Chapter 2

Codon usage bias in bacterial genomes

2.1 Introduction

The universal genetic code table defines the translational mapping of sixty-one codons into twenty amino acids. Except for Met and Trp, all other eighteen amino acids are assigned with two or more synonymous codons for which the genetic code table is called degenerate. These synonymous codons are not used randomly in genes within a genome, a phenomenon known as codon usage bias (CUB), which is common in all organisms [32].

Several mutational factors such as genome G+C%, strand asymmetric nucleotide composition [27][47][114][131][136][152] are known to influence CUB. In addition to these mutational factors, selective forces also influence CUB though in variable strength [179]. During the process of translation certain codons are used more frequently in comparison to other synonyms for faster and/or accurate translation in high expression genes than rest of the genes in a genome [5][203][160][194][227]. This is considered as the primary selection factor influencing CUB [68] which is a common phenomenon in genomes of bacteria. Based on CUB property, several mathematical formulas have been proposed for measuring codon usage bias [174]. These mathematical equations have been developed for measuring CUB from different point of view. Effective number of codons [237] measures overall degree of codon usage bias in a gene, whose value varies between 61.0 and 20.0 representing synonymous codons are being used with equal frequencies, and highly biased synonymous codon usages respectively. Measuring CUB as a departure from uniform usage of alternative synonymous codons is not always desirable. When background nucleotide composition of a gene is considered, the null distribution of codon usage is non-uniform. Novembre (2002) [137] reported a modification over effective number of codons (Nc), taking background GC content of the genome into consideration. Keeping genomic GC composition in view, improvements in these two methods have been proposed in research literate [181][208][49]. Codon adaptation index (CAI) proposed by Sharp and Li estimates the extent to which codons of a gene are adapted towards the optimal codons favored by the set of organism specific highly expressed genes [192]. CAI is a measure for finding predicted gene expression from codon usage bias. CAI value of the genes varies between 1.0 to 0.0. Higher the CAI value of the genes, higher is the adaptation of codon usage of the gene towards optimally used codons in the high expression genes and therefore considered as highly expressed genes.

CUB is distinctly different between the high expression genes (HEG) and the low expression genes (LEG) in an organism [57]. There is stronger selection for both accurate and fast codon translation in the HEG than LEG in a genome [5][203][160][194][227]. For example, among the Leucine codons, CUG is the most frequent codon in both HEG and LEG, but the frequency is higher in HEG than in LEG. This higher frequency of CUG in HEG is attributed to higher cognate tRNA gene abundance for CUG codon in *Escherichia coli* genome, facilitating faster translation [82][84][195]. Other similar notable codons most frequently used in HEG are Glycine codon GGU and Arginine codon CGU [178]. These codons used more frequently in HEG than LEG are considered the optimal codons (OCs) [68][192][195][180]. The anticodons used for decoding GGN codons are not different among archaea, bacteria and eukarya: generally, UCC and GCC anticodons are used [138]. According to the four-column theory for the origin of the genetic code [72], GGN is one of the oldest codon families in genetic code evolution. The conservation of anticodons of Gly family indicates the universal preference of GGU among bacteria. In contrast to Gly, anticodons used for decoding Arg family codons are different among the three kingdoms: UCG and GCG anticodons are used by archaea; ICG and CCG are used by bacteria; and ICG and UCG are used by eukarya (in yeast CCG). In spite of these differences in anticodons used, selection of the U-ending codons has been reported in both archaea and bacteria organisms [227].

Transfer RNA genes can influence translational selection on CUB in different ways [160]. The cytosolic abundance values of isoacceptor tRNAs differ and the synonymous codon with high cognate tRNA abundance might be preferred to the other synonymous codons with low cognate tRNA abundance. Secondly, efficiency of a tRNA molecule might differ in decoding two or more synonymous codons due to difference at the wobble position in codon-anticodon pairing. Further, the codon-anticodon pairing is also influenced by tRNA base modifications [160]. Contribution of these tRNA specific factors on translational selection across bacteria is difficult to quantify. In a comparative study between the tRNA gene number and codon usage bias across genomes in 199 bacteria Satapathy et al. (2012) [177] observed that the tRNA gene numbers may not be completely responsible for the CUB in Asp, Ile, Phe, and Tyr in the high expression genes.

Rare codons are reported to be used as a mechanism to achieve circadian clock conditionality [243] and also for translation elongation to regulate co-translational protein folding because rare codons cause ribosome stalling during translation [246][252][163].

In this regard, in a detailed study [180] observed that there is amino acid specific influence for the selection of optimal codons. There is also influence of phylogeny in the choice of OCs for some amino acids such as Glu, Gln, Lys and Leu.

Apart from the positively selected codons in HEG, specific codons can be observed to be used in lower frequency in HEG compared to LEG, possibly because of the presence of these codons having any retarding effect on optimum gene expression. Open reading frame (ORF) of a gene containing certain rare codons may be wrongly translated possibly because of ribosomal frameshifting error during translation. For example, Proline codon CCC and Glycine codon GGG are known to be prone to translational frameshifting errors [140][141]. Rare codons are reported to be used as a mechanism to achieve circadian clock conditionality [243] and also for translation elongation to regulate co-translational protein folding because rare codons cause ribosome stalling during translation [246][252][163].

For some of the amino acids, only specific synonymous codons are primarily selected as the optimal codons across bacteria: the C-ending codons of Phe, Tyr, Asn, and Ile [193]; GGY and CGY codons for Gly and Arg, respectively [178]. In contrast, none of the synonymous codons for amino acids Lysine and Cysteine is selected as OCs in *E. coli*, even though the bacterium is known to have a strong selected codon usage bias [195]. In this regard, in a detailed study [180] observed that there is amino acid specific influence for the selection of optimal codons. There is also influence of phylogeny in the choice of OCs for some amino acids such as Glu, Gln, Lys and Leu.

Despite a large number of studies on codon usage bias available in the literature, a detailed codon specific generalized study across bacteria regarding gene expression is still having ample scope. The study by Sharp et al. (2005) mainly focuses on two-fold degenerate AT rich codons in four amino acids. The study by Satapathy et al. (2014) [178] has limitations regarding number of genes used for gene expression analysis. Further, these two above studies have not emphasized on codons negatively selected in the HEG. In this study, we have used a machine learning based approach to rank codons in their effectiveness in classifying genes into the high and the low expression genes in more than 600 bacterial species, that identifies several codons that are universally influenced by gene expression and certain codons that are not influenced by gene expression in bacteria. Along with that, we have tried to solve the limitations of existing implementations of CAI.

