INNSPUB

International Journal of Biosciences | IJB |
ISSN: 2220-6655 (Print) 2222-5234 (Online)
http://www.innspub.net
Vol. 13, No. 4, p. 179-192, 2021

RESEARCH PAPER OPEN ACCESS

Analysis of codon usage pattern and gene expression in *Aspergillus fumigatus*

Satyabrata Sahoo*

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

Key words: Codon usage, Codon bias, Gene expression, *Aspergillus fumigatus*, PHE genes

http://dx.doi.org/10.12692/ijb/19.4.179-192 Article published on October 30, 2021

Abstract

The codon usage pattern and its impact on gene expression has been investigated in protein coding genes of *Aspergillus fumigatus*. Multivariate statistical analysis has been used to analyze compositional properties and codon usage pattern of genes. The results suggest that the selection of translational efficiency due to natural selection is the major factor shaping the codon usage in *Aspergillus fumigatus*. Using different codon usage indices as numerical estimators of gene expression level, a critical analysis has been performed to predict highly expressed (PHE) genes in *Aspergillus fumigatus*. The codon usage indices correlate well with the gene expression data set stem from the transcriptional responses of the micro-organisms, suggesting that codon usage is an important determinant of gene expression. We found a systematic strong correlation between Nc (effective number of codons) and different expression-measures. Our study highlights the relationship between gene expression and compositional signature in relation to the codon usage bias in *Aspergillus fumigatus*.

* **Corresponding Author:** Satyabrata Sahoo ✉ dr_s_sahoo@yahoo.com

Introduction

Aspergillus fumigatus, the most prevalent airborne pathogenic fungus, is reported to be a measure cause of invasive fungal infections in immune-compromised hosts. Several studies have demonstrated that the pathogenesis of this fungal species is multifactorial due to a combination of its biological characteristics and the immune status of patients (Abad *et al.*, 2010; Beauvais and Latgé, 2001; Osherov, 2007; Tekaia and Latgé, 2005). The availability of the whole-genome sequence of this fungal species since 2005 had developed it as a model organism for contributing to the fundamental understanding of modern genetics and molecular biology. Undoubtedly, any useful insight in understanding the expression of functional proteins of *Aspergillus fumigatus* will contribute to the development of modern biotechnology. The present study is focused on the compositional signature of gene sequence and its influence on gene expressivity. Codon usage bias (CUB), the preferential use of some types of codons over others encoding the same amino acid during protein synthesis, is now an well-established phenomenon. 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. The codon usage pattern varies significantly between different organisms(Grantham *et al.*, 1980) and also between genes that are expressed at different levels in the same organism(Ikemura, 1985). Several hypotheses prevail regarding the factors (Salim and Cavalcanti, 2008) which influence the codon usage pattern. Codon biases are mainly influenced by mutational pressure (Osawa *et al.*, 1988; Sueoka, 1988) and natural selection (Akashi, 1994; Sharp and Li, 1986; Sharp *et al.*, 1993) 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 (Duret and Mouchiroud, 1999; Lobry and Gautier, 1994; Tao and Dafu, 1998). 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 (Ikemura, 1981;Ikemura, 1985). 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.

Information on the codon usage pattern can provide significant insights into the prediction and design of highly expressed genes (Bennetzen and Hall, 1982; Carbone *et al.*, 2003; Das *et al.*, 2009; Das *et al.*, 2012; Roymondal *et al.*, 2009; Fox and Erill, 2010; Khandia *et al.*, 2019; Sharp and Li, 1987; Supek and Vlahovicek, 2005; Wright, 1990). The highly expressed genes are often characterized by strong compositional bias in terms of codon usage. Several quantitative measures have been developed in this regard to provide numerical indices to predict the expression level of genes. All these heuristic or model based approaches are based on assumption as well as contextual knowledge about the organisms under study. Some of these measures include the effective number of codons (N_c) (Wright, 1990), the frequency of optimal codons (F_{opt}), codon bias index (CBI) (Bennetzen and Hall, 1982) and relative synonymous codon usage (RSCU) that depend on the deviation from uniform codon usage. The expression measure (E_g) (Carbone *et al.*, 2003) or codon adaptation index(CAI) (Sharp and Li, 1987) rely on the knowledge of codon bias of a reference set of highly expressed genes while the score of relative codon bias (RCBS) or the score of modified relative codon bias (MRCBS) (Roymondal *et al.*, 2009; Das *et al.*, 2009; Das *et al.*, 2012) 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 *Aspergillus fumigatus* from a whole-genome perspective without any dependence on the choice of the reference set. Our objective of

this work is to perform an analysis of codon usage pattern using various codon bias indices and identify the highly expressed genes of *Aspergillus fumigatus*.

Materials and methods

The whole-genome sequence of *Aspergillus fumigatus Af 293* along with the gene annotations was retrieved from the Gen Bank database at the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). All gene sequences under study along with those annotated as hypothetical have been extracted from the Gene Bank Accession Nos: NC_007194.1 (Chromosome 1), NC_007195.1 (Chromosome 2), NC_007196.1 (Chromosome 3), NC_007197.1 (Chromosome 4), NC_007198.1 (Chromosome 5), NC_007199.1 (Chromosome 6), NC_007200.1 (Chromosome 7) and NC_007201.1 (Chromosome 8). The coding sequences with internal stop codons, not-translatable codons, and incomplete start and stop codons were excluded from the analysis. Therefore, for the present study, finally, 9623 coding sequences were considered for analysis.

Analysis of nucleotide and dinucleotide organization

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, 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 *Aspergillus fumigatus* genome. 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 the average GC content 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 dinucleotide composition also plays 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 (Lytras and Hughes, 2020) 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.

Analysis of Codons usage pattern

The Relative Synonymous Codons Usage (RSCU) (Sharp and Li, 1986) has 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 expected frequency, given that all codons for a specific amino acid are used equally. Codons showing an RSCU value of 1 means no synonymous bias in the

codon usage pattern of the gene, while codons with RSCU values >1 or <1 are showing positive or negative synonymous codon bias, are preferred or un-preferred codons for efficient translation, respectively. RSCU has been calculated by using the following equation:

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

Where *Xij* is the observed number of the i[th] codon for j[th] amino acid which has n_i number of synonymous codons for the amino acid.

