0.234	0.127	0.046	0.083	0.570	-0.114
NCU04126	Uncharacterized protein	303	0.462	0.577	0.211	0.156	0.356	0.341	0.101	0.057	0.046	0.564	-0.162
NCU02876	Uncharacterized protein	201	0.361	0.52	0.250	0.212	0.510	0.220	0.101		0.057	0.558	-0.298
NCU16022	Uncharacterized protein	546	0.215	0.366	0.490	0.140	0.449	0.138	0.172		0.136	0.481	-0.194
NCU16019	Uncharacterized protein	915	0.121	0.253	0.654	0.070	0.490	0.103	0.189		0.152	0.457	-0.322
NCU16007	Uncharacterized protein	444	0.148	0.302	0.536	0.080	0.529	0.124	0.159		0.098	0.448	-0.30

												7
NCU16027	Uncharacterized protein	225	0.155	0.356	0.394	0.076	0.548	0.113	0.111	0.066	0.410	-0.358
NCU16024	Uncharacterized protein	165	0.019	0.204	0.652	0.022	0.525	0.000	0.162	0.130	0.397	-0.279
NCU16008	Uncharacterized protein	270	0.094	0.251	0.443	0.089	0.592	0.020	0.092	0.091	0.381	-0.323

TABLE-4 A list of well characterized PHE genes in *Neurospora crassa* genome segregated into different functional categories

Protein/protein family/Protein Class	Gene Id/Gene Name
Ribosomal protein	NCU00294, NCU05804, NCU09089, NCU07057, NCU07826, NCU00315, NCU08963, NCU06843, crp-4, NCU06047, NCU02707, NCU02905, NCU00971, NCU03038, NCU01776, NCU01452, NCU05235, NCU00706, NCU02509, NCU07829, NCU06226, NCU00475, NCU03988, NCU11321, NCU03635, NCU05599, NCU03102, NCU01552, crp-3, NCU01948, NCU08620, NCU10498, NCU01221, NCU06469, NCU01827, NCU06892, NCU08389, NCU07562, NCU00464, NCU09476, NCU03757, crp-5, NCU00634, NCU01317, NCU08960, crp-7, NCU00618, NCU06066, NCU02181, NCU08951, NCU07857, NCU08344, NCU08990, NCU05813, NCU04779, NCU00979, un-25, NCU08964, NCU03738, NCU06210, rap-1, NCU07182, crp-10, NCU03703, NCU06743, NCU03565, NCU06661, NCU02744, NCU06432, crp-15, NCU08502, NCU01966, NCU08500, NCU09475, NCU03150, NCU06431, NCU07408, ubi::crp-6, NCU05275
Transcription/translation/ elongation /initiation factor	NCU07922, tef-1, NCU03737, NCU05270, NCU03826,NCU07831, NCU06279, NCU06307, NCU01021, hex-1, NCU02810, NCU05889, NCU07929, NCU07437, NCU02813, NCU02076, NCU06035, NCU07954, NCU00366, cot-3, NCU07380, NCU08920, NCU05274, NCU02208, NCU04640, NCU09707, NCU02955, gst-4, NCU07420, NCU03148, NCU00635,
Histone/chaperone/Heat shock protein/ chaperonin	hsp80, NCU04334, NCU01740, NCU05778, hsp88, hsp70-5, fkr-4, fkr-2, csr-1, NCU09700, hH3, NCU01200, hH2B, NCU09265, NCU09223, NCU03009, hsp70-2, NCU09602, hH4-1, hH1, hsp60, NCU00692
Oxidoreductase/dehydrogenase/reductase	NCU04098, NCU08402, fdh, NCU02407,xr,adh-1, gpd-1, qcr7, NCU06652, NCU00904, tca-10, tca-7, tca-4, NCU03415, NCU01824, ncw-4, leu-1, qcr8, NCU08272, NCU04462, NCU04823, NCU10029, tca-12,tca-5, gcy-1, NCU06543, cys-2, sod-1, NCU03233, NCU03362, ndi-1, am, ppm-2, mig-4, NCU03603, NCU07887, tca-16, NCU03112, NCU02580, NCU00891, NCU03748, mig-2, trx, tca-15, cys-4, NCU05989, acd-2, acd-1, NCU04768, tca-6, ace-2, NCU03031, NCU03935,
Binding protein	cpc-2, tpm, NCU01290, NCU04799, NCU01587, NCU06464, NCU00243, NCU06397, NCU05289, ypt-1, NCU03092, NCU16466, NCU03600, ran, NCU08923
Transferase/acyltransferase/methyltransferase/glycosyltransferase	NCU05680,gel-4, NCU06781, NCU08002, NCU05290, met-8, fpp, cut-1, NCU08976, spe-3, acu-9, NCU00168, cit-1, erg-4, gel-3, for, lys-5, NCU06694, NCU04796, dpm, gst-1, NCU09646, NCU07659, eth-1, NCU05301
Oxidase/peroxidase	NCU03340, rhd, cya-4, NCU05816, NCU06741, cox-6, cat-3, NCU08931, cox-4, NCU06402, NCU04108, eat-5, NCU03297, NCU04114
Aminoacyl-tRNA synthetase	NCU04020, NCU08888, NCU02380, NCU08894, NCU05095, NCU08195, NCU07926, NCU06457, NCU01443, NCU04449, NCU07451, leu-6, NCU06914, NCU03575, NCU07755,
ATP synthase	NCU03199, NCU00636, NCU00644, oli, NCU00502, NCU08093, NCU01606, NCU09119, des, NCU05220,
Ligase	tca-8, NCU07982, arg-1, arg-3, tca-9, NCU04303, NCU00261, gua-3, acu-5, NCU09789, NCU10477
Lyase	tca-14, cys-16, emp-7, ad-4, NCU08216, acu-3
Oxygenase	NCU07808, NCU04072, NCU08062, NCU09931, inl, gpi-1,
Transaminase	ser-7, NCU06189, NCU08411, NCU04292, ala, ilv-6,
Protease/metalloprotease	NCU07159, NCU05071, NCU09992, NCU07913, NCU00477, NCU06923, NCU11288, NCU07200,

Transfer/carrier protein/transporter	NCU06346, NCU03561, NCU01516, tom20, NCU04127, NCU08897, cdt-2, tim10, NCU06643, NCU04837, aac, NCU05008, NCU04537, tim9, NCU05390, tom70, NCU06804, NCU08743
Cell wall/membrane /inner membrane/inter membrane	NCU04304, NCU09175, acw-3, NCU00399, tim8, NCU04945, erg-1, NCU06702, prm-1, NCU06771, ccg-14, ccg-15, ccg-4
Domain containing protein	nfh-1, NCU05488, prm-1, NCU05800, NCU07153, nfh-2, NCU00422, NCU07536, NCU00443, NCU00225, NCU07127, NCU03515, NCU02765, NCU05542, NCU02124
Mitochondrial	NCU08898, NCU09816, NCU08794, NCU09250, pep, NCU03155, fes-1, NCU16844, tca-3, NCU03559
Kinase//dehydratase//hydrolase//aldolase//glucosidase//amylase//decarboxylase//mutase/ /transketolase//phosphodiesterase	acu-6, ace-8, pgk, NCU01550// leu-2, NCU08133, NCU04579, NCU00680// cys-18, acu-8// fba, NCU02136// NCU05974, gh1-1// gh13-1// cfp// emp-6// NCU01328// NCU09659
G-protein//tubulin //hydrophobin//calmodulin// chromatin//RNA metabolism protein//	NCU02044, rdi-1, NCU05288// tba-2, NCU04054// eas, NCU08192// NCU04120// naf-1// NCU03396, NCU06943, NCU08903//
Others	NCU08877, NCU08595, acw-9, NCU07547, gh61-9, NCU08518, NCU08720, NCU09528, ilv-2, NCU05304, acw-6, NCU01424, NCU05404, NCU09442, NCU07859, ods, ad-5, gh6-2, NCU04047, NCU06495, NCU05782, cyc-1, NCU02736, ccg-1, mig-6, NCU10020, gh61-2, NCU01849, con-6, fox-2, NCU05259, NCU09497, NCU06086, NCU03922, NCU09521, tom6, NCU08507, cse-1, NCU02263, NCU00935, NCU02623, NCU02797, gh11-1, NCU02887, fes-1, NCU05495, NCU09349, NCU09764, con-10, NCU07550, NCU09076, trp-3, pcn, NCU08030
Uncharacterized protein	NCU07177, NCU01001, NCU05080, NCU00811, NCU09395, NCU02807, NCU09929, NCU05122, NCU03873, NCU08550, NCU04148, NCU00265, NCU08657, NCU04605, NCU09693, NCU06740, NCU02944, NCU05163, NCU04620, NCU02016, NCU16329, NCU07972, NCU04169, NCU08949, NCU07439, NCU08992, NCU02784, NCU07287, NCU00766, NCU01548

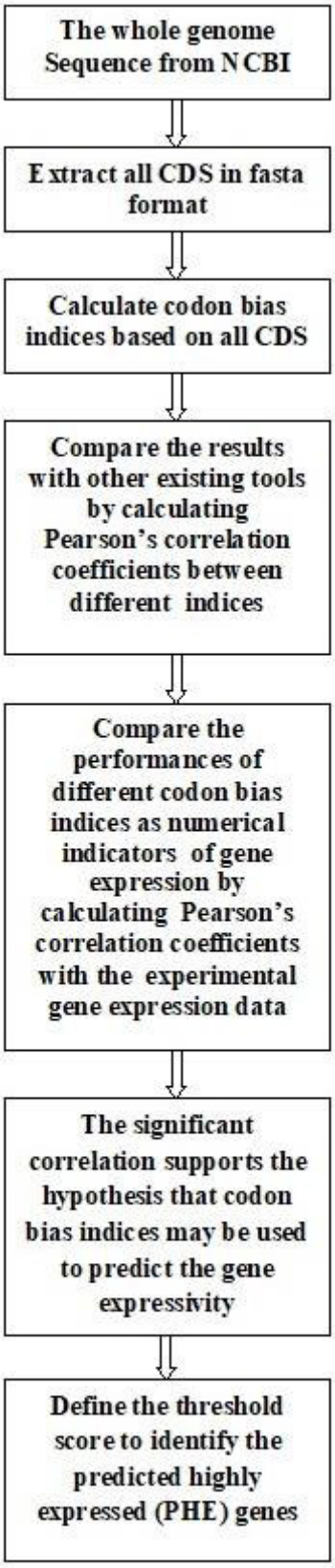

Fig. 1

Fig 1: Diagrammatic representation of the methodology.