2.2 Importance of the codons towards classifying high and low expression genes: a machine learning-based analysis

We have carried out a detailed codon usage analysis in *Escherichia coli* K-12 MG1655 genome. Gene sequences of *E. coli* were downloaded from the NCBI database. To avoid any biased result due to missing amino acids, we have considered larger genes with a size of more than 100 amino acids (aa). Out of all the *E. coli* genes, we have considered 893 genes for our analysis. These genes are more than 100 aa length whose protein level expressions have been reported [85]. These genes were further arranged in decreasing order of expression level. The top and bottom one-third genes were considered high expression genes (HEG) and low expression genes (LEG). In total, there were 297 genes each in HEG and LEG sets (Appendix A.1.2). For codon usage analysis, we also considered another 683 bacteria species (Appendix A.1.3). As gene ex- pression was not available for these organisms, we estimated HEG and LEG gene sets for each organism, considering codon adaptation index (CAI) values [192]. Considering organism-specific ribosomal protein-coding genes as the reference set, we calculated CAI values of the genes in each bacterium using a web-based tool [192][185]. As per the mathematical formula used for CAI, maximum and minimum possible theoretical values of CAI for a gene are 1.0 and 0.0 respectively. CAI value of a gene is equal to 1.0 when all the codons present in the genes are such that they are the optimally used in the reference set of high expression genes. For these codons w value is 1.0 in equation 1.3. In the other extreme case, when all the non-optimal codons are used for which w values are 0.0, CAI value equals to 0.0. However, in reality, the range of the CAI values differ from organism to organism. Distributions of CAI values for the genes of two bacteria, *Bradyrhizobium japonicum* (Genome G+C 64.06%), and *Staphylococcus aureus* (Genome G+C 32.84%) are given in Figure 2-1, Figure 2-2. For the machine learning analysis, in each bacterium, we calculated CAI values for the genes with size more than 100 amino acids, arranged genes in the decreasing order of CAI values and considered top 100 genes as HEG and bottom 100 genes as LEG. We ignored smaller genes because some of amino acids might be missing in these genes leading to biased result. CAI values in HEG set were always more than the values in LEG set among the bacteria studied here. For example, in *Bradyrhizobium japonicum*, average CAI value for HEG was 0.781 and LEG was 0.212 and in *Staphylococcus aureus*, average CAI value for HEG was 0.750 and LEG was 0.362. Therefore, machine learning based study with other different sets of HEG and LEG genes lead to similar conclusion on codon importance values (data not shown).

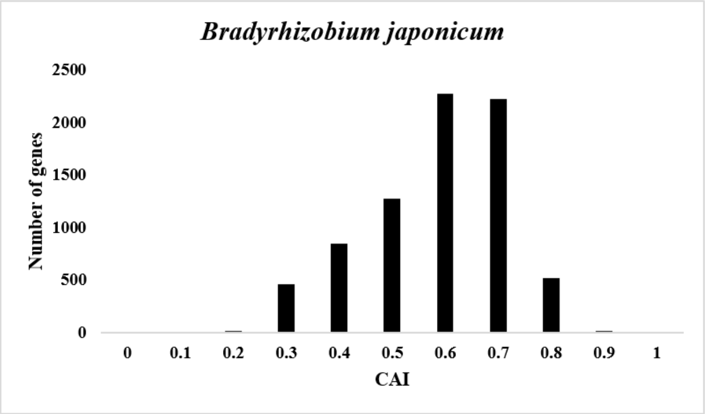


Figure 2-1: Distribution of CAI values in *Bradyrhizobium japonicum* bacteria

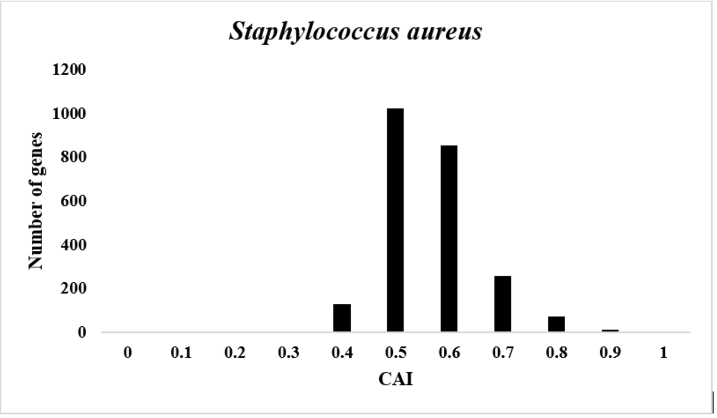


Figure 2-2: Distribution of CAI values in *Staphylococcus aureus*

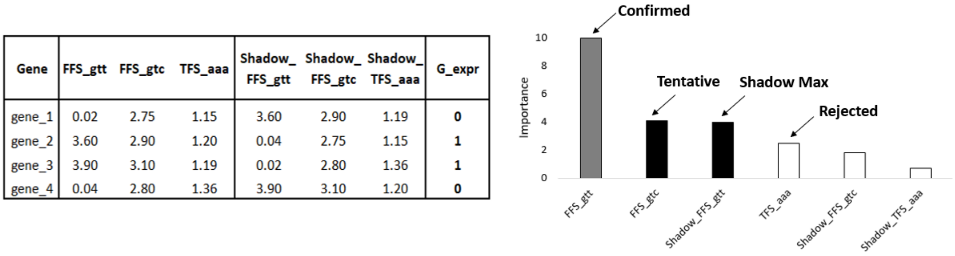
2.2.1 Machine learning-based estimation of the importance of the codons in classifying high and low expression genes

The dataset of our work can be represented as D={X,y}, where X is a 2-dimensional matrix and any particular row r of X, $x_r = [x_r^1, x_r^2, x_r^3, , x_r^{(n-1)}, x_r^n]$, consists of n RSCU values. Here the value of n is 59 (except codons of Met and Trp of degeneracy one and 3 stop codons). The target variable y is a one-dimensional integer-valued vector such that y∈0,1, where 0 represents class of LEG and 1 class of HEG. Number of synonymous codons encoding one amino acid differs among amino acids: Met and Trp are encoded by one codon each; Phe, Tyr, His, Gln, Asn, Lys, Asp, Glu, and Cys are encoded by two codons each; Val, Pro, Thr, Ala, and Gly are encoded by four codons each; Three codons encode Ile; Leu, Ser and Arg are encoded by six codons each. Accordingly, codon degeneracy for amino acids is 1/2/3/4/6 depending on the number of synonymous codons for an amino acid. RSCU values of these codons are used to estimate the codons' importance towards classifying high and low expression genes, as described in the following section.

We have used the Boruta algorithm [101][102] to estimate the importance of the codons towards classifying high and low expression genes. This algorithm is a wrapper built around the random forest classifier [19][108]. In this algorithm, values of a feature are randomized to create shadow features of each attribute (Figure 2-3) to remove any correlation with the target variable. Due to the randomization of the feature values, there is a loss in accuracy incurred in random forest classification. The extent of accuracy loss in terms of Z-scores is considered for estimating the codons' importance. Z-score is defined as the ratio between the mean and standard deviation of accuracy losses for the decision trees for a particular feature generated in the random forest classification. Once the importance of each of the original and shadow features are calculated, the shadow feature with the highest importance is taken as the threshold (shadow Max Z-score). Any feature with an importance value more than the threshold is marked as a confirmed attribute and the one less than the threshold is as a rejected one. If the importance value is close to the threshold for any feature, it is marked as tentative (Figure 2-3). The process is repeated for a maximum number of iterations (say 1000 iterations) to calculate feature importance. Considering the default random forest classifier parameters, importance of the codons is generated using Boruta package in R language [101]. Boruta algorithm has also been used to determine the important features instead of Principal Component Analysis (PCA) in different domains that reports the useful features without reducing the dimensions and having better results [189][55]. A hypothetical example is presented in Figure 2-3, and a flowchart of the algorithm is illustrated in Figure 2-4.