The Relative codon adaptation (RCA) (Fox and Erill, 2010) has been proposed to describe the codon usage pattern under the assumption of random codon usage in genes under study. RCA was calculated by determining the ratio of 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 RCA value of 1 means no codon bias or the codon usage frequency is similar to the expected value, while codons with RCA values >1 or <1 are showing overrepresented or underrepresented codons (with respect to a randomized sequence) respectively in respect of compositional bias of nucleotides.

RCA has been calculated by using the following equation:

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

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 RCA 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 RCA>1, whereas for rare codons both RSCU <0.5 and RCA <0.5. The effective number of codons (ENC) (Wright, 1990) is a simple and absolute measure of codon usage bias in the use of synonymous codons. The value of ENC ranges from 20 to 61, with lower ENC values (<35) indicating strong codon usage bias of a gene. The effective number of codons has been calculated by

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

Where F_K values for k-fold degenerate amino acids can be estimated by

$$F_k = (m\sum_{i=1}^{k}\left(\frac{m_i}{m}\right)^2 - 1)/(m-1)$$

Where m_i is the number of occurrences of ith codon for the k-fold degenerate amino acid having total m number of synonymous codons.

The neutrality plot is an analytical method to analyze the influence of mutation bias and natural selection on codon usage. In neutrality plot, a regression line was plotted between average GC contents at the first and second synonymous codon positions (GC12) and GC3 contents at the third synonymous codon position. 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 (Khandia *et al.*, 2019).

The ENC-Plot (ENC vs GC3s) is commonly used to determine whether the codon usage of a gene is affected by mutation or selection.

The ENC-plot is the comparison of the observed and expected distribution of genes based on GC3s on a single plot. Expected ENC values for all GC3s compositions were calculated using the equation (Chen, 2013)

$$ENC_{exp} = 2 + S + \frac{29}{S^2 + (1-S)^2}$$

Where S corresponds to the GC3s value and used to plot standard curve. Data points located on or just below the standard curve (ENC_{exp}) indicate mutational pressure determines the codon usage bias, while data points located far away from the standard curve indicate that factors other than mutational pressure are affecting the codon usage bias.

The Codon Adaptation Index (CAI) a measure of codon bias based on RSCU values of the codons in reference to a set of highly expressed genes is given by (Sharp and Li, 1987),

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

Where N is the number of codons in the gene and relative adaptiveness of an ith 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 jth amino acid and $RSCU_{i,,max}$ is the RSCU value of the most frequent codon used for encoding jth amino acid in a reference to a set of highly expressed genes.. The score measured by CAI ranges from 0 to 1 indicating that the higher is 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 a reference set of highly expressed genes.

The codon bias index (CBI) is another measure of codon bias based on the degree of optimal codons used in a gene. The value of CBI can be expressed by the following formula Bennetzen and Hall, 1982):

$$CBI = \frac{N_{opt} - N_{ran}}{N_{tot} - N_{ran}}$$

Where, N_{opt} represents the total number of occurrences of the superior codons in the gene; N_{ran} represents the sum of the number of occurrences of the superior codon when all the synonymous codons are random in a certain protein; N_{tot} represents the occurrence number of the amino acid corresponding to the superior codons in the gene.

The Frequency of Optimal Codons (FOP) (Ikemura, 1981) is a species-specific measure of bias towards optimal **codons**. It is measured by the number of opimal codons divided by the sum of the optimal and non-optimal codons. The optimal codons have been defined as codons which are recognized by the most abundant isoaccepting tRNAs.

Results and discussion

In the present study, we have analyzed the codon usage pattern of the *Aspergillus fumigatus* genome with respect to nucleotide compositions at synonymous and non-synonymous sites of the codons, dinucleotide composition to understand the compositional properties that greatly influence the codon usage bias. 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. The preference for one type of codon over another can be greatly influenced by the nucleotide composition of the genome. We found the nucleobases C (27.52%) and G (26.68%) occurred more frequently than nucleobases A (23.68%) and T (22.12%). Although the occurrences of the nucleobases C and G were slightly higher and followed by A and T, the base usages are almost uniform. The *Aspergillus fumigatus* genome is rich with C content having a mean value of 0.275. The same trend was found for nucleotide composition at the synonymous 3rd codon position and observed the dominance of $C3_s$. The nucleobase C occurred most frequently at the third codon position (39.69%) and A occurred the least (21.38%). The overall nucleotide composition and the composition at the third codon position in the coding sequences of genes suggested that compositional constraint might influence the codon usage pattern of these genes. The average GC content of the protein coding genes was 54.30%, i.e., genes were GC-rich, which was different from GC content at the first (57.29%), second (44.83%), and third (60.50%) position of codons. The GC content at the third position was higher than GC content at the first and second codon positions and the greatest difference of GC content was found between second and third codon position. For the genome under study, the average GC value of the genes lies between 0.685 to 0.374 [Fig. 1] and the $GC3_s$ score varies between 0.888 to 0.195 [Fig.1]. Many researchers have argued that GC content or $GC3_s$ may be viewed as

the primary influence on the codon usage pattern and thus on the expression profile.

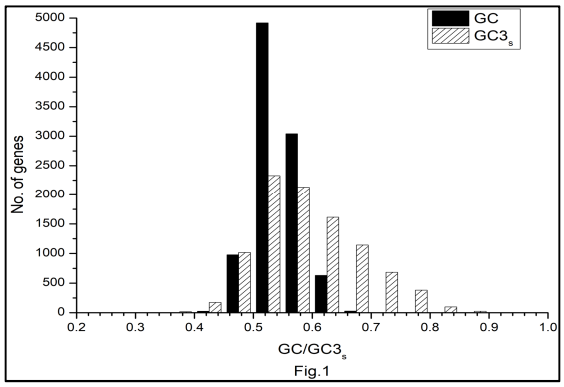

Fig. 1. Distribution of GC and $GC3_s$ in protein-coding genes of *Aspergillus fumigatus* genome under study.