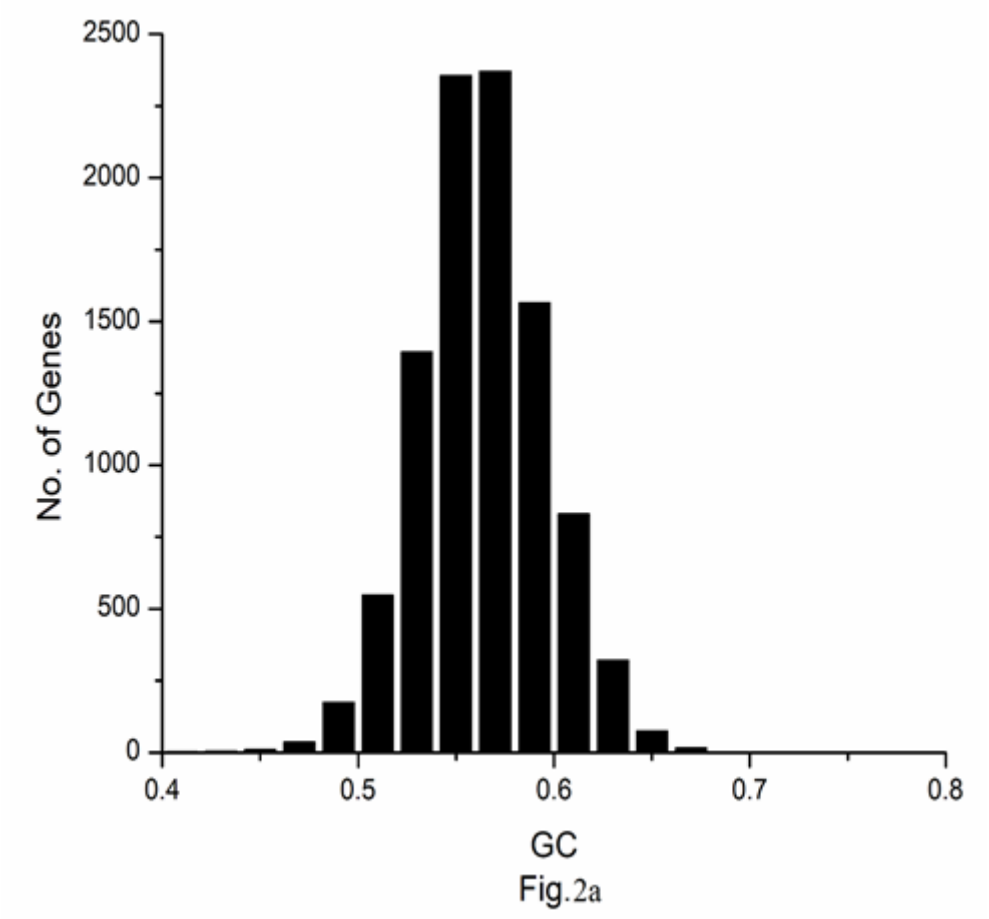

Fig 2a: Distribution of GC content in protein-coding genes of *Neurospora crassa* genome under study.

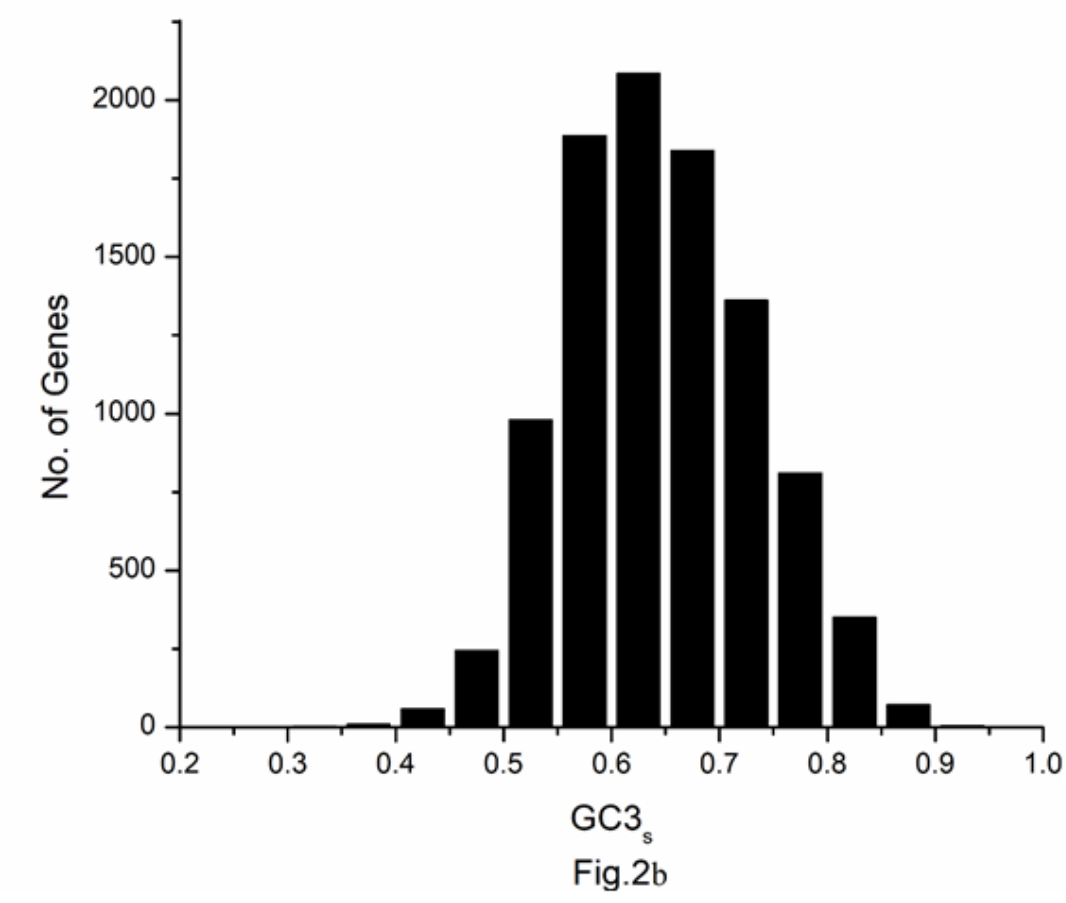

Fig 2b: Distribution of GC3s in protein-coding genes of *Neurospora crassa* genome under study.

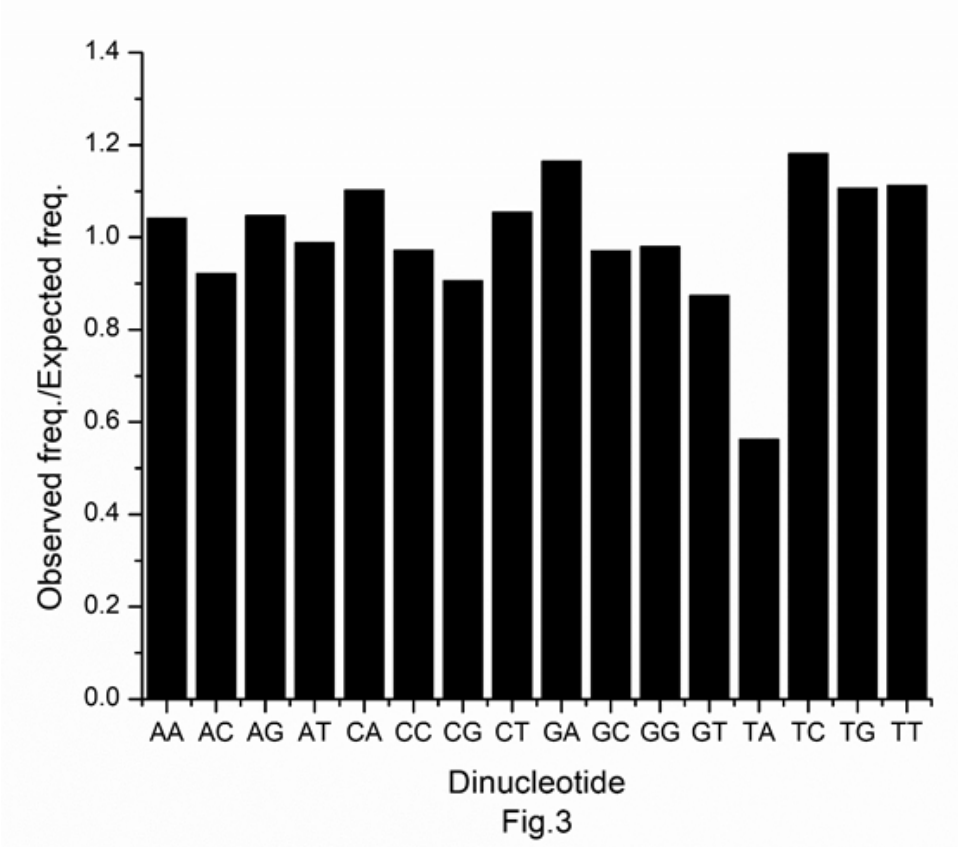

Fig 3: The ratio of observed and expected frequencies of dinucleotides in protein-coding genes of *Neurospora crassa* genome under study.

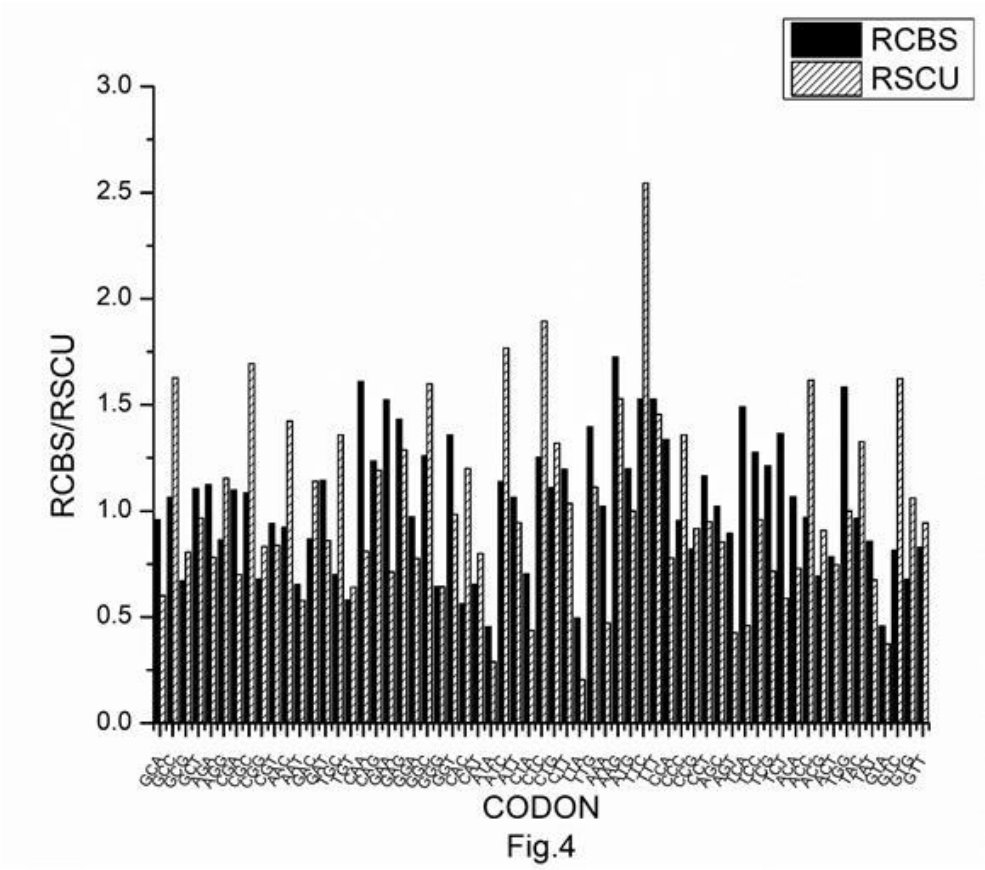

Fig 4: The RCBS and RSCU values of 61 codons of *Neurospora crassa* genes under study.