2.2.2 Machine learning-based analysis of CUB between high and low expression genes

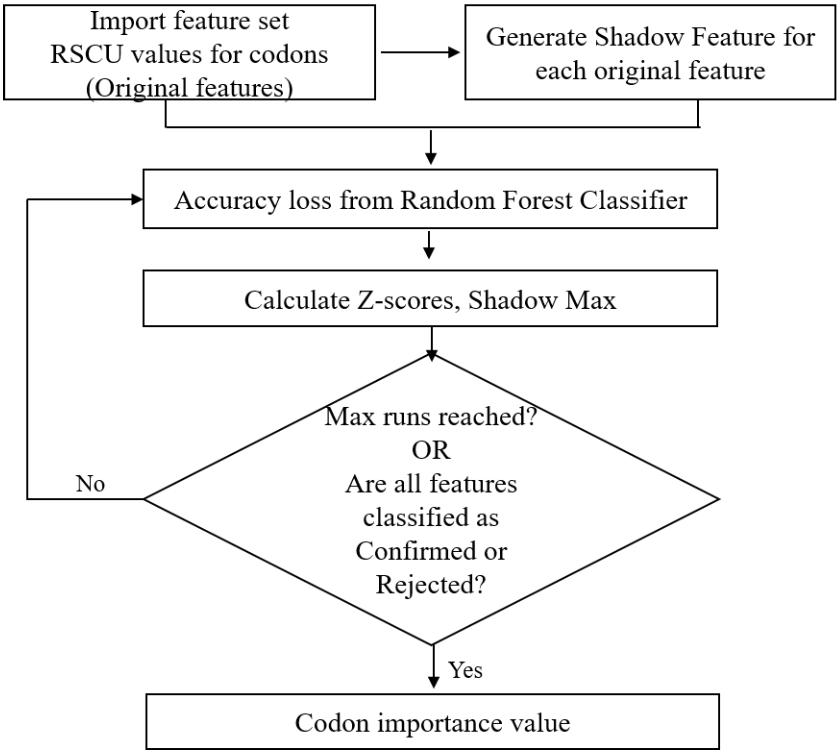
To understand the role of codon degeneracy, we considered two popular classification algorithms, Random Forest described earlier and another Extreme Gradient Boost (XGBoost) [28]. The python language-based sklearn.ensemble [146] package for Random Forest and xgboost [28] package for XGBoost are used in our analysis. The default parameters are considered both in RF and XGBoost classifiers. For



Gene	FFS_gtt	FFS_gtc	TFS_aaa	Shadow_ FFS_gtt	Shadow_ FFS_gtc	Shadow_ TFS_aaa	G_expr
gene_1	0.02	2.75	1.15	3.60	2.90	1.19	**0**
gene_2	3.60	2.90	1.20	0.04	2.75	1.15	**1**
gene_3	3.90	3.10	1.19	0.02	2.80	1.36	**1**
gene_4	0.04	2.80	1.36	3.90	3.10	1.20	**0**

Figure 2-3: Codon importance estimation using Boruta Algorithm with the help of a hypothetical set of genes and RSCU values of a few codons

Figure presents codon importance estimation using Boruta Algorithm with the help of a hypothetical set of genes and RSCU values of a few codons. The table on left hand side shows genes with original and shadow RSCU values for three codons. The shadow codons are randomly shuffled values of original codons and are initialized with the name "Shadow_" for each codon. The last column indicates if a gene is from HEG class or LEG class with 1 and 0 respectively. The figure on the right hand side shows the importance graph of all the original and shadow codons using the method described in section 2.2.2. Any codon with importance value more than threshold (Shadow_FFS_gtt or Shadow_max) is considered as important or confirmed (FFS_gtt); any codon having importance value less than threshold (Shadow_FFS_gtt) is considered as non-important or rejected (TFS_aaa)

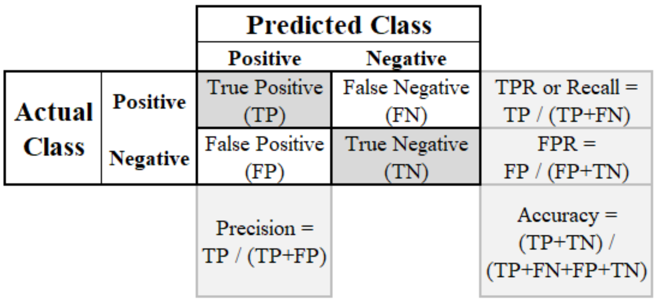


Import feature set
RSCU values for codons
(Original features)
Generate Shadow Feature for each original feature
Accuracy loss from Random Forest Classifier
Calculate Z-scores, Shadow Max
Max runs reached?
OR
Are all features classified as Confirmed or Rejected?
No
Yes
Codon importance value

Figure 2-4: Workflow of Boruta Algorithm

Figure presents workflow of Borutas algorithm. The feature set is imported and corresponding shadow features are created. All of these features are passed to the random forest classifier to retrieve the accuracy loss of each feature. Z-score is calculated from the accuracy loss for each feature. The process is repeated for maximum number of runs or until all the features are classified as confirmed or rejected. Threshold is chosen as the shadow feature having maximum importance value. Based on this threshold, the codons are classified as confirmed or rejected performance measurement, the precision, recall, accuracy, and F1 score were estimated for both classifiers, and ROC curves were also plotted.

Precision measures how accurate the model is out of those predicted positive. Recall measures how accurate our model is out of those actual positive. Accuracy measures the correctly predicted observations out of total observations. F1 score is the weighted average of Recall and Precision. F1 Score = 2*Recall*Precision / (Recall + Precision). All these measures are summarized in Figure 2-5. The higher these values, the better the model is. We also chose the Receiver Operating Characteristics curve (ROC) and its Area Under the Curve (AUC) values to compare the models or feature sets. ROC is a probability curve set on various threshold settings. ROC is a plot against True positive Rate (TPR) or False Positive Rate (FPR). AUC measures how good the model is in account of the separability of classes. AUC values range from 0.0 to 1.0. Higher the AUC value, the better the model [42][214]. For statistical analysis and determining p-value for significance test, Mann Whitney test is used [125].