Dinucleotide composition may have consequences on the intrinsic characteristics of the codon usage pattern. In the case of the *Aspergillus fumigatus* genome, the calculated frequency ratios did not deviate much from 1 for most dinucleotides, but there are some exceptions [Fig. 2]. The dinucleotides TpC and GpA showed slight over-representation in *Aspergillus fumigatus*. 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.

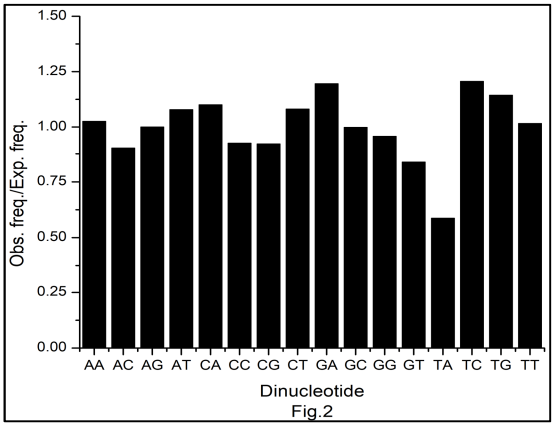

Fig. 2. The ratio of observed and expected frequencies of dinucleotides in protein-coding genes of *Aspergillus fumigatus* genome under study.

Codon usage profile of the *Aspergillus fumigatus* genome has been described in terms of the relative synonymous codon usage, RSCU, and the relative strength codon adaptation, RCA of protein-coding sequences of the genome [Fig. 3]. 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. RCA and RSCU of 61 codons have been displayed in Table 1. The codons having RSCU greater than 1.0 are preferred codons for increasing translational efficiency of the protein-coding genes, whereas codons having RCA greater than 1.0 are overrepresented codons for the organism under study. The preferred codons in *Aspergillus fumigatus* are found to be used in coding Ala (GCC, GCT), Arg (CGG, CGC), Asn (AAC), Asp (GAC), Cys (TGC), Gln (CAG), Glu (GAG), Gly (GGC), His (CAT), Iln (ATC, ATT), Leu (CTC, CTG, CTT, TTG), Lys (AAG), Phe (TTC, TTT), Pro(CCC, CCT), Thr (ACC), Tyr (TAC), Val (GTC, GTG). Importantly, these codons reflect a simple compositional bias. Most of the preferred codons have C or G at the 3rd codon position. Out of 26 preferred codons, 13 are C ending codons. Whereas, Ala (GCA, GCT), Arg(AGA, CGA, CGC), Asp(GAT), Lys(AAG), Gln(CAA, CAG), Glu(GAA, GAG), Gly(GGC, GGT), Iln(ATC, ATT), Leu(CTC, CTG, CTT, TTG), Lys(AAA, AAG), Met(ATG), Phe(TTC, TTT), Pro(CCA, CCT), Ser(AGC.TCA, TCC, TCG.TCT), Thr(ACA), Trp(TGG), Tyr(TAC). 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 (Zhao *et al.*, 2017).

In the present study, we observed that GCT(Ala), CGC(Arg), CAG(Gln), GAG(Glu), GGC(Gly), ATC, ATT(Iln), CTC, CTG, CTT, TTG(Leu), AAG(Lys), TTC, TTT(Phe), CCT(Pro), TAC(Tyr) are optimal (RSCU>1 and RCA>1) whereas, GTA(Val) is the rare codon (RSCU<0.5 and RCA<0.5). The low relative abundance of TpA in rare codons, and the relatively

high abundance of TpC, TpT, ApG, CpT, and TpG in the choice of preferred codons indicate the influence of selection pressure in codon usage bias in the *Aspergillus fumigatus* 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 (Zhou *et al.*, 2016). To explore the amino acid usage trend in *Aspergillus fumigatus* genes, we calculated the number of each amino acid for all ORFs across the genome. The wide variation of amino acid usage was observed among genes. The mean amino acid usages of leucine, alanine, and serine were high for the novel virus, while amino acids such as tyrosine, histidine, methionine, tryptophan, and cysteine were low [Fig. 4].

Table 1. The RCA and RSCU of 61 codons of *Aspergillus fumigatus* genes understudy.

Codon	RCA	RSCU	Codon	RCA	RSCU	Codon	RCA	RSCU
GCA	1.18569	0.81787	GGC	1.21574	1.40772	CCG	0.80452	0.96038
GCC	0.97537	1.2571	GGG	0.71109	0.72646	CCT	1.08309	1.07681
GCG	0.79883	0.90838	GGT	1.13744	0.96781	AGC	1.07223	0.7497
GCT	1.07347	1.01666	CAC	0.53152	0.99613	AGT	0.92299	0.47422
AGA	1.03366	0.77113	CAT	0.72894	1.00387	TCA	1.43737	0.52815
AGG	0.54652	0.67213	ATA	0.51743	0.35235	TCC	1.21694	0.8355
CGA	1.37269	0.99138	ATC	1.27472	1.62192	TCG	1.22151	0.73992
CGC	1.10236	1.48758	ATT	1.09706	1.02572	TCT	1.33305	0.67252
CGG	0.95341	1.13514	CTA	0.69615	0.49363	ACA	1.15552	0.88521
CGT	0.95063	0.94266	CTC	1.13923	1.50938	ACC	0.92247	1.32041
AAC	0.87179	1.13905	CTG	1.38723	1.6216	ACG	0.70991	0.89654
AAT	0.89672	0.86095	CTT	1.07414	1.04576	ACT	0.85361	0.89785
GAC	0.92847	1.02957	TTA	0.6071	0.30518	TGG	1.62414	1
GAT	1.19095	0.97043	TTG	1.23623	1.02446	TAC	1.01859	1.17014
TGC	0.77565	1.22237	AAA	1.25841	0.68153	TAT	0.98306	0.82986
TGT	0.67151	0.77763	AAG	1.47676	1.31847	*GTA*	*0.42646*	*0.40377*
CAA	1.25769	0.75273	ATG	1.15218	1	GTC	0.86429	1.52897
CAG	1.26412	1.24727	TTC	1.64832	2.53854	GTG	0.72564	1.1326
GAA	1.52009	0.81524	TTT	1.29139	1.46146	GTT	0.71899	0.93465
GAG	1.34003	1.18476	CCA	1.19173	0.86294			
GGA	0.95245	0.89801	CCC	0.81293	1.09987			