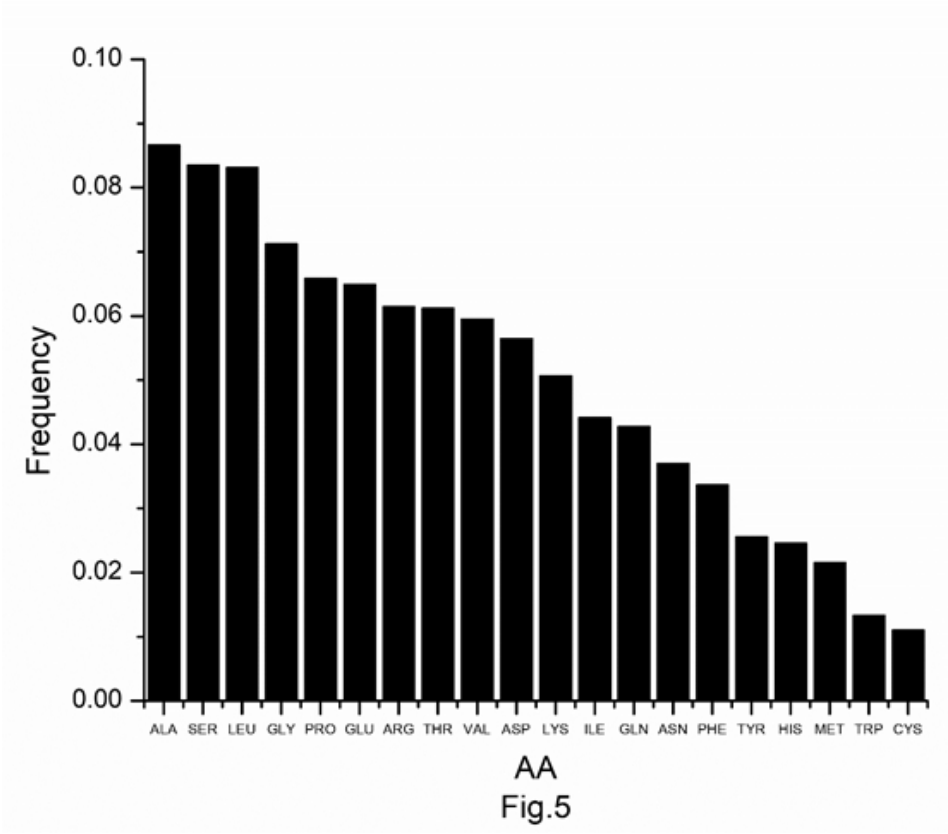

Fig 5: The frequencies of amino acids in protein-coding genes of *Neurospora crassa* genome under study.

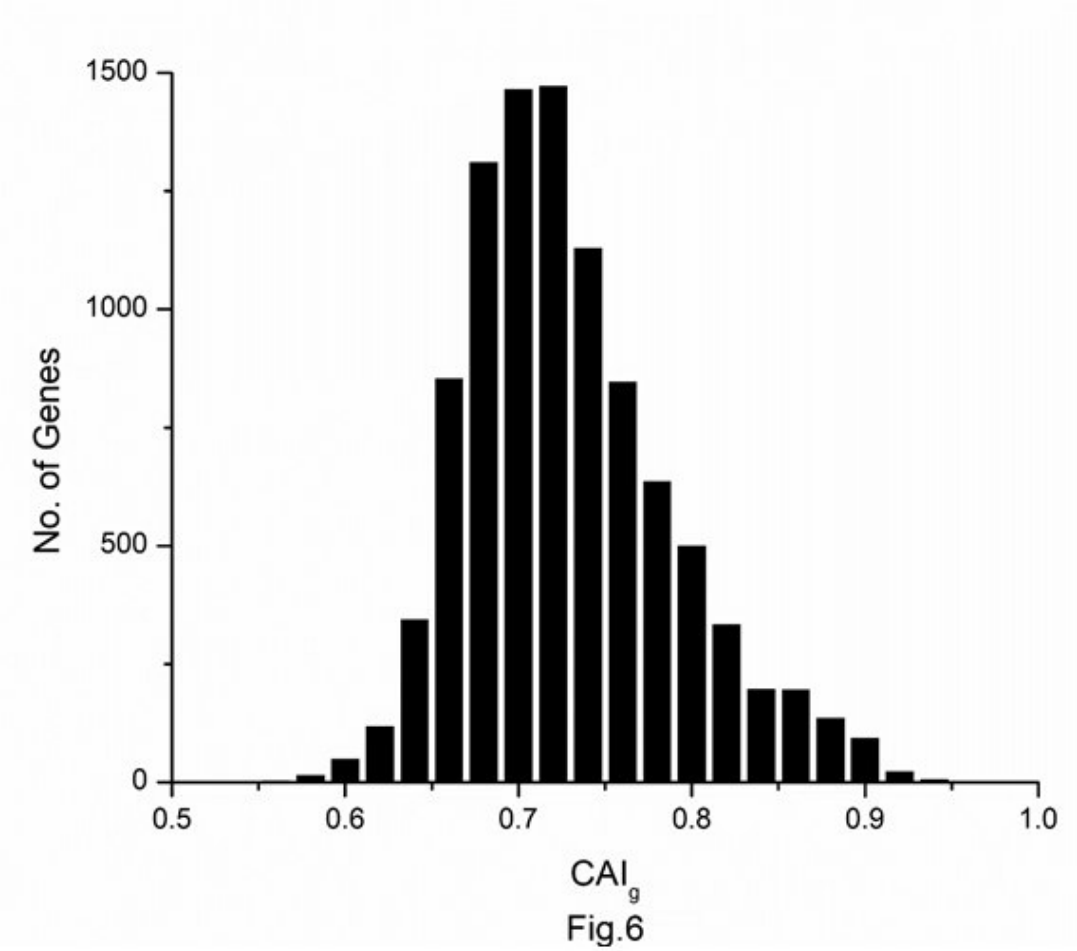

Fig 6: Distribution of CAIg of all protein-coding genes in *Neurospora crassa* genome under study.

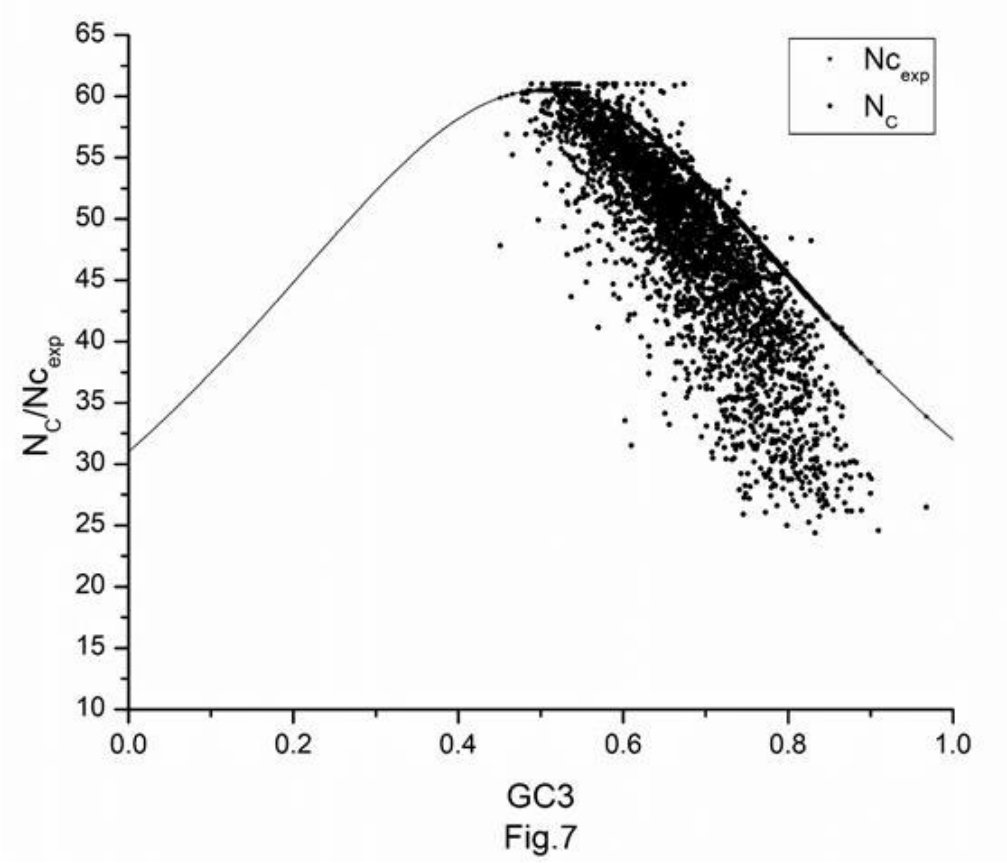

Fig 7: NC–GC3s plot for all protein coding sequences of *Neurospora crassa* genome.

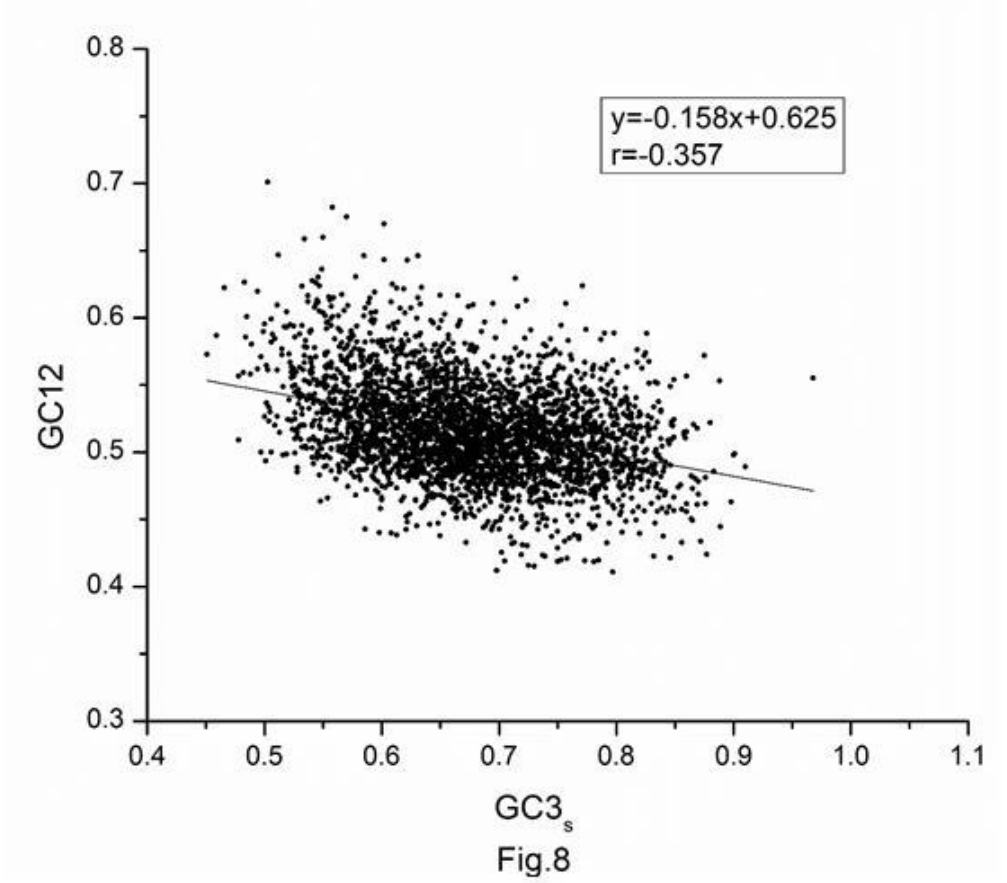

Fig 8: The neutrality plot (GC12 vs GC3s) for all protein coding sequences of *Neurospora crassa* genome.