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN)	TPR or Recall = TP / (TP+FN)
	Negative	False Positive (FP)	True Negative (TN)	FPR = FP / (FP+TN)
		Precision = TP / (TP+FP)		Accuracy = (TP+TN) / (TP+FN+FP+TN)

Figure 2-5: Confusion Matrix along with Precision, TPR or Recall, FPR, and Accuracy
Figure presents confusion matrix and the performance metrices Precision, TPR or Recall, FPR, and Accuracy

2.2.3 Results and Discussion

2.2.3.1 RSCU values of the codons do not contribute equally towards classifying high and low expression genes

It is known that the codon usage is optimized in HEG for faster or accurate translation. In comparison to LEG, some codons are used more frequently in HEG, whereas some other codons are avoided. This difference is prominently visible for some amino acid codons, whereas the difference is not significant for some other codons. Keeping this difference in view, we tried to access the contribution of individual codons towards classifying HEG and LEG using the Boruta algorithm and assigned Z-score-based importance values to each of the codons. The result is presented in Figure 2-6. The range of codon importance values varies between 0.37 and 16.23. As per the Boruta algorithm result, codons are categorized into two groups. First, the codons having Z-score less than the shadow Max Z-score. These codons had similar usage both in HEG and LEG and accordingly had no significance in classification. The range of the Z-scores for the codons in this group was between 0.37 to 2.42. Two-fold degenerate amino acids codons Cys (TGT and TGC) and Lys (AAG and AAA) and split box Arg codons (AGA and AGG) were belonging to this group. The other codons belonging to this group were AGC(Ser), GGC(Gly), CTT(Leu), CCT(Pro), CCA(Pro), ACA(Thr). The remaining codons were categorized under the second group. In the second group, codons had Z-score more than the shadow Max Z-score. The range of the Z-scores for the codons in this group was between 2.67 to 16.24. These codons had different codon usage in HEG and LEG sets and significantly classified HEG and LEG.

In support of the above result, we did a simulation study. Instead of considering two distinct sets of HEG and LEG, we randomly categorized genes into two groups and estimated the importance of the codons using the Boruta algorithm. The distribution of the Z-scores is presented in Appendix A.1.1. It can be seen from the figure that almost all the codons were of very low Z-scores. This

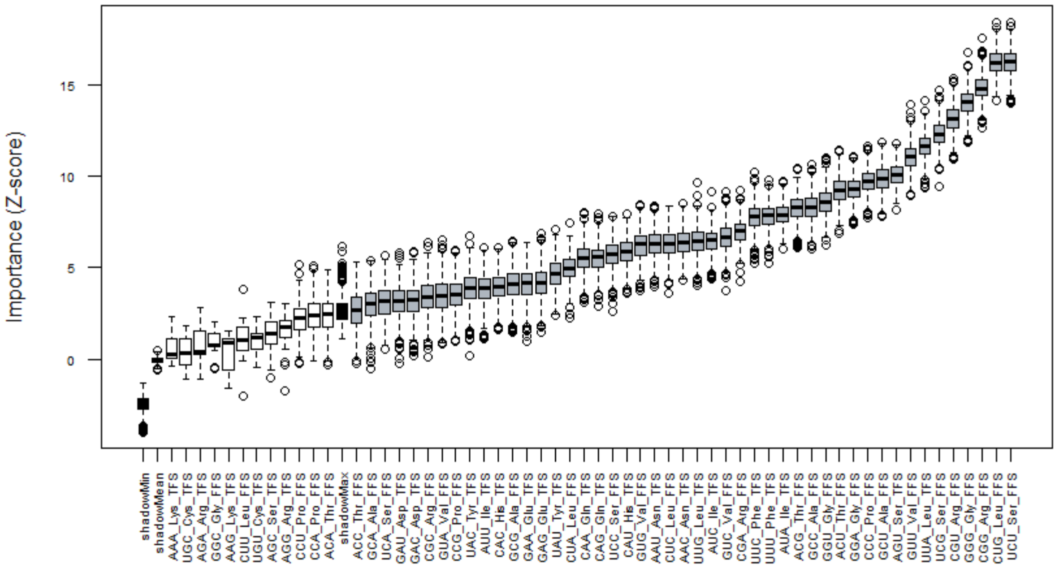


Figure 2-6: Importance of the codons in *E. coli* genome

Figure presents importance of the codons in E. coli genome estimated in terms of Z-scores using Boruta algorithm. The algorithm was run 1000 times. The distribution of Z-scores for the codons was presented in box plots. Codons that contribute to classifying HEG and LEG are shown with gray boxes, and those with negligible contribution are shown with white boxes. The distribution of the Z-score for shadow boxes are shown in black.

confirmed estimation of codon importance is actually because of the difference between high and low expression gene sets.

Further, in support of the result on codon importance, we did a classification study considering set of confirmed codons and set of rejected codons separately. In both the sets of codons, we split the genes in the ratio of 70:30, where 70% of data was used for training purposes and 30% of data for testing purposes to predict the two classes of genes (HEG and LEG). On the training data, a random forest classifier was employed. The testing data was used for the prediction of a class of genes i.e., HEG or LEG. Based on the predictions and unseen target variable, TPR and FPR is calculated on different threshold values. ROC curves with AUC values are generated based on TPR and FPR values to exhibit the difference in classifier performance in the two data sets. The classification result is presented in the ROC curve in Figure 2-7. A considerable difference in AUC values of both sets of codons can be observed.

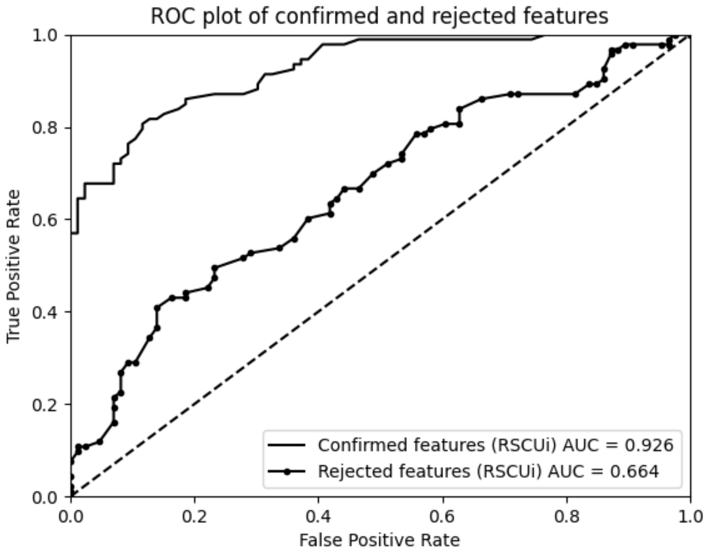


Figure 2-7: ROC curve of confirmed and rejected features using RF classifier Figure presents ROC curves with AUC values generated based on TPR and FPR values to exhibit the difference of classifier performance in the two sets of confirmed and rejected features. A considerable difference in AUC values of both sets of codons can be observed.