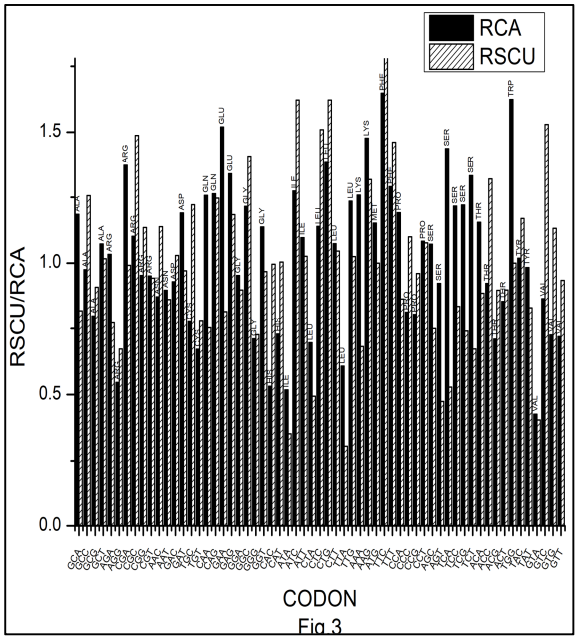

Fig. 3. The RCA and RSCU values of 61 codons of *Aspergillus fumigatus* genes under study.

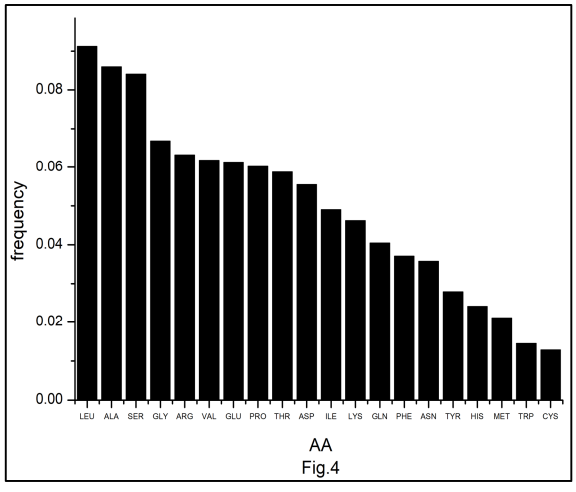

Fig. 4. The frequencies of amino acids in protein-coding genes of *Aspergillus fumigatus* genome under study.

In this study, we compared the performances of several commonly used computation tools for measuring prevalent codon bias in an organism.

The codon usage bias of the *Aspergillus fumigatus* genome were analyzed in terms of CAI, N_c, F_{op}, and CBI. The CAI and CBI scores depend on the reference set of highly expressed genes. using CodonW (available at http://sourceforge.net/projects/codonw) the CAI scores were calculated in reference to *e. coli*.

The CBI scores and F_{op} were calculated in reference to *Aspergillus nidulans*. The value of codon-based expression indicators can perhaps be appreciated by comparing results with the experimental gene expression data in general. In this study, we compared our results with the experimental dataset that stems from transcriptional response of *Aspergillus fumigatus* strains upon exposure to human airway epithelial cells which led to the identification and quantification of about 7600 proteins using $\log_2$(TPM) (transcripts per kilobase million) values. To assess the codon based indicators for predicting protein expression levels, we derived the pair wise correlation coefficient of codon based indicators with the expression data set.

We observed that the agreement of predicted and actual protein expression level varied greatly between all examined combinations of prediction method and data set (r_{CAI}=0.379 [Fig.5], r_{CBI}=0.493 [Fig.6], and rN_c=-0.416 [Fig. 7] and rF_{op}=0.489 [Fig.8]).

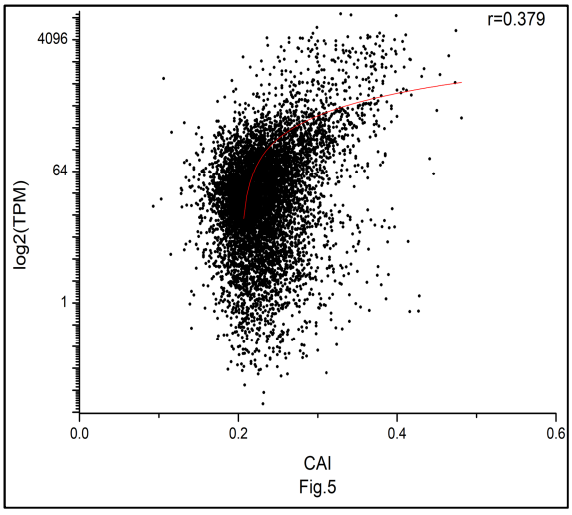

Fig. 5. Transcripts per kilobase million [$\log_2$(TPM)] plotted against CAI in Log2 scale for a set of 7600 identified genes in *Aspergillus fumigatus* genome

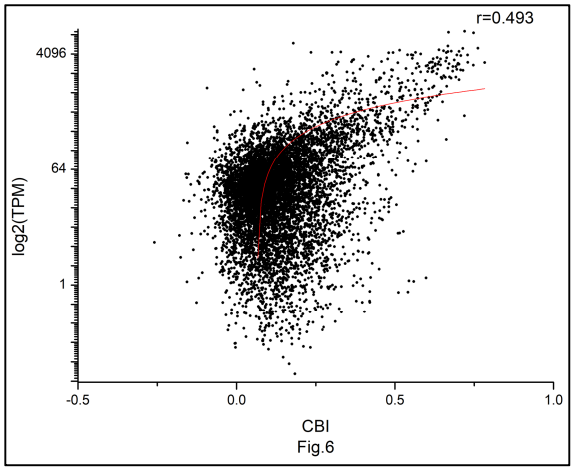

Fig. 6. Transcripts per kilobase million [$\log_2$(TPM)] plotted against CBI in Log2 scale for a set of 7600 identified genes in in *Aspergillus fumigatus* genome.