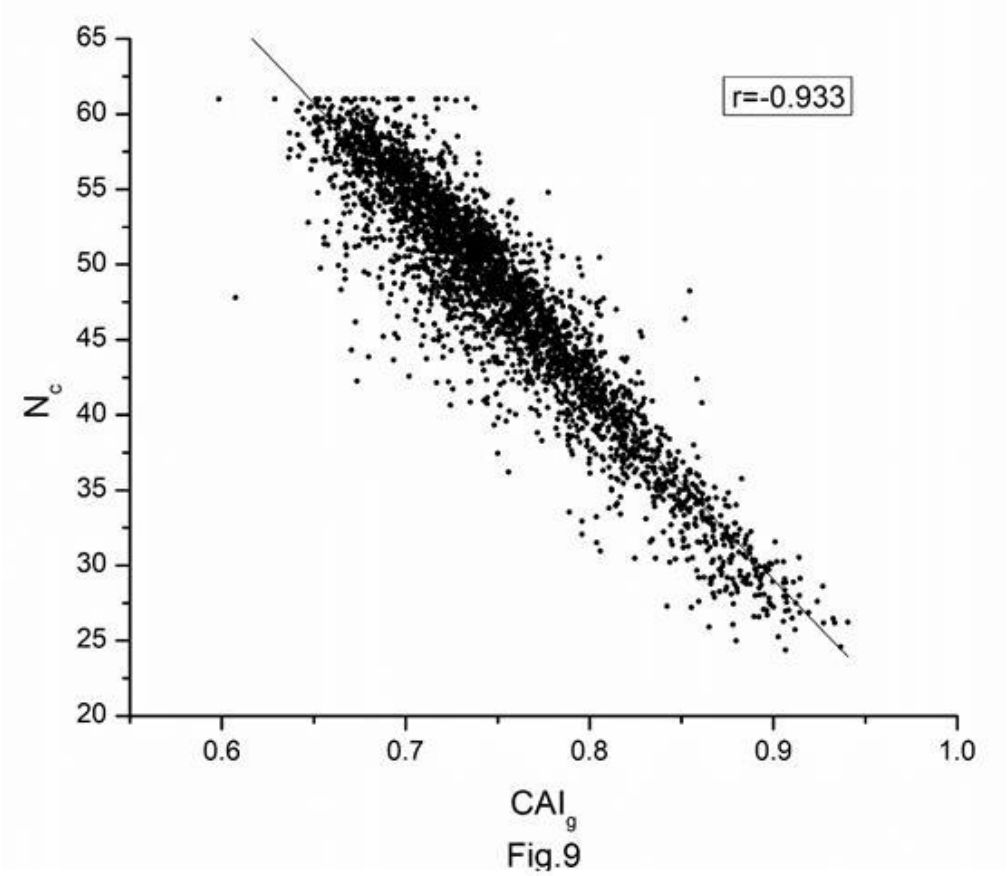

Fig 9: N_C plotted against CAIg for each protein coding-genes in *Neurospora crassa* genome under study.

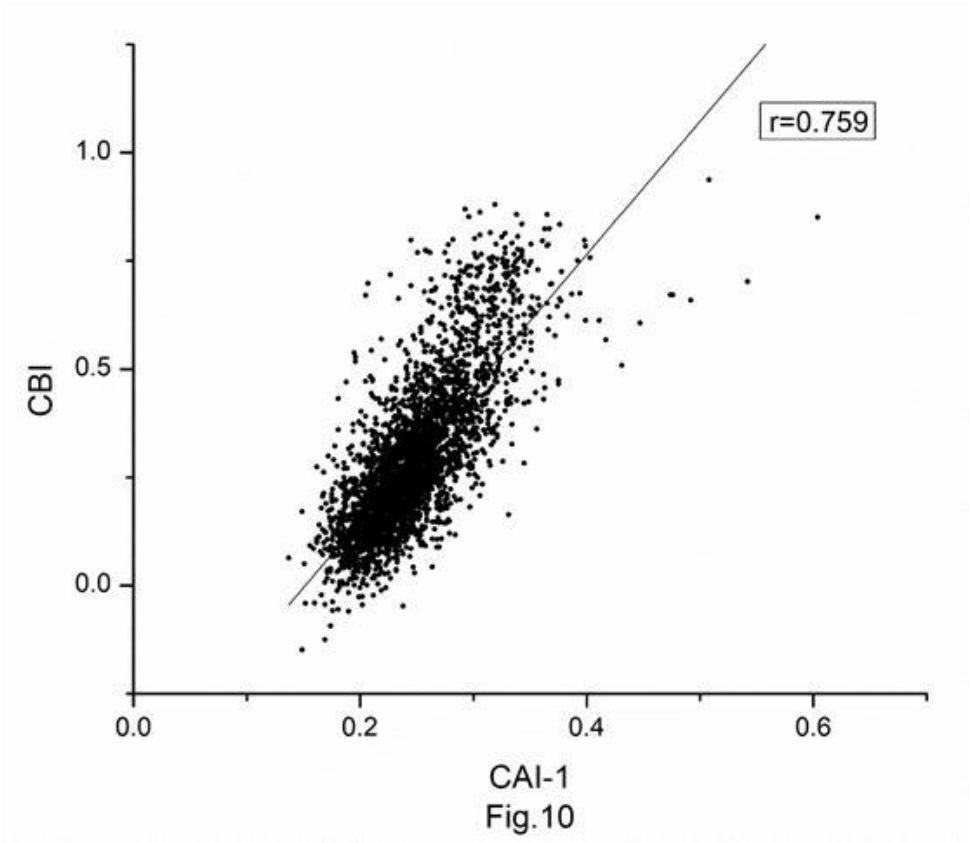

Fig 10: CBI plotted against CAI-1 for each protein-coding genes in *Neurospora crassa* genome under study. CAI-1 is the codon adaptation index calculated in reference to set of highly expressed genes of *e.coli*.

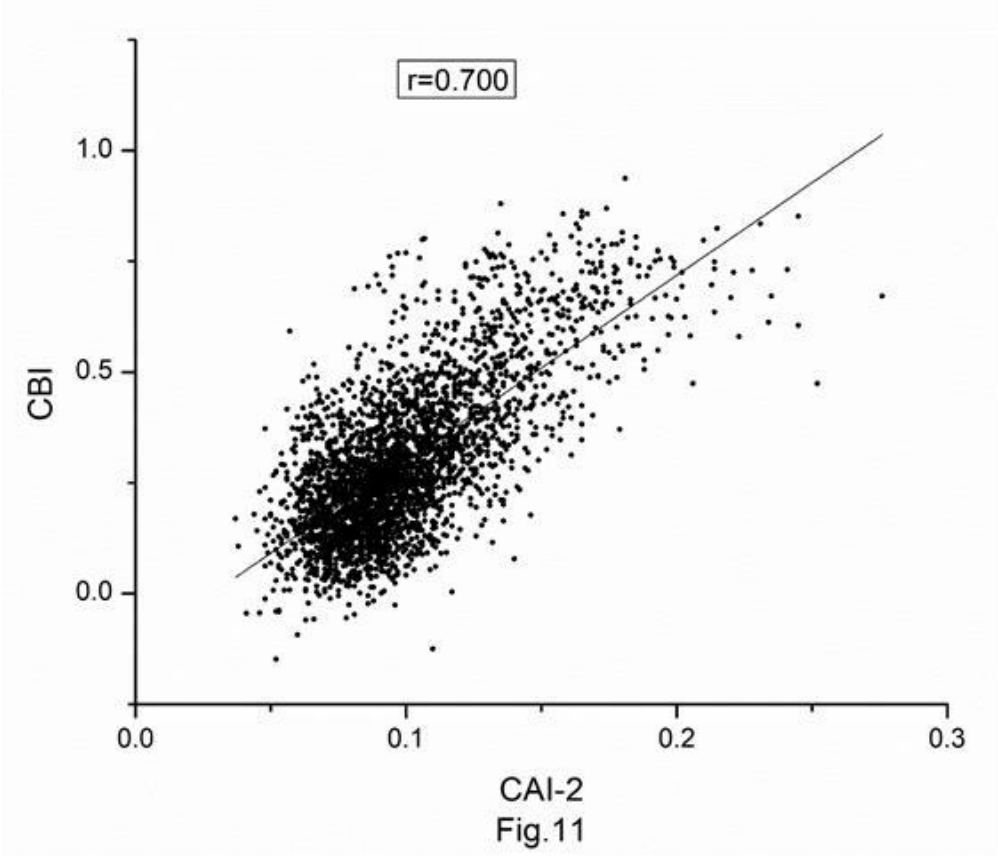

Fig 11: CBI plotted against CAI-2 for each protein-coding genes in *Neurospora crassa* genome under study. CAI-2 is the codon adaptation index calculated in reference to set of highly expressed genes of *S. cerevicsiae*.