Codon importance estimated using the Boruta algorithm is summarized in Table 2.1. Codons that classify HEG and LEG are shown in shaded

Table 2.1: RSCU values in high and low expression genes in *E. coli*

AA	Codon	RSCU LEG	RSCU HEG	AA	Codon	RSCU LEG	RSCU HEG	AA	Codon	RSCU LEG	RSCU HEG	AA	Codon	RSCU LEG	RSCU HEG
Phe	UUU	1.10	0.69	Ser	**UCU**	0.81	1.79	Tyr	UAU	1.13	0.80	Cys	UGU	0.85	0.77
	UUC	0.90	1.31		UCC	0.92	1.44		**UAC**	0.87	1.20		UGC	1.15	1.23
Leu	UUA	0.67	0.30		UCA	0.63	0.40	TER	UAA	x	x	TER	UGA	x	x
	UUG	0.75	0.39		UCG	1.01	0.47		UAG	x	x	Trp	UGG	1.00	1.00
	CUU	0.61	0.42	Pro	CCU	0.59	0.49	His	CAU	1.12	0.73	Arg	**CGU**	2.48	3.64
	CUC	0.67	0.50		CUC	0.67	0.50		**CAC**	0.88	1.27		CGC	2.54	2.13
	CUA	0.19	0.07		CCA	0.76	0.63	Gln	CAA	0.66	0.45		CGA	0.31	0.08
	CUG	3.10	4.31		**CCG**	2.23	2.75		**CAG**	1.34	1.55		CGG	0.48	0.08
Ile	AUU	1.58	1.15	Thr	**ACU**	0.60	1.06	Asn	AAU	0.83	0.48	Ser	AGU	0.84	0.37
	AUC	1.27	1.82		ACC	1.82	2.10		**AAC**	1.17	1.52		AGC	1.78	1.53
	AUA	0.15	0.03		ACA	0.43	0.24	Lys	AAA	1.52	1.58		AGA	0.12	0.05
Met	AUG	1.00	1.00		ACG	1.14	0.61		AAG	0.48	0.42		AGG	0.07	0.02
Val	**GUU**	0.99	1.52	Ala	**GCU**	0.59	1.02	Asp	GAU	1.26	1.01	Gly	**GGU**	1.40	1.93
	GUC	0.86	0.57		GCC	1.08	0.73		**GAC**	0.74	0.99		GGC	1.68	1.71
	GUA	0.59	0.76		GCA	0.81	0.94	Glu	**GAA**	1.37	1.50		GGA	0.35	0.13
	GUG	1.56	1.15		GCG	1.52	1.30		GAG	0.63	0.50		GGG	0.58	0.23

Table presents RSCU values in two sets of *E. coli* genes (i) High Expressed Genes (HEG) (ii) Low Expression Genes (LEG). The list of genes considered in the two sets is given in Appendix A.1.2. In general, RSCU values are more variable among synonymous codons in HEG than LEG. For example, RSCU values of Leu codons vary between 0.07 and 4.31 in HEG, whereas the same values vary between 0.19 and 3.10 in LEG. The higher codon usage bias in HEG is attributed to facilitating faster translation. In contrast to this observation, certain codons are avoided in HEG. For example, RSCU value Gly codon GGG are 0.23 and 0.58 respectively in HEG and LEG sets. This lower RSCU value in HEG is attributed to the high probability of frameshift error in GGG codons. Similar is the case for Pro codon CCC. Lys and Cys amino acid codons are exceptions regarding the above observations, as codon usage is similar in both the gene sets. Codons contributing towards classifying HEG and LEG are highlighted with dark shade. Based on the codon importance estimated using Borutas algorithm, codons preferred (positively selected) and avoided (negatively selected) in HEG are shown in large and small fonts in the shaded boxes

boxes, and others are shown in white boxes. Among the codons in shaded boxes, codons used more frequently in HEG in comparison to LEG, termed as positively selected codons, are shown in large font. In contrast, codons used less frequently in HEG than LEG, termed as negatively selected codons, are displayed in small font. Codons in large font are known to be optimal codons Table 2.1 [180]; some of the codons known to be beneficial for optimum translation across bacteria species are CGU and GGU [178] and UUC, UAC, CAC and AAC [193]. In addition to these optimal codons, several other codons are being avoided in HEG, for example, CCC and GGG, possibly because these codons are prone to translational frameshift [140][141].

2.2.3.2 Codons with higher degeneracy are more important toward gene expression prediction

Codon usage bias is influenced by codon degeneracy [181]. To understand the relationship between codon importance values and codon degeneracy, we compared them. We considered all the amino acid codons of degeneracy >= 2 into two groups: (i) 2-fold degenerate codons (TFS) and 4-fold degenerate codons (FFS). Family box and split box codons of Ser, Arg and Leu were separated into FFS and TFS, respectively. Ile codons AUU and AUC were also considered in TFS. In general, codons with higher degeneracy were with higher importance values than codons with lower degeneracy. Accordingly, codon importance values for TFS were significantly different from those of FFS codons (p-value <0.05).

Further, we considered RSCU values of FFS and TFS into two separate datasets and employed RF and XGBoost classifiers to classify high and low expression genes. Both the classifiers were trained separately by dividing the dataset into 70%:30%, where 70% of data were used for training purposes and 30% for testing purposes. Once the classifier was trained, then the prediction of high and low expression genes was performed on the testing data for several iterations. Based on the confusion matrix values (Figure 2-5), the average accuracy, precision, recall, and F1 score values were calculated (Table 2.2). The accuracy score in terms of AUC and ROC curve is plotted in Figure 2-8. It can be observed that the accuracy, precision and F1 scores for FFS were higher than TFS according to both the classifiers. This result further supports our observation concerning the role of codon degeneracy.

Codon adaptation index (CAI) correlates with gene expression in *E. coli* significantly. When considering all the genes (MM section), Pearson r(CAI, gene expression) was 0.713. To understand the role of codon degeneracy, we segregated *E. coli* genes into two datasets: compared to average codon composition value, (i) genes richer in FFS and (ii) genes poorer in FFS. Pearson r(CAI, gene

Table 2.2: Confusion matrix matrices of Random Forest (RF) and XGBoost (XGB) Classifier

Metrices	Random Forest		XG Boost	
	FFS	TFS	FFS	TFS
Accuracy	0.837	0.789	0.827	0.792
Precision	0.936	0.760	0.854	0.790
Recall	0.733	0.813	0.797	0.773
F1 Score	0.822	0.786	0.823	0.781