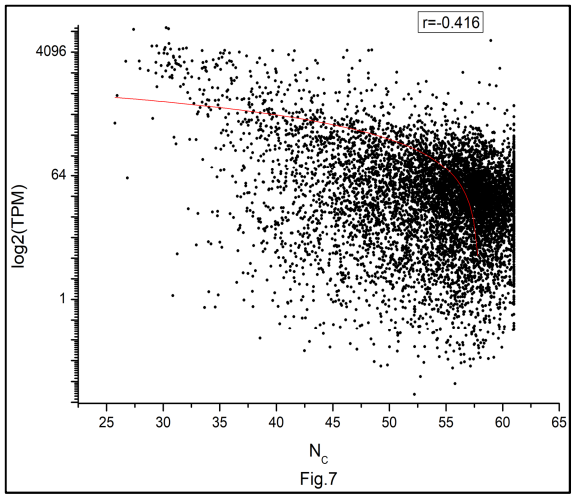

Fig. 7. Transcripts per kilobase million [$\log_2$(TPM)] plotted against N_C in Log2 scale for a set of 7600 identified genes in in *Aspergillus fumigatus* genome.

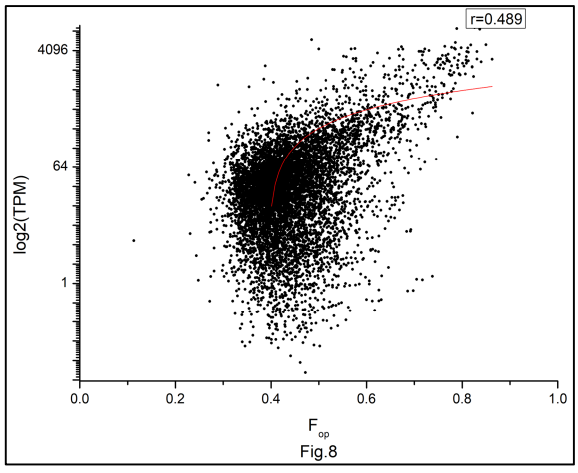

Fig. 8. Transcripts per kilobase million [$\log_2$(TPM)] plotted against F_{op} in Log2 scale for a set of 7600 identified genes in in *Aspergillus fumigatus* genome.

The correlation of CBI with CAI is 0.734 [Fig.9], whereas the correlation between F_{op} and CAI is 0.756 [Fig.10]. The novel method of quantitatively predicting gene expressivity by CBI is then compared with F_{op} and the correlation between them is found to be surprisingly good (r=0.990)[Fig. 11].

The correlation of the codon usage index with Nc is very much significant. The correlations of N_C with CAI is -0.615. The strong negative correlation between CBI and Nc (r=-0.797) [Fig.12] compared to F_{op} and N_c (r=-0.785) [Fig.13] indicates that synonymous codon preferences have been taken into account in CBI.

These correlation coefficients can be used to express the strength of the existing prediction methods.

It can be seen that CBI consistently yields a better correlation than others. We also observe that there are clear correlations between CBI with $GC3_s$ (r=0.740), $C3_s$(r=0.905) and A3(r=-0.813) [Fig. 14-16], where correlation with GC is not much significant (r=0.548). 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.

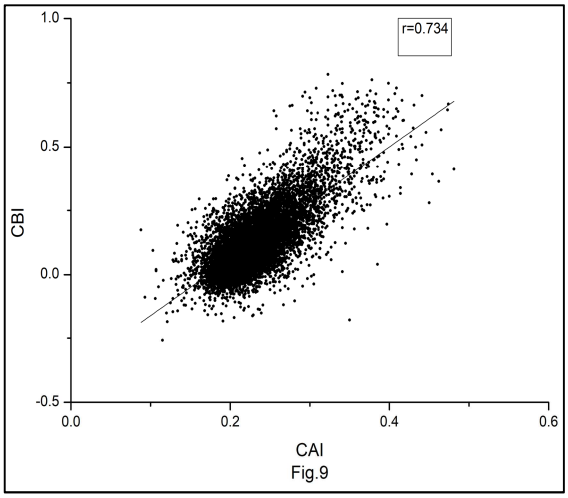

Fig. 9. CBI plotted against CAI for each protein-coding genes in *Aspergillus fumigatus* genome under study.

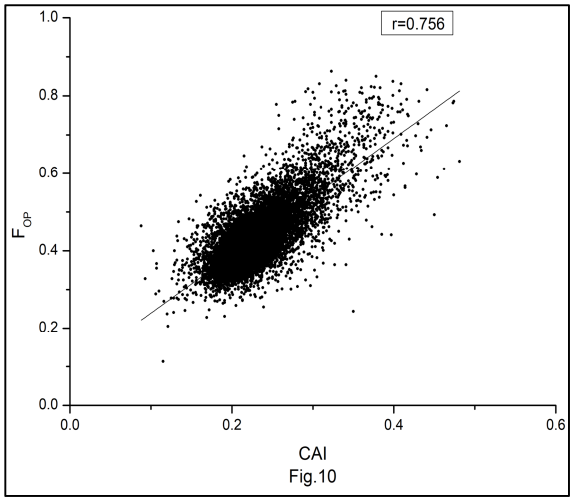

Fig. 10. F_{op} plotted against CAI for each protein-coding genes in *Aspergillus fumigatus* genome under study.