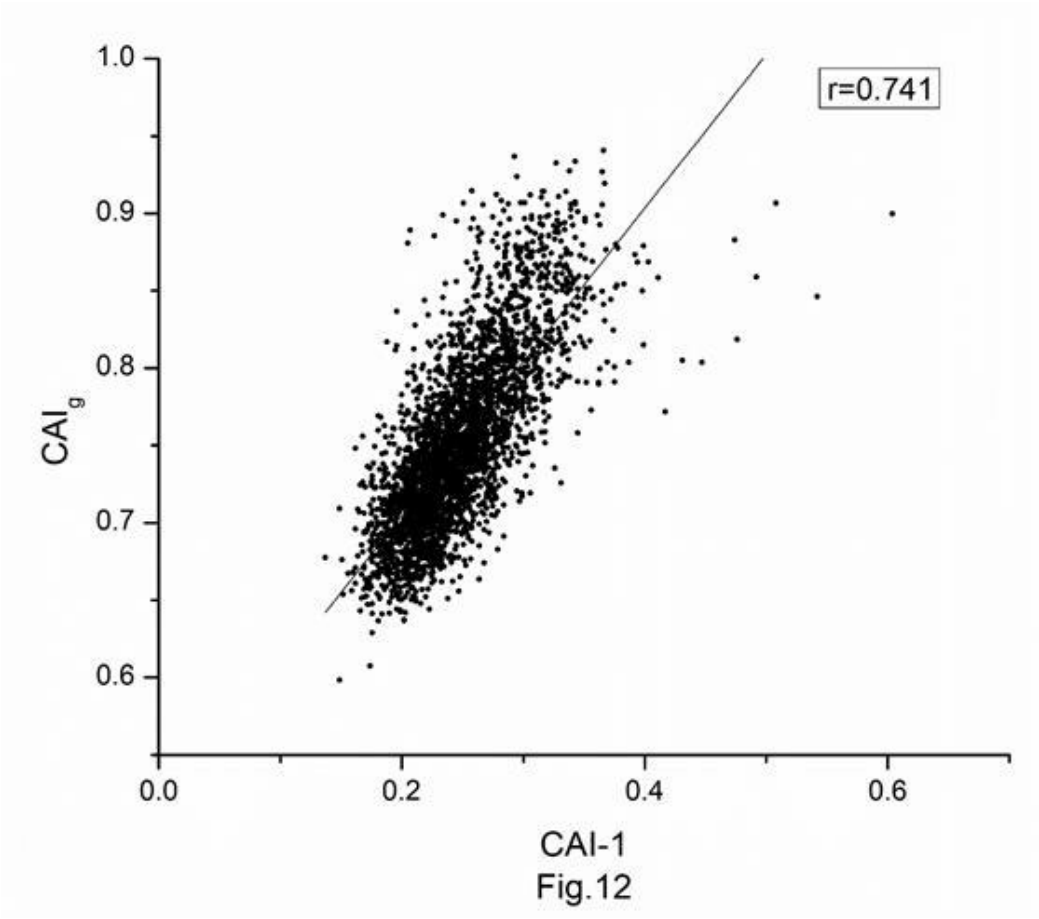

Fig 12: CAIg plotted against CAI-1 for each protein-coding genes in *Neurospora crassa* genome under study. CAI-1 is the codon adaptation index calculated in reference to set of highly expressed genes of *e.coli*.

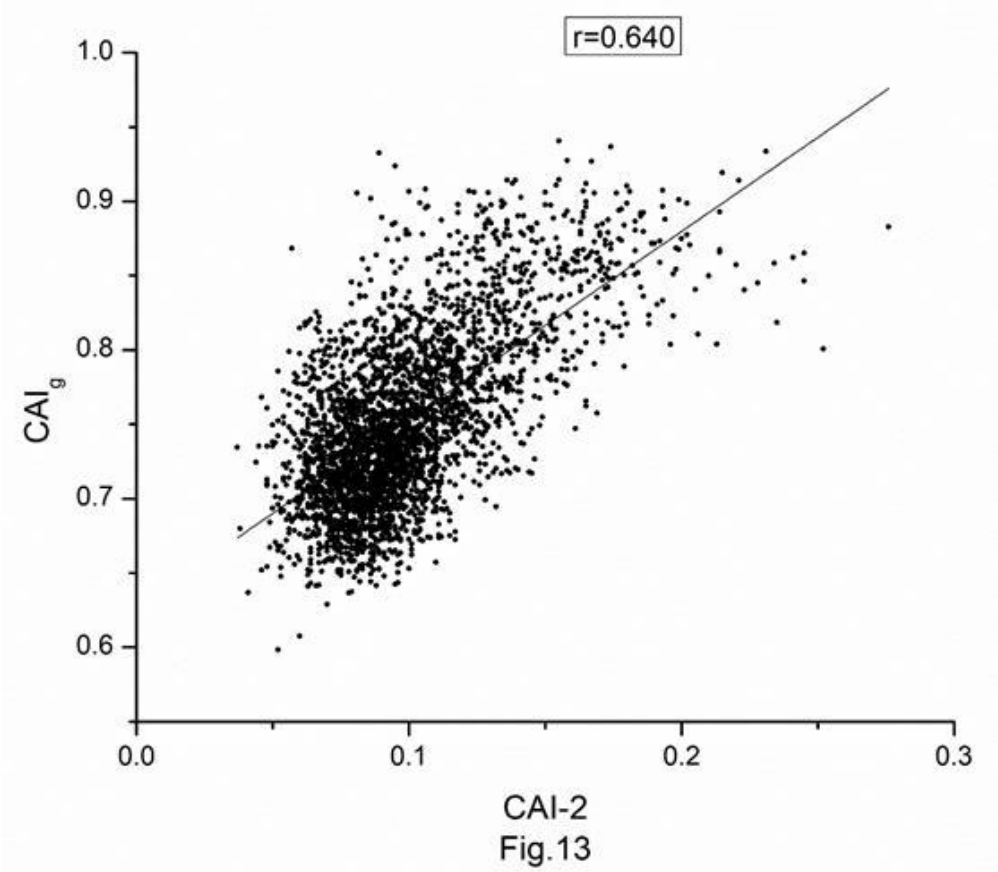

Fig 13: CAIg plotted against CAI-2 for each protein-coding genes in *Neurospora crassa* genome under study. CAI-2 is the codon adaptation index calculated in reference to set of highly expressed genes of *S. cerevicsiae*.

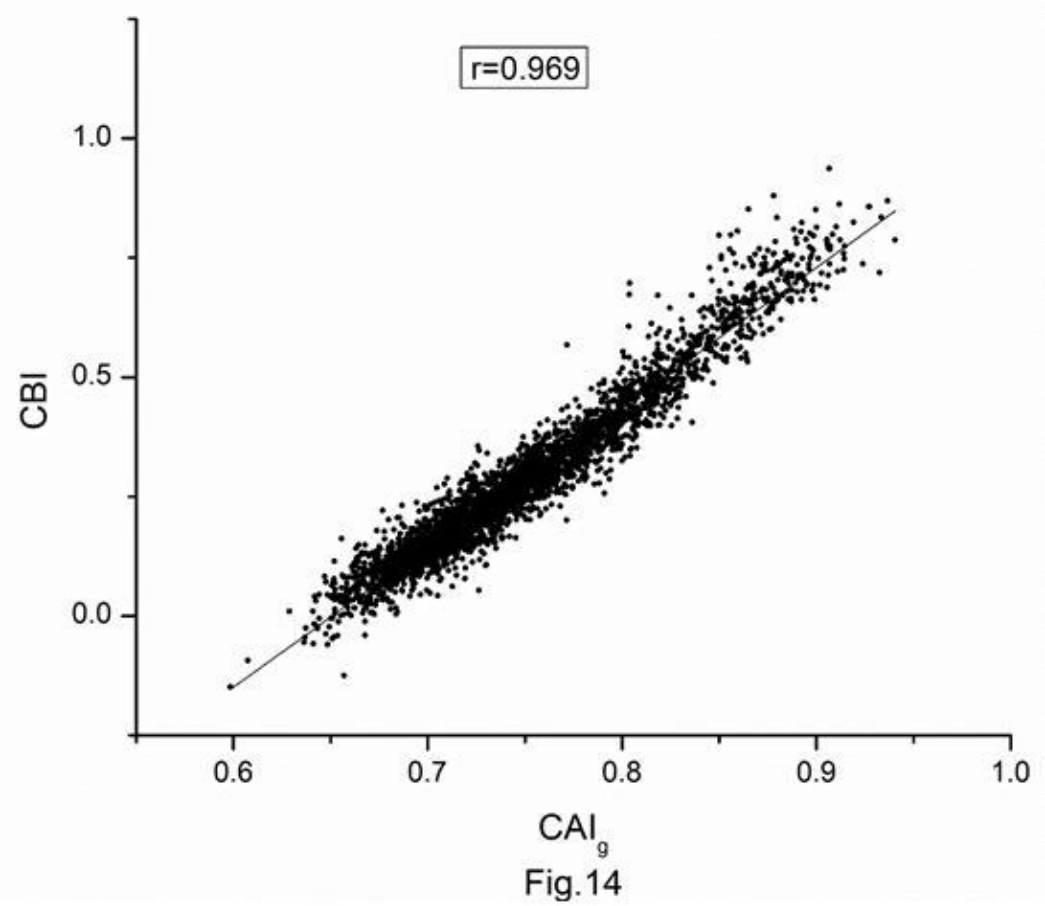

Fig 14: CBI plotted against CAIg for each protein coding-genes in *Neurospora crassa* genome under study.

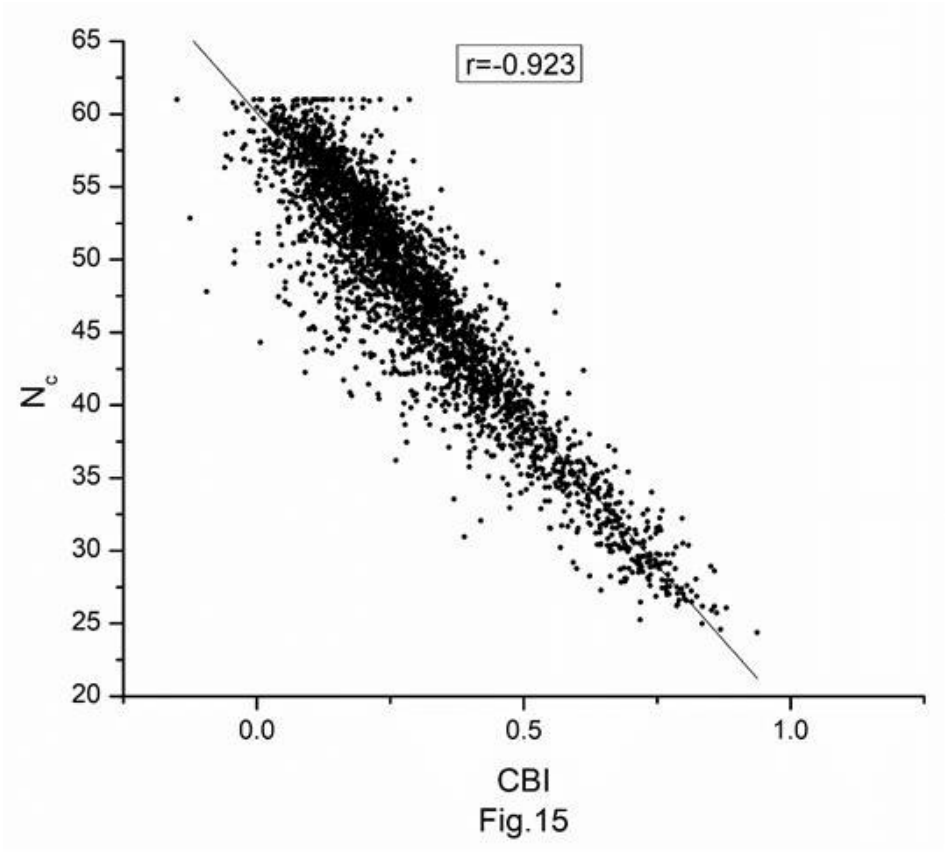

Fig 15: N_C plotted against CBI for each protein coding-genes in *Neurospora crassa* genome under study.

