Introduction to Codon Usage Bias

During the translation process from mRNAs to proteins, information is transmitted in the form of triple nucleotides named codons encode. Amino acids are degenerate, have more than one codon to be encoded except for methionine (Met) and tryptophan (Trp), thus codons encode one amino acid are known as synonymous codons. Many studies on different organisms reported that synonymous codons are not used uniformly within and between genes of one genome, this phenomenon called `synonymous codon usage bias (SCUB) or Codon usage bias (CUB) [1–4]. Further, the degree of the unequal use of synonymous codons differ between species [5,6]. Hence each organism has its optimal codons, an optimal codon is defined as the codon which is more frequently used in highly expressed genes than in the low expressed genes [7]. Two main factors shaping the codon usage of an organism are mutation and selection [8–10] , other factors also described to influence CUB of an organism as nucleotide composition [11], synonymous substitution rate [12], tRNA abundance [13], codon hydropathy and DNA replication initiation site [14], gene length [15] and expression level [16]. Investigating the CUB of an organism could tell about genes molecular evolution, expression, and host-pathogen coadaptation. Further in biotechnology applications CUB and predicted optimal codons could help to design highly expressed genes and constructing suitable cloning vectors [17,18].

Equations used for codon usage bias analysis

Effective number of codons

The The effective number of codons (ENc) measures the bias of using a smaller subset of codons apart from the equal use of synonymous codons. In addition, the ENc measures the codon usage unbalance among genes, where an amino acid is encoded by one codon in a gene would be biased and negatively correlated with the ENc value. this measureENc ranges from 20-61 with higher values indicating more codons being used for each amino acid i.e less bias and vice versa, ENcthis measures codon bias irrespective to gene length, it can be an indicator of codon usage with reference to mutational bias [19]. ENc was calculated using the equation given by [20]:

$$F_{CF} = \sum_{i=1}^{m} \left(\frac{n_i + 1}{n + m}\right)^2$$

Then ENc could be calculated by:

$$N_{c.CF} = \frac{1}{F_{CF}}$$

Where n_i is the count of codon i in m amino acid family and m is the number of codons in an amino acid family

Codon adaptation index

Codon adaptation index (CAI) uses a reference set of highly expressed genes (e.g. ribosomal genes), this measure is an indicator of gene expression levels and natural selection; it ranges from 0 to 1 with higher values indicating stronger bias with respect to the reference set, therefore this method is an indicator of selection for a bias toward translational efficiency [21,22]. Codon adaptation index (CAI) was calculated by the equation given by [23][24]:

$$CAI = exp\frac{1}{L}\sum_{k=1}^{L} \ln w_{c(k)}$$

Where, L is the count of codons in the gene and *wc*(k) is the relative adaptiveness value for the *k*-th codon in the gene.

Relative synonymous codon usage

Relative synonymous codon usage (RSCU) calculate the observed count to the expected count of each codon to be analyzed, as RSCU less than one means codon observed less frequently than the average codon usage, while RSCU more than one means codon is observed more frequently than the average codon usage[21]. The equation to calculate the RSCU is [21]:

$$RSCU = \frac{O_{ac}}{\frac{1}{k_a}\sum_{c \in C_a} O_{ac}}$$

Where Oac is the count of codon c for amino acid and ka is the number of synonymous codons.

Translational selection index

Translational selection (P2) index measure the bias of anticodon-codon interactions, from which we can indicate the translation efficiency[25]. Generally P2-index range from 0 to 1 also Also its values have been noted to be high for highly expressed genes and low for lowly expressed genes[7,26]. i [25]By taken the averages of numbers resulted from this equation for each CDS:

$$P2 = \frac{WWC + SSU}{WWY + SSY}$$

Where, W = A or U(T), S = G or C, and Y = C or U(T).

Hydropathicity (Gravy) and Aromaticity (Aroma) indices

In analyzing the natural selection for shaping the codon usage bias, two indices, including Gravy and Aroma scores, were used in many studies [27–29]. Thus, the variation of the two indices reflects the amino acid usage. A higher Gravy or Aroma value suggests a more hydrophobic or aromatic amino acid product.

Neutrality Plot

average

Parity Rule 2 -plot Analysis

All the nucleotides content at the third codon position (A3, T3, G3, and C3) were calculated, then for each gene AT-bias (A3/(A3 + T3)) and GC-bias (G3/(G3 + C3)) were estimated and used as the ordinate and the abscissa respectively in the plot, with both coordinates equal to 0.5, where A = T and G = C[32]. The genes positions on the plot along the ordinate and the abscissa tell about factors influence the CUB. If genes over the plot view are scattered equally, then the CUB is likely to be solely caused by the mutation[33].

Establish reference genes set.

Genes with high expression level within the organims genomes are ussually named Reference genes set. Reference genes set are used to calculate the CAI, in BCAWT two options may be used;

1- Reference genes set given by the users.

2- An auto option where BCAWT generates a genes reference set using 10% of genes have the lowest ENc values (highest biased genes).

Determination of putative optimal codons

BCAWT use the correlation method described here [20] to determine the putative optimal codons. Where each synonymous codon RSCU in one amino acid family correlated with all genes ENc, and optimal codon for each amino acid family was defined as the codon which has the strongest negative correlation RSCU with ENc values, and with a significant p-value less than 0.05/n and n is equal to the number of synonymous codons in such amino acid family.

Correspondence analysis

excluding Met and Trp codons, it is an advantage to perform multivariate statistical analysis on the rest of 59 codons to examine the variations in the codon usage bias among all the CDS. One way to do that is correspondence analysis (COA)[34,35], by plotting group of genes on continuous axes in multidimensional space according to the trends affecting the synonymous codon usage within the genes group.