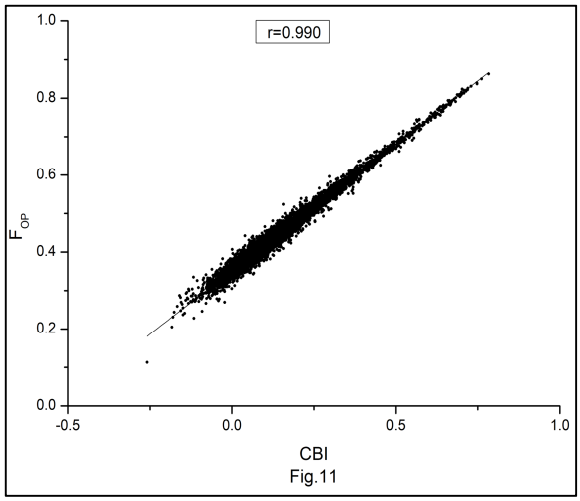

Fig. 11. F_{op} plotted against CBI for each protein-coding genes in *Aspergillus fumigatus* genome under study.

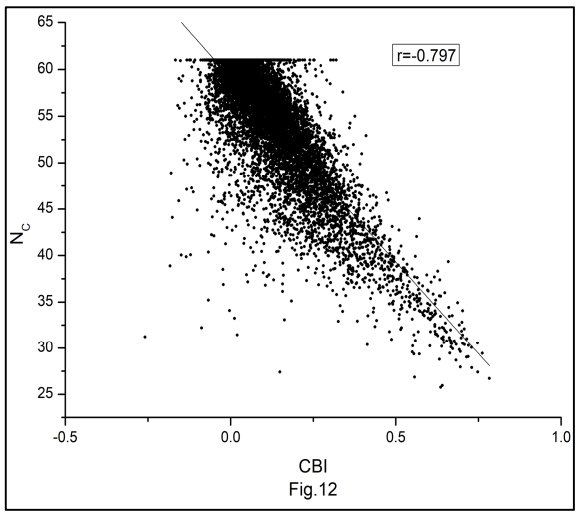

Fig. 12. CBI plotted against N_C for each protein-coding genes in *Aspergillus fumigatus* genome under study.

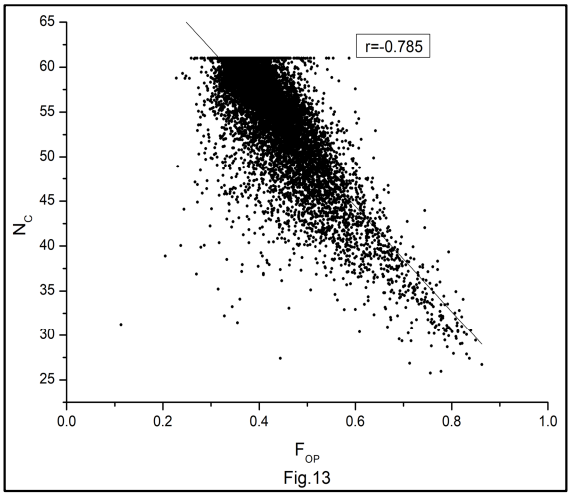

Fig. 13. F_{op} plotted against N_C for each protein coding-genes in *Aspergillus fumigatus*genome under study.

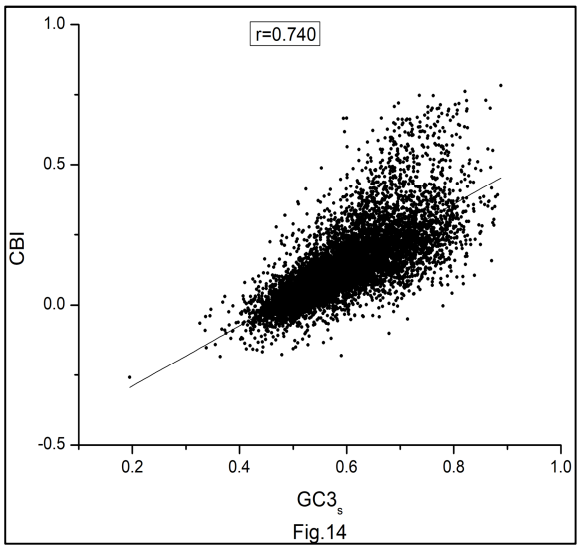

Fig. 14. CBI plotted against $GC3_s$ for each protein-coding genes in *Aspergillus fumigatus*genome under study.

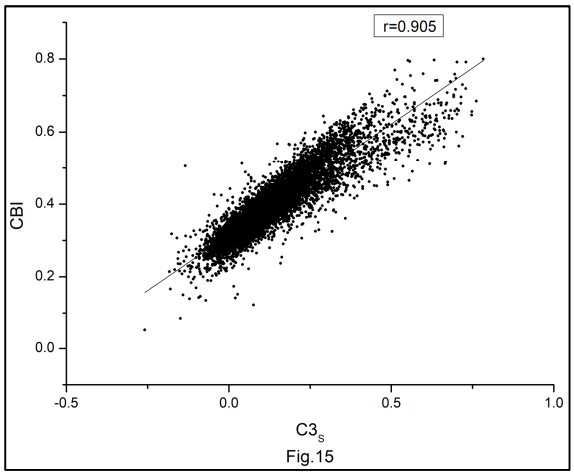

Fig. 15. CBI plotted against $C3_s$ for each protein-coding genes in *Aspergillus fumigatus*genome under study.

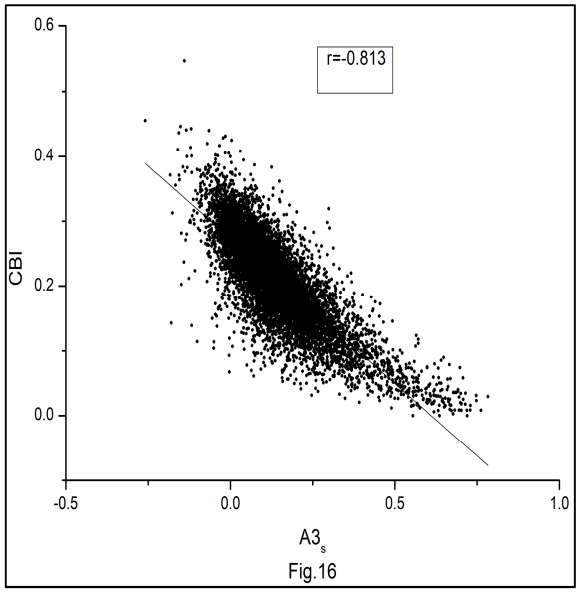

Fig. 16. CBI plotted against $A3_s$ for each protein-coding genes in *Aspergillus fumigatus*genome under study.