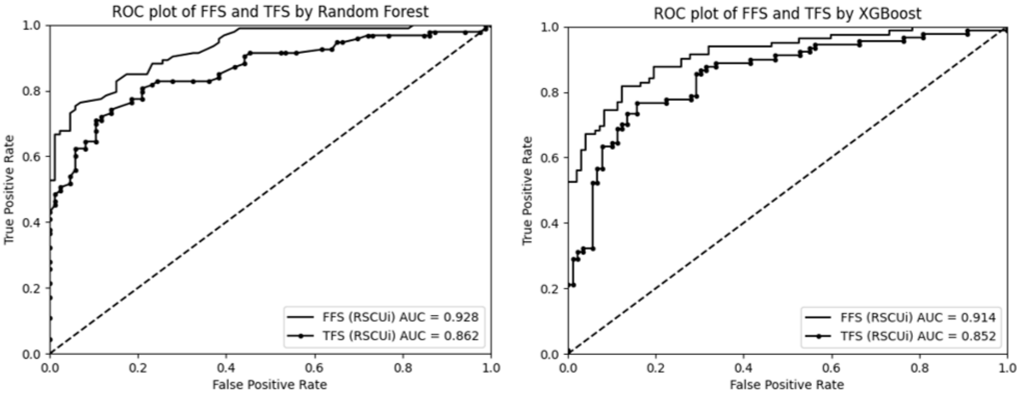


Figure 2-8: ROC curve of FFS and TFS features using RF and XGB

Figure presents ROC curves with AUC values generated based on TPR and FPR values. RF and XGBoost classifiers are employed to exhibit the difference in classifier performance in the two sets of FFS and TFS codons. Both the graphs show FFS with better accuracy.

expression) value was found to be 0.753 for gene set richer in FFS whereas, on the other hand, Pearson r(CAI, gene expression) value was found to be only 0.643 for the second set of genes (Figure 2-9). Further, we segregated *E. coli* genes into two datasets: compared to average codon composition value, (i) genes richer in TFS and (ii) genes poorer in TFS. In concordance with the above result, Pearson r(CAI, gene expression) values were found to be 0.640 and 0.751, respectively, for the two sets (Figure 2-9).

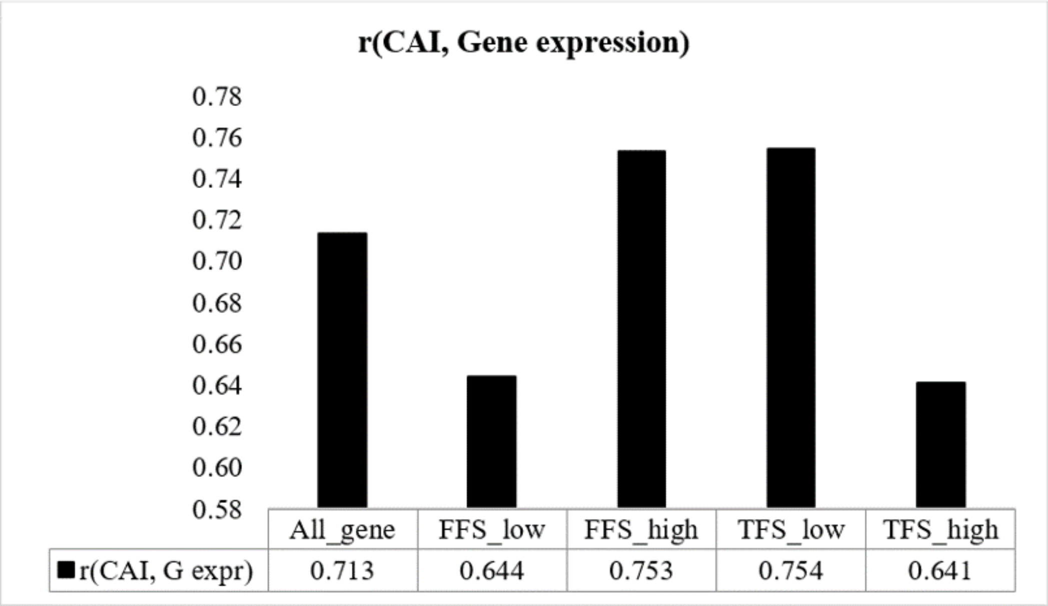


Figure 2-9: Correlation of CAI and gene expression based on the composition of FFS and TFS in *E. coli*

Figure presents correlation of CAI and gene expression based on the high or low composition of FFS codons and TFS codons. A better correlation is observed in case of high FFS and low TFS in *E. coli*

2.2.3.3 Codon importance features of 683 bacteria species

We further extended our study to find Boruta algorithm-based codon importance values in genomes of other bacteria. For this study we considered genomes of 683 bacteria which are summarized in Table 2.3. These bacteria belong to twenty nine phylogenetic groups and with coding region (G+C)% between 23.65 and 74.68. The estimated codon importance values among the bacteria found to have both similarities and differences. The importance values for the two bacteria with extreme (G+C)%, *Wigglesworthia glossinidia* with (G+C)% equal to 23.65 and *Anaeromyxobacter dehalogenans* with (G+C)% equal to 74.68 are given for example in the Figure 2-10 and Figure 2-11. In both the bacteria, several codons such as UUC(Phe), AUC(Ile), GGU(Gly), GGG(Gly) and CUG(Leu) were found to be important and some other codons such as AUA(Ile) not important towards classifying high and low expression genes. Considering codon importance values in individual bacterium, we have summarized the result for all the 683 bacteria in the following section. The codons selected as important codons (marked as 'C') for each organism were presented in Appendix A.1.4.

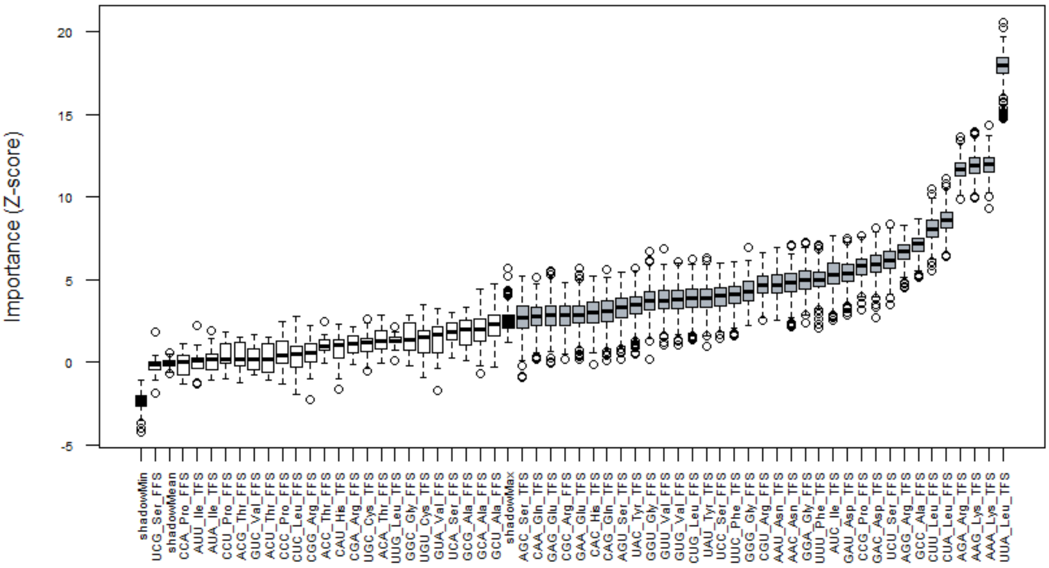


Figure 2-10: Importance of the codons in *Wigglesworthia glossinidia* genome

A frequency distribution graph of the importance value of all 683 organisms is presented in Figure 2-12. Here, a high importance value indicates a substantial difference in usage of the corresponding codon in HEG and LEG. A low importance value indicates almost equal codon usage in both HEG and LEG. We can observe a significant variation in the selection of these codons as important to classify HEG and LEG. Most prominently, the two cysteine codons have been equally used in HEG and LEG across the majority of the organisms. The ATA codon of Ile has been selected as most important by 40 organisms, indicating a