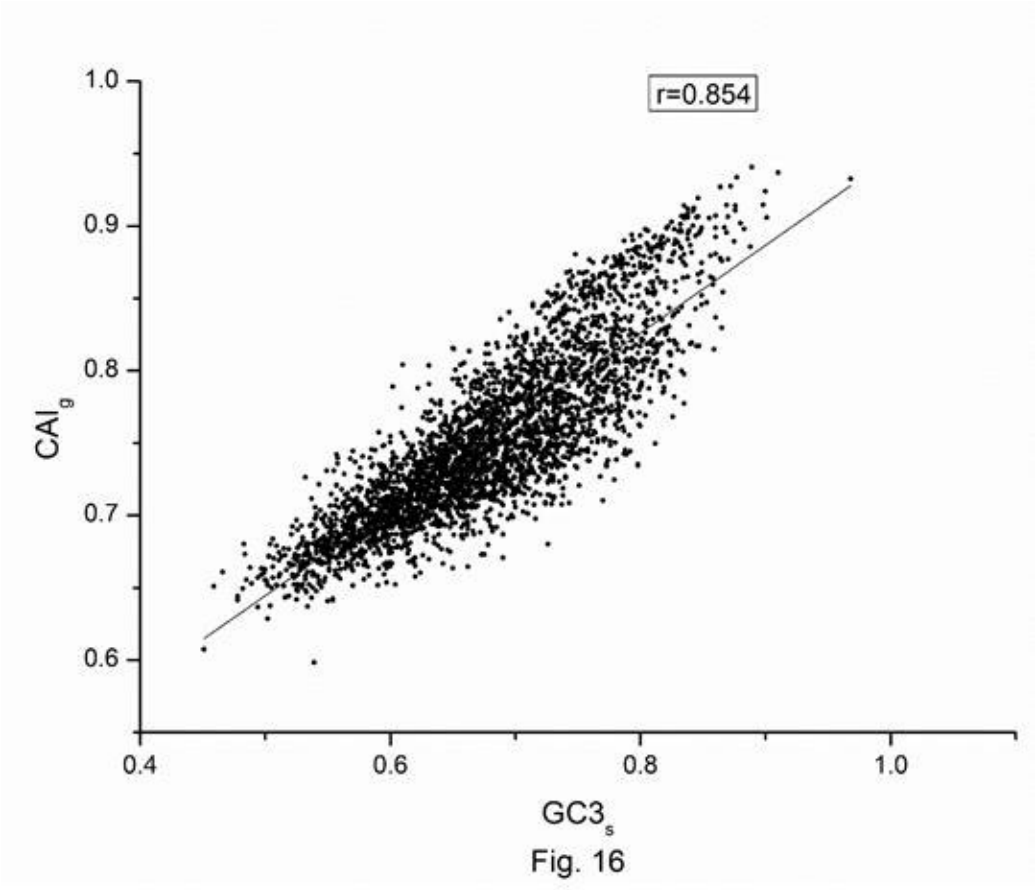

Fig 16: CAIg plotted against GC3s for each protein-coding genes in *Neurospora crassa* genome under study.

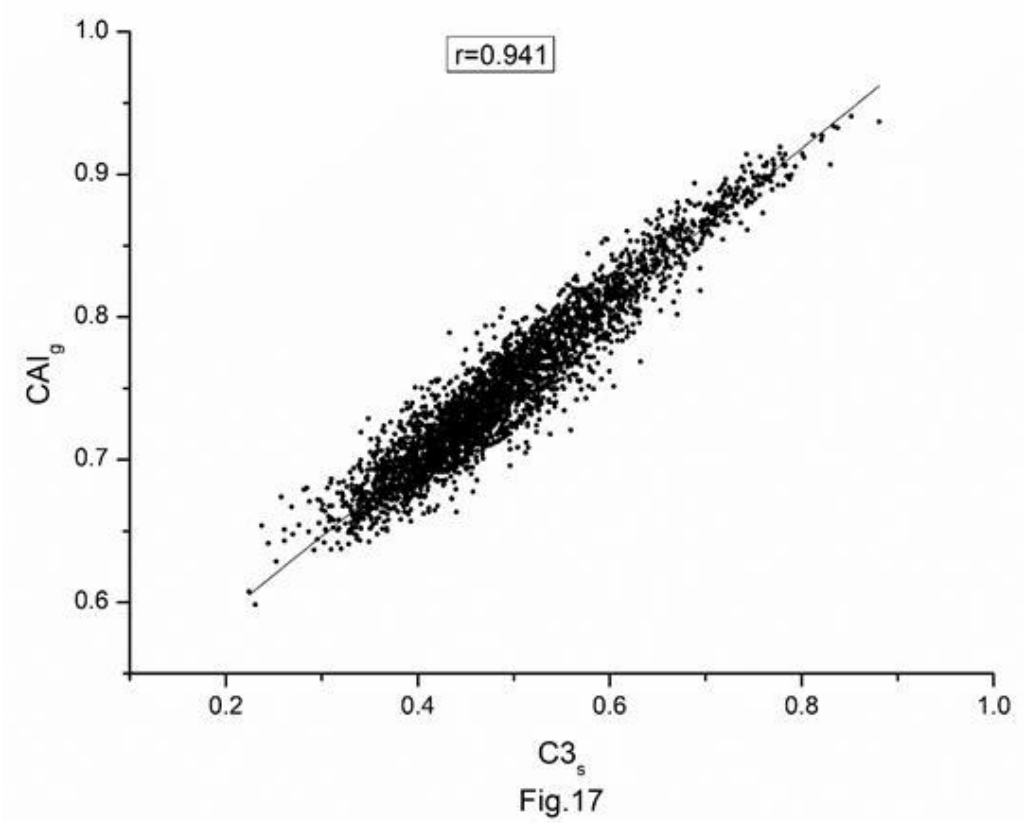

Fig 17: CAIg plotted against C3s for each protein-coding genes in *Neurospora crassa* genome under study.

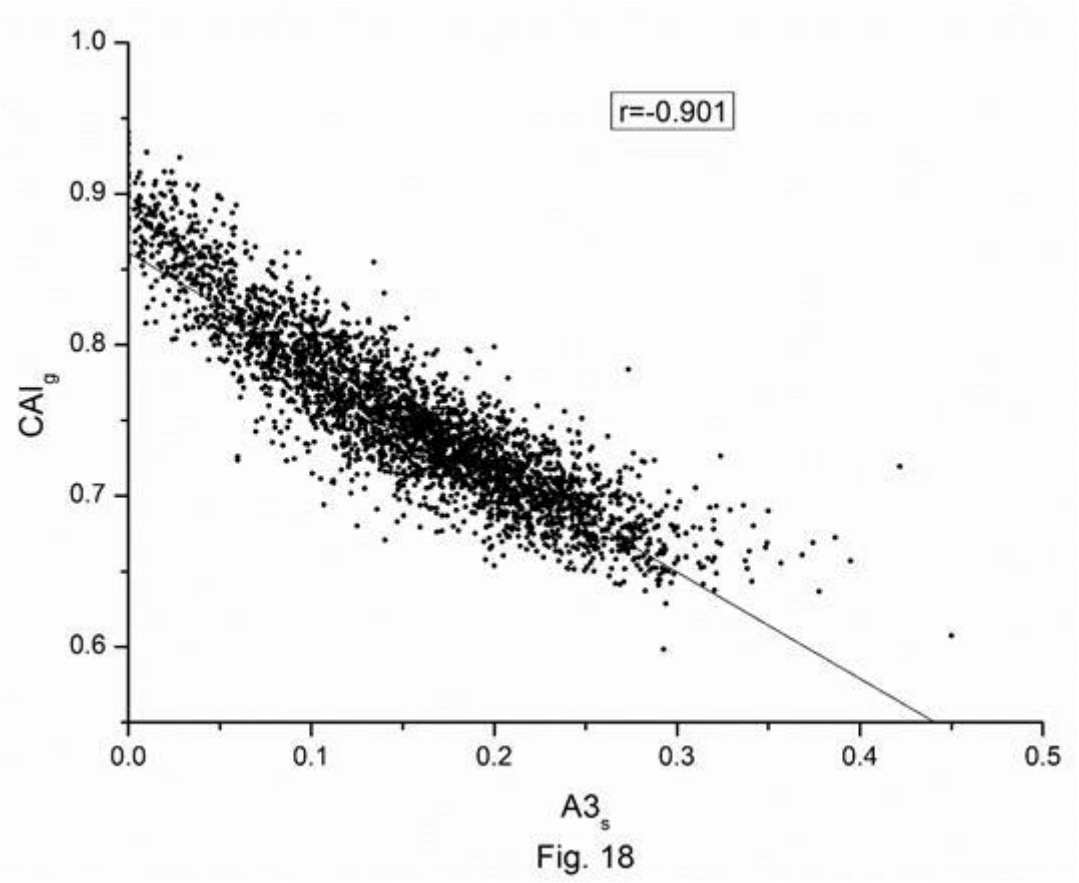

Fig 18: CAIg plotted against A3s for each protein-coding genes in *Neurospora crassa* genome under study.

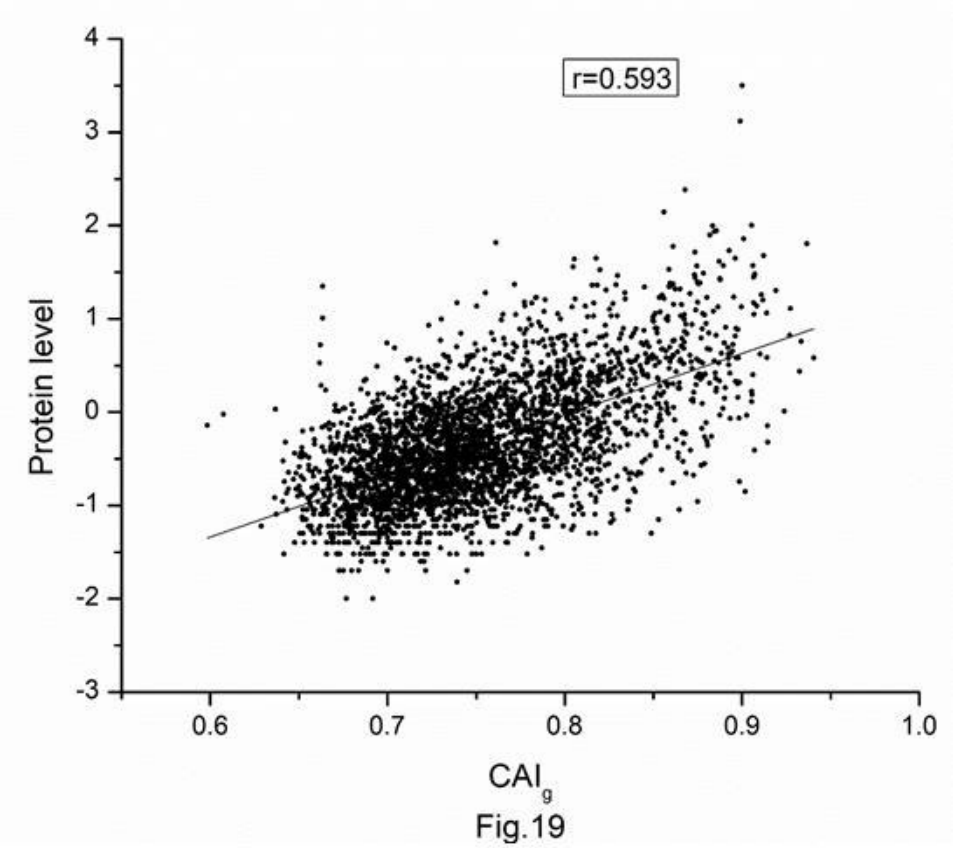

Fig 19: Protein levels(emPAI) plotted against CAIg for a set of 3200 identified genes in *Neurospora crassa* genome.