Expression profiles of the genes are determined by calculating CBI for each gene and their distributions are shown in Fig. 17. The majority of genes (92%) have CBI lying between -0.05 and 0.35.

The threshold score for identifying highly expressed genes has been determined by the z score of CBI 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 CBI has been calculated to be 0.384. Only 4.5% of genes of the *Aspergillus fumigatus* genome have CBI values greater than 0.384. The genes with z score less than -2.00 may be assumed to have low expression levels.

The overall variation of GC or GC3 content of the genes is depicted in Fig. 1. It indicates that the majority of genes (90%) have a GC3 score lying between 0.48 to 0.76 and 97% of genes have GC content lying between 0.46 to 0.62. Table 3 displays the statistics of PHE genes and the top 20 PHE genes of the *Aspergillus fumigatus* genome along with their functions and gene expression level (CAI_g).

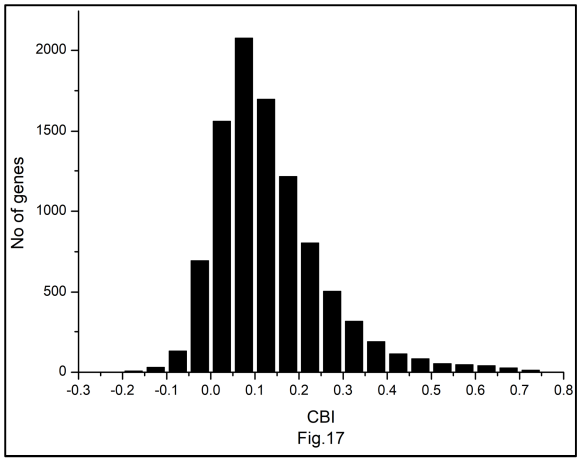

Fig. 17. Distribution of CBI of all protein-coding genes in *Aspergillus fumigatus* genome under study.

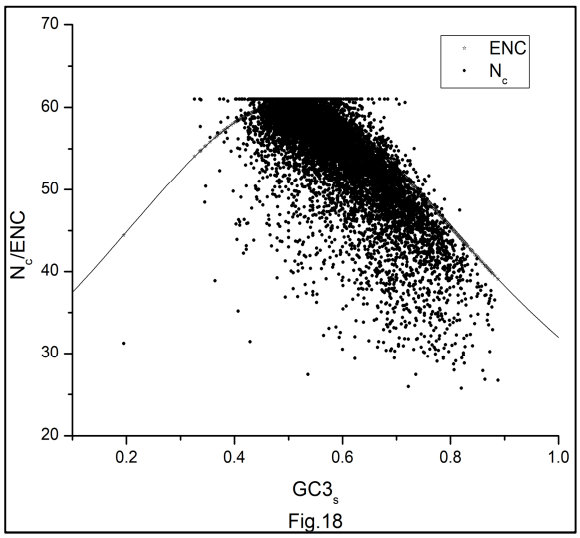

Fig. 18. N_C–$GC3_s$ plot for all protein coding sequences of *Aspergillus fumigatus* genome.

N_C is a measure of bias from equal codon usage in a gene. N_C values are calculated to determine the inter-genic codon bias. In our study, we observed that N_C of *Aspergillus fumigatus* genes ranges from 25.75 to 61 with a mean value of 53.85±6.23. The more biased a gene is, the smaller is the N_C value. In general, codon usage bias in *Aspergillus fumigatus* is low. Only 1.4% of 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. 18] 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 are likely to be involved in determining the codon usage in *Aspergillus fumigatus*. The relative magnitude of magnitude of mutation pressure and natural selection on codon usage bias has been investigated by constructing a neutrality plot (GC12 vs $GC3_s$) [Fig. 19]. The weak correlation (r=0.033) between $GC3_s$ and GC12 suggests that codon usage is influenced by natural selection. In the neutrality plot [Fig. 19], most of the genes are far from the regression line. The slope (0.013) of the regression line indicates the relative neutrality (mutation pressure) was 1.3%, and the relative constraint on $GC3_s$ (natural selection) was 98.7%, indicating the dominant influence of natural selection on the codon usage patterns of *Aspergillus fumigatus*. To find out the putative relationship of base composition at different synonymous codon position, when the values were compared at third synonymous codon positions, the significant correlations were observed only between $A3_s$ and $C3_s$ (r =-0.755), $C3_s$ and $GC3_s$(r=0.825), and $A3_s$ and $AT3_s$(r=0.835).

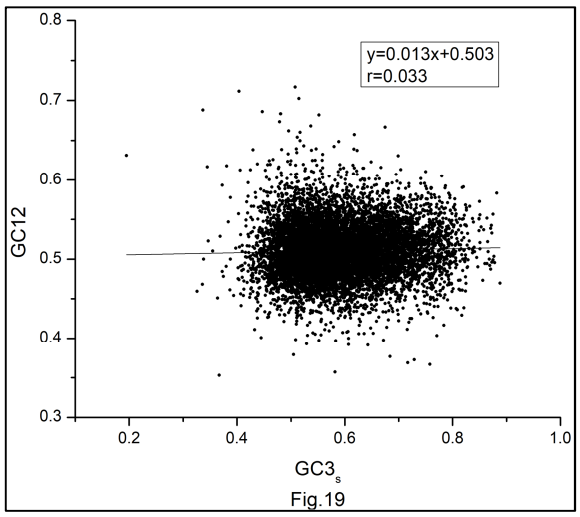

Fig. 19. The neutrality plot (GC12 vs $GC3_s$) for all protein coding sequences of *Aspergillus fumigatus* genome.

The large data set analyzed here revealed that selective constraints play a major influencing factor towards a strong codon usage bias of different set of preferred codons in genes with high cytoplasmic mRNA levels. In contrast, genes with low mRNA levels showed very little synonymous codon usage bias.