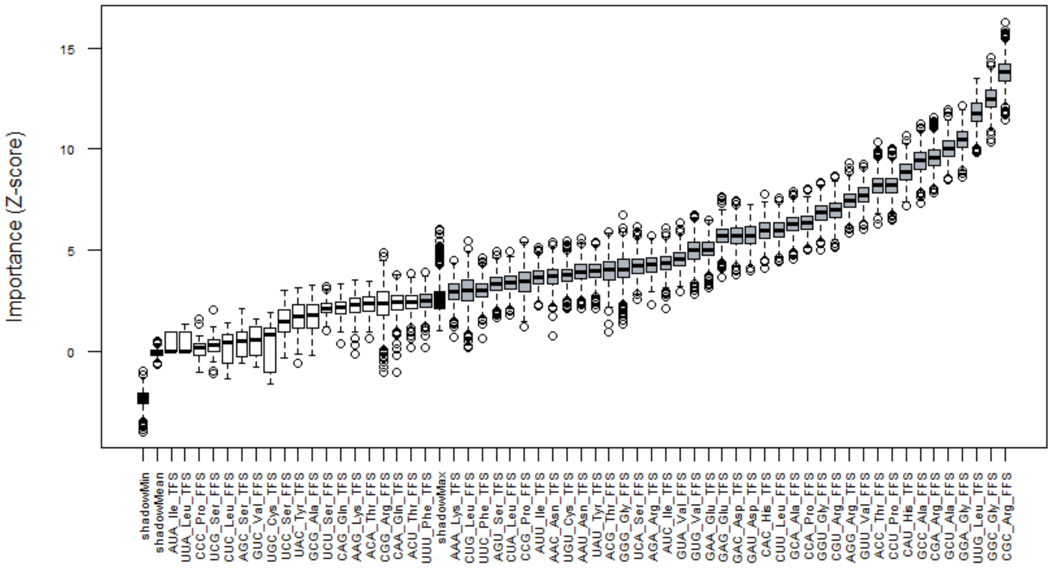


Figure 2-11: Importance of the codons in *Anaeromyxobacter dehalogenans* genome

high usage difference of this codon in HEG and LEG. Codons like CCC of Pro and GGG of Gly have been selected by most organisms with importance value above 5. This observation indicates a high usage difference in HEG and LEG, most probably due to frameshift mutation. Similarly, CTG of Leu, CGA of Arg, GCC of Ala, ACC of Thr, ATC of Ile, and ATA of Ile have been selected as important codons by most organisms. We further calculated the RSCU values of HEG and LEG of all those organisms whose importance value of Leu CTG codon is more than 5. We observed that Leu CTG codon is used extensively in HEG and is thus positively selected. Arg CGA codon is also positively selected. In the case of Thr ACC codon, it is also extensively used in HEG, the codon is positively selected. Whereas Ile ATC is primarily used in LEG, and thus the codon is negatively selected. Similar is the case with Ile ATA, Gly GGG, and Pro CCC and is negatively selected.

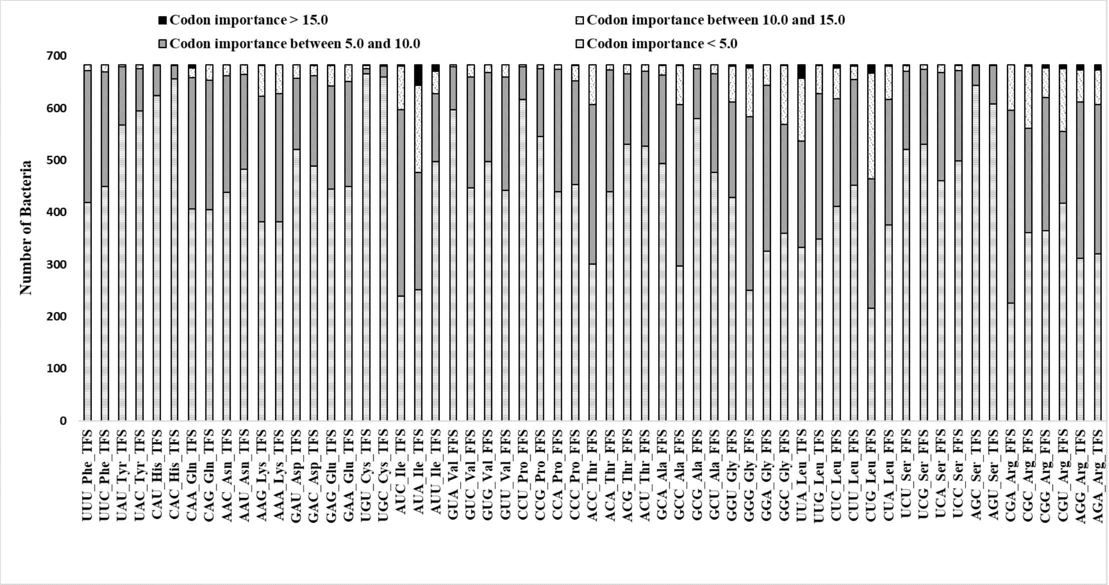


Figure 2-12: Frequency distribution of importance value in 683 organisms
Figure presents the frequency distribution of the mean of importance values of 683 organisms. The X-axis represents 59 codons arranged degeneracy-wise. The y-axis shows the number of bacteria. The grey bar presents the number of organisms that has mean importance value ranging from 5 to 10. In general, there is a wide variation in the codons with respect to individual organisms.

2.3 Improved Implementation of CAI

2.3.1 Limitations in existing implementation of CAI

Implementation of CAI is available in several software such as CodonW [145], INCA [209], CAIcal [155], EMBOSS [167], CAI Calculator 2 [239] and DAMBE7 [241]. However, there are several difficulties in using existing software for calculating CAI. One of the crucial step in calculating CAI is to select the reference set of high expression genes and then to calculate the relative adaptedness of the codons, i.e the w_k values as given in equation 1.3. The above methods have employed various approaches to do so, and have their own limitations. In CAIcal, there is dependency of reference database tables. If the organism for which CAI value is to be calculated is not present in the database, then the user have to generate the reference database table. This approach is not user friendly for providing information about codon usage in reference set of high expression genes. It has been reported by Xia [240] that the EMBOSS [167] and CAI Calculator 2 [239] software provide erroneous result, might be because of some implementation issues.