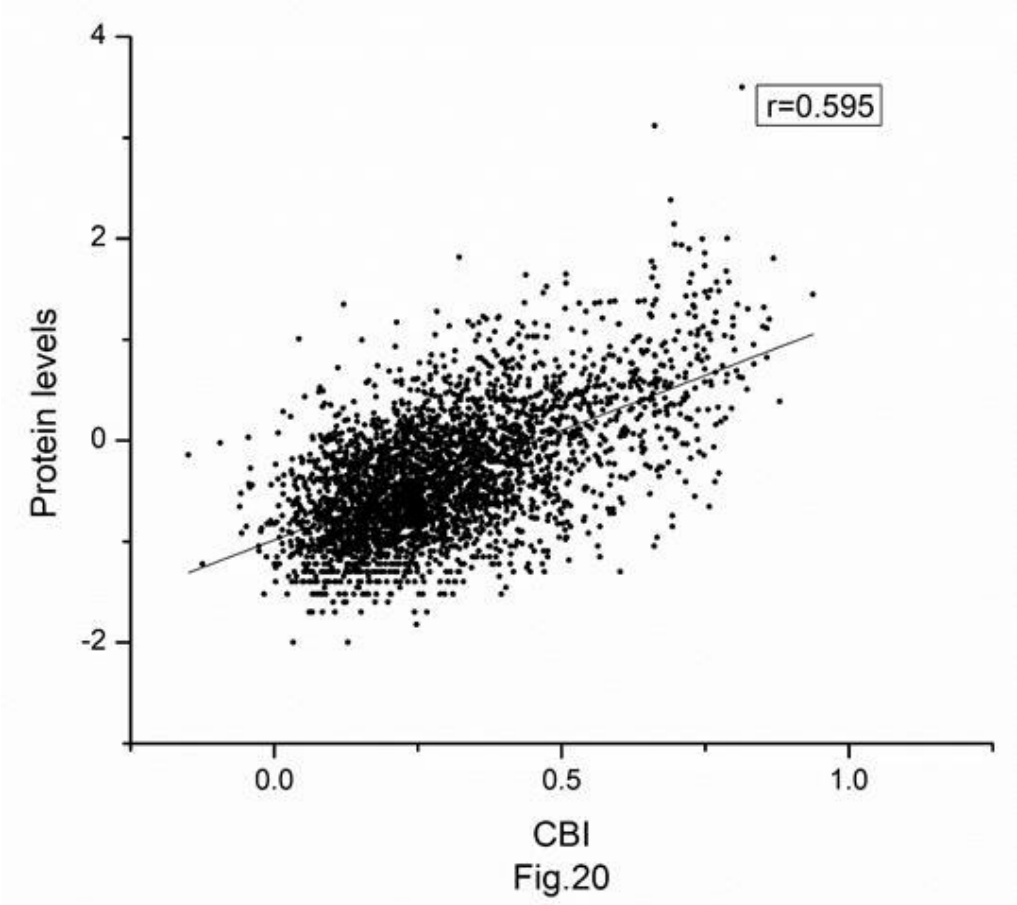

Fig 20: Protein levels(emPAI) plotted against CBI for a set of 3200 identified genes in *Neurospora crassa* genome.

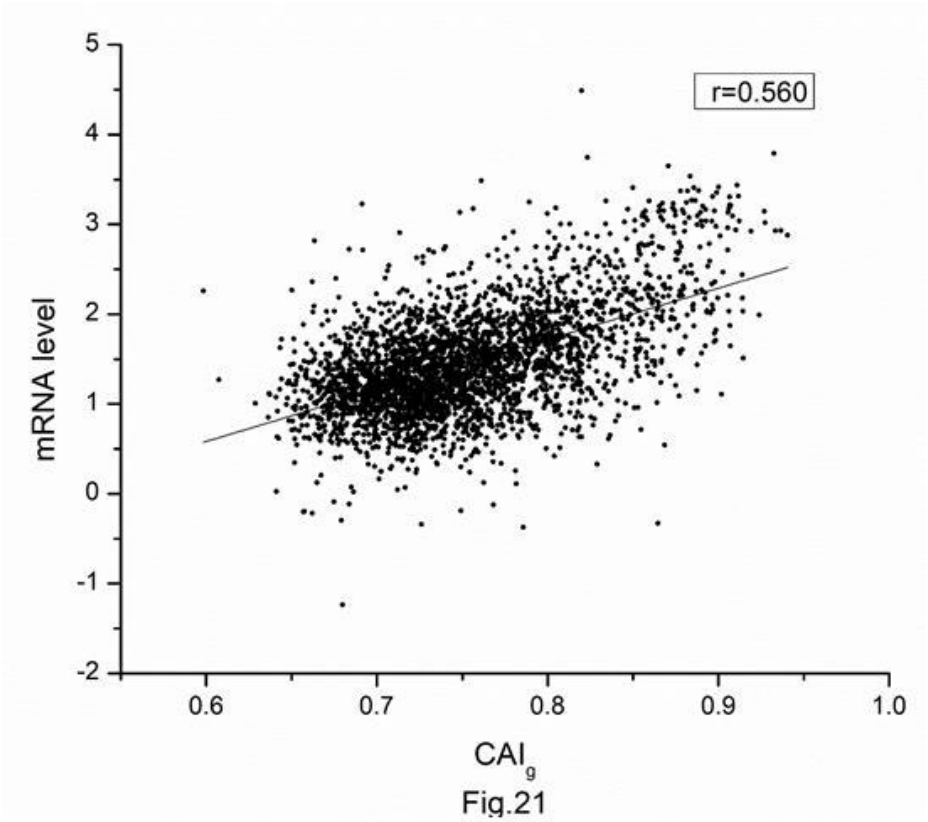

Fig 21: mRNA levels plotted against CAIg for a set of 3200 identified genes in *Neurospora crassa* genome.

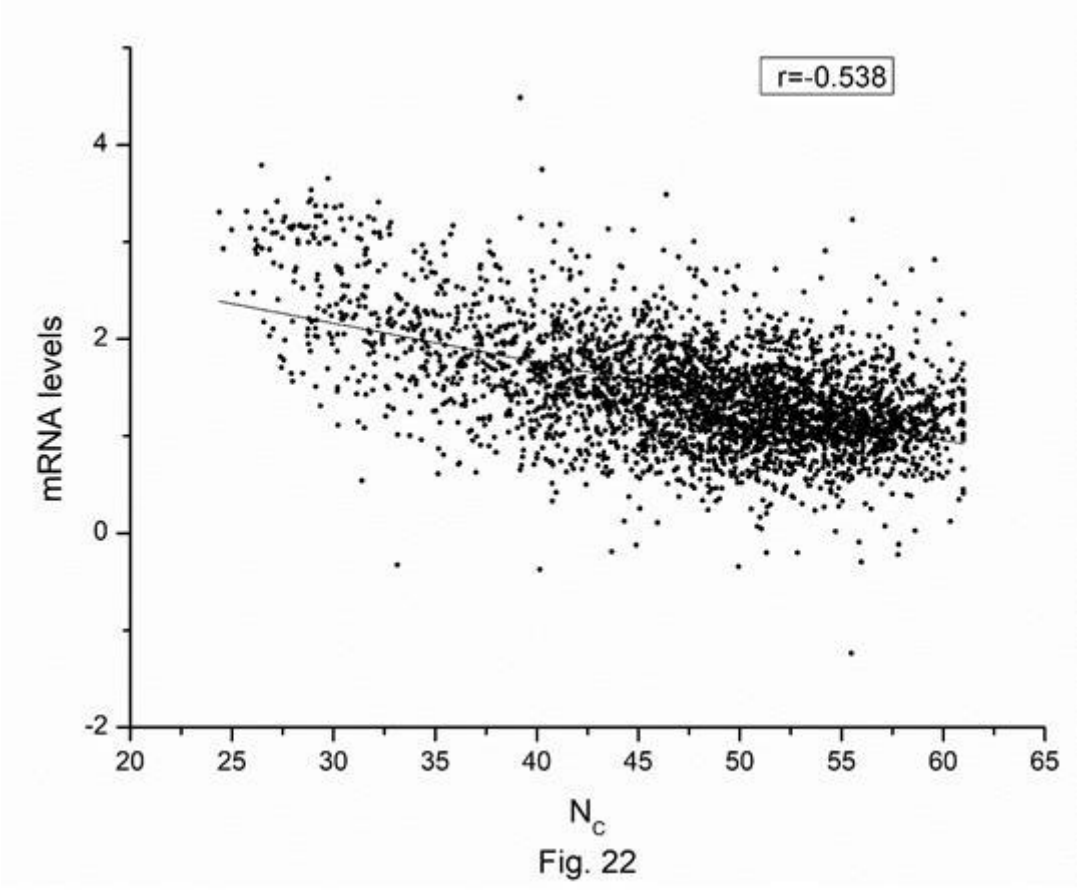

Fig 22: mRNA levels plotted against N_C for a set of 3200 identified genes in *Neurospora crassa* genome.