The average Nc value of the PHE genes is found to be 37.82 whereas, that for the predicted genes having low expression is 51.97. The codon usage bias in *Aspergillus fumigatus* was proposed as a result of translational selection, since using a codon that is translated via an abundant tRNA species were 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 4.5% of genes of *Aspergillus fumigatus* belong to PHE genes. The preferred sets 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 CBI and GC content is not significant (r=0.548). However, a strong negative correlation between CBI and N_c (r=-0.797) suggests that highly expressed genes display more biased codon usage than the lowly expressed genes. We observed that PHE genes of *Aspergillus fumigatus* mostly include ribosomal protein (RP) genes, translation initiation factors, translation elongation factors, transcription factor, chaperon, heat shock protein, histone, and many binding protein genes. Our analysis predicted 435 highly expressed genes in *Aspergillus fumigatus*. It has been observed that PHE genes belonged to various functional classes and variably represented in the genome.

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.

Conclusion

The present study demonstrates that the codon usage pattern is largely responsible for the regulation of gene expression and CBI may be a useful tool for predicting highly expressed genes. Our results indicate the dominant influence of natural selection on the codon usage patterns and less bias in codon usage among genes in the *Aspergillus fumigatus* genome. In this study, various approaches to estimating gene expression levels based on codon usage have been applied to the *Aspergillus fumigatus* genome. The predicted gene expression level using the quantitative measure by CBI was found to correlate well with CAI, F_{op} and N_c. The strong negative correlation between CBI and N_c supports the hypothesis that highly expressed genes are strongly biased. Given that a complete genome sequence is available, such computational tools may be helpful in extracting meaningful information for understanding the details of functional genomics and sets the ground for future investigation in biotechnological application regarding heterologous protein expression.

Conflicts of interest

The authors declare that they have no conflicts of interest related to this study.

Acknowledgement

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

References

Abad A, Fernández-Molina JV, Bikandi J, Ramírez A, Margareto J, Sendino J, Hernando FL, Pontón J, Garaizar J, Rementeria A. 2010. What makes *Aspergillus fumigatus* a successful pathogen? Genes and molecules involved in invasive aspergillosis. Revista Iberoamericana de Micología **27(4),** 155-82.

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

Bennetzen JL, Hall BD.1982. Codon selection in yeast. Journal of Biological Chemistry **257,** 3026-3031.

Beauvais A, Latgé JP. 2001. Membrane and cell wall targets in Aspergillus fumigatus. Drug Resistance Updatates **4(1),** 38-49.

Carbone A, Zinovyev A, Fékèps F. 2003. Codon adaptation index as a measure of dominating codon bias. Bioinformatics **19,** 2005-2015.

Chen Y. 2013. A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection. BioMed Research International, Article Id: 406342. https://doi.org/ 10.1155/2013/406342.

Das S, Roymondal U, Sahoo S. 2009. Analyzing gene expression from relative codon usage bias in *Yeast* genome: a statistical significance and biological relevance. Gene **443,** 121-131.

Das S, Roymondal U, Chottopadhyay B, Sahoo S. 2012. Gene expression profile of the *cynobacterium synechocystis* genome. Gene **497,** 344-352.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. . Proceedings of the National Academy of Science of the United States of America **96,** 4482-4487.

Fox JM and Erill I. 2010. Relative Codon Adaptation: A Generic Codon Bias Index for Prediction of Gene Expression. DNA Research **17,** 185-196.

Grantham R, Gautier C, Gouy M, Mercier R, Pave A. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Research **8(1),** r49-62.

Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. Journal Molecular Biology **151,** 389-409.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Molecular Biology and Evolution **2(1),** 13-34.

Khandia R, Singhal S, Kumar U, Ansari A, Tiwari R, Dhama K, Das J, Munjal A, Singh RK. 2019. Analysis of Nipah Virus Codon Usage and Adaptation to Hosts. Frontiers in Microbiology **10,** 886.

Lobry JR, Gautier C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Research **22(15),** 3174-3180.

Lytras S, Hughes J. 2020. Synonymous Dinucleotide Usage: A Codon-Aware Metric for Quantifying Dinucleotide Representation in Viruses. Viruses **12,** 462.

Osawa S, Ohama T, Yamao F, Muto A, Jukes TH, Ozeki H, Umesono K. 1988. Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets Proceedings of the National Academy of Science of the United States of America **85,** 1124-1128.

Osherov N. 2007. The virulence of Aspergillus fumigatus. In: Kavanagh K, Ed. New Insights in Medical Mycology. Springer, Dordrecht p.185-212.

Roymondal U, Das S , Sahoo S. 2009. Predicting gene expression level from relative codon usage bias:an application to *Escherichia coli* genome. DNA Research **16,** 13-30.

Salim HMW, Cavalcanti ARO. 2008. Factors Influencing Codon Usage Bias in Genomes. Journal of the Brazilian Chemical Society **19(2),** 257-262.

Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. Journal of Molecular Evolution **24,** 28-38.

Sharp PM, Li WH. 1987. The codon adaptation index -a measure of directional synonymous codon usage bias and its potential applications. Nucleic Acids Research **15,** 1281-1295.

Sharp PM, Stenico M, Peden JF, Lloyd AT. 1993. Codon usage – mutational bias, translational selection, or both. Biochemical Society Transaction **21(4),** 835-841.

Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. Proceedings of the National Academy of Science of the United States of America **85,** 2653-2657.

Supek F , Vlahovicek K. 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics **6,** 182.

Tao X, Dafu D. 1998. The relationship between synonymous codon usage and protein structure. FEBS Letters **434,** 93-96.

Tekaia F, Latgé JP. 2005. Aspergillus fumigatus: saprophyte or pathogen? Current Opinion Microbiology **8(4),** 385-92.

Wright F. 1990. The 'effective number of codons' used in a gene. Gene **87,** 23-29.

Zhao F, Yu CH, Liu Y. 2017. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. Nucleic Acids Research **45(14),** 8484-8492.

Zhou Z, Dang Y, Zhou M, Li L, Yu C-H, Fu J, Chen S, Liy Y. 2016.Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proceedings of the National Academy of Sciences of the United States of America **113(41),** E6117-E6125.