These organism specific w_k values are available in existing softwares only for few organisms. Therefore, users can directly calculate CAI values for the genes of only those organisms. For calculating CAI for the genes of an organism, CodonW [145] provides alternative indirect approach using correspondence analysis [61].

2.3.1.1 Considering default *E. coli* reference set may generate erroneous result

High expression gene set of *E. coli* is also suggested to be used as a default set for calculating CAI. *E. coli* is an organism with strong selection on codon usage [57] and (G+C)% around 50.0. However (G+C)% of the organism varies widely among bacteria from as less as around 17.0% to more than 75.0% [159]. Furthermore, though the selected codon usage bias is universal among organisms [210], it varies from organism to organism [194][179] and also differs among bacteria phylogeny [180]. While selection on codon usage is very strong in *E. coli*, *Bacillus subtilis* and *Saccharomyces cerevisiae* (yeast), it is very low in several other organisms [179]. Therefore calculating CAI considering default *E. coli* reference set can generate erroneous result.

Alternatively, if the user is familiar with any high level computer programming languages, they may calculate w_k values from high expression gene sequences separately and input the values to CodonW to calculate CAI. These approaches may be complicated for naive users and not suitable for researchers unfamiliar with high level programming languages. Keeping these constraints in calculating CAI in the command driven CodonW and other software in view, we developed a web portal that provides online tool for calculating CAI.

2.3.2 Improved Codon Adaptation Index web portal

2.3.2.1 Reference set of high expression genes available in the web portal

We downloaded the bacterial genomes from the NCBI site[154]. Then,we extracted set of gene known to be highly expressed and widely conserved across organism [192] from these genomes using python scripts and made available in our portal as the reference gene sets. Ribosomal protein genes, outer membrane protein genes such as *rplA, rpmB, rpsA*, etc, elongation factor genes such as *tufA, tufB, fusA* etc, regulatory/repressor genes such as *dnaG, araC* etc are some of the example genes considered in the high expression gene sets. At present we have provided high expression gene sets for 684 unique species of bacteria in our database. These bacteria belong to 29 different phylogenetic groups and with coding region (G+C)% between 23.65 and 74.68 as shown in Table 2.3. High expression gene sets for *E. coli*[85], *Saccharomyces cerevisiae*(yeast)[54] and *Homo sapiens*(human)[149][176] available in our portal are based on the experimental expression data.

Server configuration and language used for the web portal

Our web portal is launched in an IBM System x3630 M4 server with CentOS 6.10 operating system. The web portal is developed using Python programming language.

Description of how to use our web portal

Keeping limitations of the available softwares and lack of reference set of high expression genes for large number of organisms in view, we envisaged this web portal. It is designed to simplify the computation. It is very simple to use and accessible in Internet from any computer. It is designed not to have any limitation on the input genome sequences length. The user interface of the portal provides a two-step process to calculate CAI.

Table 2.3: Details of bacteria whose reference set of high expression genes available in the web portal

Sl No.	Bacterial Group	No. of Organisms	Maximum (G+C)%	Minimum (G+C)%
1	Acidobacteria	1	61.1	61.1
2	Actinobacteria	80	74.5	46.31
3	Alphaproteobacteria	86	72.09	30.37
4	Aquificae	7	52.24	32.03
5	Bacteroidetes	35	66.73	27.25
6	Betaproteobacteria	59	70.36	37.78
7	Chlamydiae	10	44.31	36.1
8	Chlorobi	8	57.66	45.06
9	Chloroflexi	5	60.94	47.85
10	Cyanobacteria	10	62.86	40.4
11	Deferribacteres	3	43.2	31.08
12	Deinococcus-Thermus	11	70.23	63.01
13	Deltaproteobacteria	28	74.68	37.48
14	Dictyoglomi	2	33.99	33.81
15	Elusimicrobia	1	40.69	40.69
16	Epsilonproteobacteria	16	44.89	27.19
17	Fibrobacteres	1	48.89	48.89
18	Firmicutes	124	69.29	28.34
19	Fusobacteria	3	34.69	26.2
20	Gammaproteobacteria	137	70.46	23.65
21	Gemmatimonadetes	1	64.49	64.49
22	Nitrospirae	1	34.16	34.16
23	Planctomycetes	2	57.91	55.46
24	Spirochaetes	17	52.08	27.7
25	Synergistetes	2	64.45	45.75
26	Tenericutes	18	40.66	23.96
27	Thermodesulfobacteria	2	42.61	30.67
28	Thermotogae	11	47.08	30.73
29	Verrucomicrobia	3	65.47	45.85

The first step is to input the nucleotide sequence of the genes whose CAI values are to be calculated in a single file in fasta format.

The second step of the calculation is to provide additional input file of reference set of high expression genes. The web portal provides three simple options for input of this reference set.

- The user can input the reference set of high expression genes in the form of a fasta file.

- At present the web interface provides a list of 684 bacteria species whose high expression gene set is available in our database. User can select the name of the organism from this list corresponding to the organism whose gene sequences were uploaded in the first step.

- In the third alternative, the web interface shows the gene informations from the uploaded file in the first step. User can select multiple genes know to be highly expressed from the displayed list. Those selected genes will be considered as the reference set while calculating CAI.

Integrity checks: Before processing, the web portal examines the accuracy of the input sequences. These include presence of internal stop codons, presence of accepted start (i.e. NTG, ATN) [142] and stop codons, presence of non-IUPAC characters. If any of these problems are found, CAI is calculated for those sequences with potential errors with appropriate warning messages. Therefore, sequences that generate warnings should be carefully checked.

Based on these, output file is generated. Along with the CAI values, the output file also contains additional information about the genes such as length in terms of number of amino acids, G+C%. Once the result is produced, no input sequence files are retained in the server to avoid any possible misuse of the users data.

2.4 Conclusion

Codon usage bias is an important genomic feature extensively used for understanding evolution at molecular level. Machine learning based study in this manuscript demonstrates a commonality among bacteria regarding behaviour of certain codons with regard to gene expression. Initial study in *E. coli* genome distinctly identified the positively and negatively selected codons for optimum translation and the codons least influenced by gene expression. Codon adaptation index (CAI) was used to predict high and low expression genes among 683 bacterial species. The machine learning based analysis prominently identified certain codons being influenced by gene expressed across a majority of these bacterial species. Further, a higher proportion of 4-fold and 6-fold degenerate codons than the 2-fold degenerate codons were observed to be influenced by the gene expression in these bacteria. Some of the codons being least influenced by gene expression across the bacterial species is an interesting codon usage feature to be investigated in coming future.

The web portal to calculate CAI is freely available for academic and research purpose in our web server at http://14.139.219.242:8003/cai. At present the web portal can be used for calculating CAI as per universal genetic code table. Future scope lies for consideration of other available versions of genetic code. We believe our web portal will be helpful for biologists working on molecular evolution.