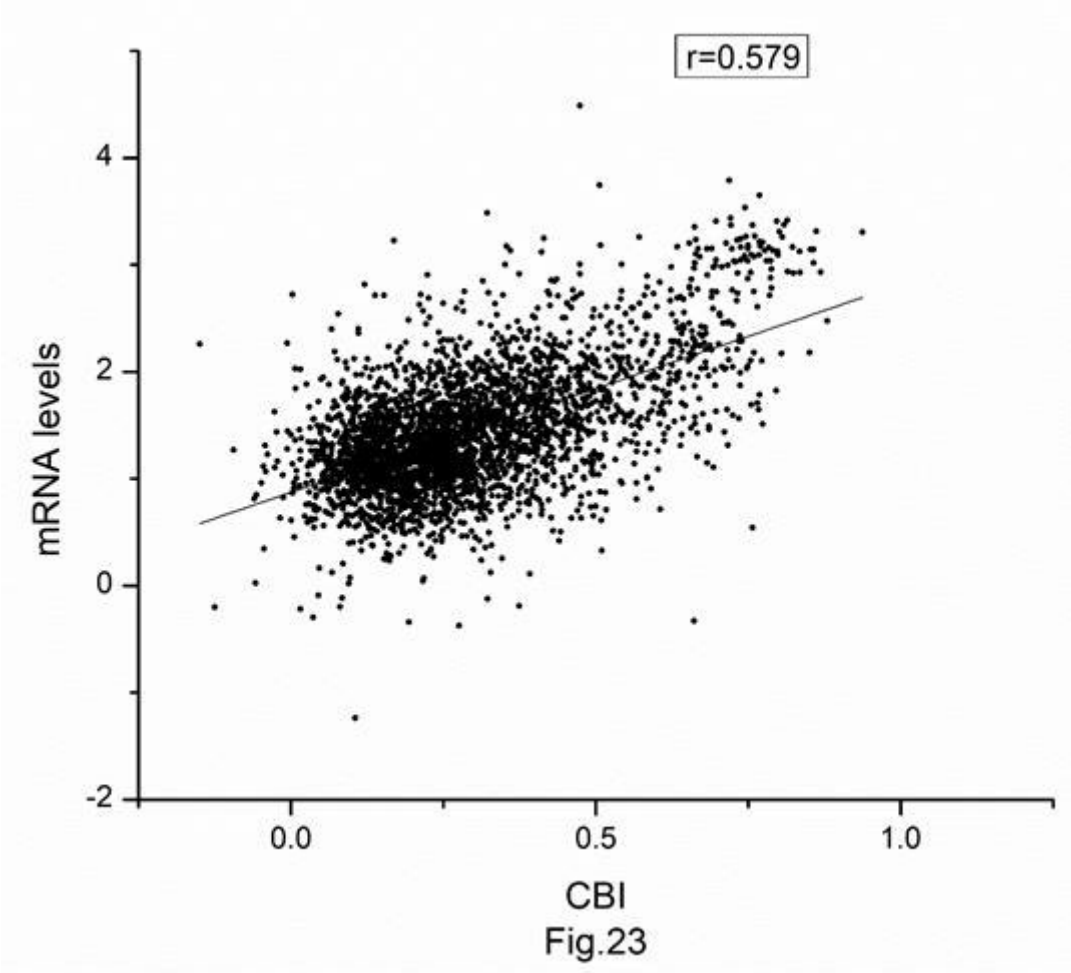

Fig 23: mRNA levels plotted against CBI for a set of 3200 identified genes in *Neurospora crassa* genome.

4. CONCLUSION

In summary, the present study describes the influence of synonymous codons on gene expression in *Neurospora crassa* and supports the hypothesis that the codon usage pattern is an important factor in translational dynamics regulating gene expressivity. The protein production is controlled by translation initiation and elongation efficiency, and the codon usage and tRNA anticodons coevolve to adapt to each other, resulting in increased production of correctly translated proteins. CAI as an indicator for translation elongation efficiency is inadequate and can lead to serious bias because it does not account for the mutation bias in characterizing codon adaptation. Further studies are essential for understanding the joint effect of mutation and selection on codon usage and to design improved computational tools for characterizing codon usage and codon-anticodon adaptation yielding an adequate index for quantifying gene expressivity.

5. FUNDING ACKNOWLEDGEMENT

The author acknowledges the Science and Engineering Research Board, DST, Govt. of India for the financial support under fixed grant scheme MATRICS[File No: MTR/2019/000274].

6. CONFLICT OF INTEREST

Conflict of interest declared none.

7. REFERENCES

1. Davis RH. Neurospora: contributions of a Model Organism. New York: Oxford University Press; 2000.
2. Davis RH, Perkins DD. Timeline: Neurospora: a model of model microbes. Nat Rev Genet. 2002;3(5):397-403. doi: 10.1038/nrg797, PMID 11988765.
3. Quax TEF, Claassens NJ, Söll D, Oost JVD. Codon Bias as a Means to Fine-Tune Gene Expression. Mol Cell. 2015;59(2): 149–161 doi: 10.1016/j.molcel.2015.05.035, PMID: 26186290
4. Brule CE, Grayhack EJ. Synonymous Codons: Choose Wisely for Expression. *Trends Genet*. 2017;33(4):283-297. doi:10.1016/j.tig.2017.02.001
5. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 1980;8(1):r49-62. doi: 10.1093/nar/8.1.197-c, PMID 6986610.
6. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 1985;2(1):13-34. doi: 10.1093/oxfordjournals.molbev.a040335, PMID 3916708.
7. Salim HMW, Cavalcanti ARO. Factors influencing codon usage bias in genomes. J Braz Chem Soc. 2008;19(2):257-62. doi: 10.1590/S0103-50532008000200008.

8. Osawa S, Ohama T, Yamao F, Muto A, Jukes TH, Ozeki H, Umesono K. Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. Proc Natl Acad Sci U S A. 1988;85(4):1124-8. doi: 10.1073/pnas.85.4.1124, PMID 2448791.
9. Sueoka N. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci U S A. 1988;85(8):2653-7. doi: 10.1073/pnas.85.8.2653, PMID 3357886.
10. Akashi H. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics. 1994;136(3):927-35. doi: 10.1093/genetics/136.3.927, PMID 8005445.
11. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol. 1986;24(1-2):28-38. doi: 10.1007/BF02099948, PMID 3104616.
12. Sharp PM, Stenico M, Peden JF, Lloyd AT. Codon usage – mutational bias, translational selection, or both. Biochem Soc Trans. 1993;21(4):835-41. doi: 10.1042/bst0210835, PMID 8132077.
13. Xie T, Ding D. The relationship between synonymous codon usage and protein structure. FEBS Lett. 1998;434(1-2):93-6. doi: 10.1016/s0014-5793(98)00955-7, PMID 9738458.
14. Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci U S A. 1999;96(8):4482-7. doi: 10.1073/pnas.96.8.4482, PMID 10200288.
15. Lobry JR, Gautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. 1994;22(15):3174-80. doi: 10.1093/nar/22.15.3174, PMID 8065933.
16. Ikemura T. 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. 1981;151(3):389-409. doi: 10.1016/0022-2836(81)90003-6, PMID 6175758.
17. Sharp PM, Li WH. The codon adaptation index -a measure of directional synonymous codon usage bias and its potential applications. Nucleic Acids Res. 1987;15(3):1281-95. doi: 10.1093/nar/15.3.1281, PMID 3547335.
18. Wright F. The 'effective number of codons' used in a gene The 'effective number of codons' used in a gene The 'effective number of codons' used in a gene. Gene. 1990;87(1):23-9. doi: 10.1016/0378-1119(90)90491-9, PMID 2110097.
19. Bennetzen JL, Hall BD. Codon selection in yeast. J Biol Chem. 1982;257(6):3026-31. doi: 10.1016/S0021-9258(19)81068-2, PMID 7037777.
20. Carbone A, Zinovyev A, Képès F. Codon adaptation index as a measure of dominating codon bias. Bioinformatics. 2003;19(16):2005-15. doi: 10.1093/bioinformatics/btg272, PMID 14594704.
21. Supek F, Vlahovicek K. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics. 2005;6:182. doi: 10.1186/1471-2105-6-182, PMID 16029499.
22. Roymondal U, Das S, Sahoo S. Predicting gene expression level from relative codon usage bias:an application to *Escherichia coli* genome. DNA Res. 2009;16(1):13-30. doi: 10.1093/dnares/dsn029, PMID 19131380.
23. Das S, Roymondal U, Sahoo S. Analyzing gene expression from relative codon usage bias in *Yeast* genome: a statistical significance and biological relevance. Gene. 2009;443(1-2):121-31. doi: 10.1016/j.gene.2009.04.022, PMID 19410638.
24. Das S, Roymondal U, Chottopadhyay B, Sahoo S. Gene expression profile of the *cynobacterium Synechocystis* genome. Gene. 2012;497(2):344-52. doi: 10.1016/j.gene.2012.01.023, PMID 22310391.
25. Fox JM, Erill I. Relative codon adaptation: A generic codon bias index for prediction of gene expression. DNA Res. 2010;17(3):185-96. doi: 10.1093/dnares/dsq012, PMID 20453079.
26. Khandia R, Singhal S, Kumar U, Ansari A, Tiwari R, Dhama K, Das J, Munjal A, Singh RK. Analysis of Nipah virus codon usage and adaptation to Hosts. Front Microbiol. 2019;10:886. doi: 10.3389/fmicb.2019.00886, PMID 31156564.
27. Lytras S, Hughes J. Synonymous dinucleotide usage: A codon-aware metric for quantifying dinucleotide representation in viruses. Viruses. 2020;12(4):462. doi: 10.3390/v12040462, PMID 32325924.
28. Yang X, Luo X, Cai X. Analysis of codon usage pattern in Taenia saginata based on a transcriptome dataset. Parasit Vectors. 2014;7:527. doi: 10.1186/s13071-014-0527-1, PMID 25440955.
29. Jia X, Liu S, Zheng H, Li B, Qi Q, Wei L, Zhao T, He J, Sun J. Non-uniqueness of factors constraint on the codon usage in Bombyx mori. BMC Genomics. 2015;16:356. doi: 10.1186/s12864-015-1596-z, PMID 25943559.
30. Zhao Y, Zheng H, Xu A, Yan D, Jiang Z, Qi Q, Sun J. Analysis of codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus (NPV) and its relation to evolution. BMC Genomics. 2016;17:677. doi: 10.1186/s12864-016-3021-7, PMID 27558469.
31. Sémon M, Mouchiroud D, Duret L. LRelationship between gene expression and GC-content in mammals: statistical significance and biological relevance. Hum Mol Genet. 2005;14(3):421-7. doi: 10.1093/hmg/ddi038, PMID 15590696.
32. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. High guanine and cytosine content increases mRNA levels in mammalian cells. PLOS Biol. 2006;4(6):e180. doi: 10.1371/journal.pbio.0040180, PMID 16700628.
33. Arhondakis S, Clay O, Bernardi GS. GC level and expression of human coding sequences. Biochem Biophys Res Commun. 2008;367(3):542-5. doi: 10.1016/j.bbrc.2007.12.155, PMID 18177737.
34. Fryxell KJ, Zuckerkandl E. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol Biol Evol. 2000;17(9):1371-83. doi: 10.1093/oxfordjournals.molbev.a026420, PMID 10958853.
35. Zhao F, Yu CH, Liu Y. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. Nucleic Acids Res. 2017;45(14):8484-92. doi: 10.1093/nar/gkx501, PMID 28582582.
36. Jansen R, Bussemaker HJ, Gerstein M. Revisiting the codon adaptation index from a whole-genome

perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res*. 2003;31(8):2242-2251. doi:10.1093/nar/gkg306

37. Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J, Chen S, Liu Y. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proc Natl Acad Sci U S A. 2016;113(41):E6117-25. doi: 10.1073/pnas.1606724113, PMID 27671647.
38. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics 2005; **4**(9):1265–1272. doi: 10.1074/mcp.M500061-MCP200. Epub 2005 Jun 14. PMID: 15958392.
39. Zhou Z, Dang Y, Zhou M, Yuan H, Liu Y. Codon usage biases co-evolve with transcription termination machinery to suppress premature cleavage and polyadenylation. eLife. 2018;7:e33569. doi: 10.7554/eLife.33569, PMID 29547124.
40. Lyu X, Liu Y. Non-optimal codon usage is critical for protein structure and function of the master general amino acid control regulator CPC-1. mBio. 2020;11(5):e02605-20. doi: 10.1128/mBio.02605-20, PMID 33051373.