



# BCAW(BIO CODON ANALYSIS WORKFLOW)

# AGENDA

Introduction to Codon Usage Bias

Factors that shape & influence the CUB

Usage of CUB

Equations used for codon usage bias analysis

Dependencies

# INTRODUCTION TO CODON USAGE BIAS

During the translation process from mRNA to protein, information is transmitted in the form of nucleotide triplets named codons. Amino acids are degenerate, having more than one codon representing each except for methionine (Met) and tryptophan (Trp). Thus, codons encoding the same amino acid are known as synonymous codons

Many studies on different organisms show 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). Further, the degree of the unequal use of synonymous codons differs between species. Hence, each organism has its optimal codons, where an optimal codon is defined as a codon which is more frequently used in highly expressed genes than in the lowly expressed genes.

# FACTORS THAT SHAPE & INFLUENCE THE CUB

---

- ❑ Two main factors shaping the codon usage of an organism are mutation and selection
- ❑ Other factors also known to influence the CUB of an organism are nucleotide composition, synonymous substitution rate, tRNA abundance, codon hydropathy and DNA replication initiation sites, gene length and expression level

## Usage of CUB

Investigating the CUB of an organism can tell about genes' molecular evolution, expression, and host-pathogen coadaptation. Furthermore, in biotechnology applications, CUB and predicted optimal codons could help to design highly expressed genes and construct suitable cloning vectors



# EQUATIONS USED FOR CODON USAGE BIAS ANALYSIS

## ❑ Effective number of codons

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 measure ranges from 20-61 with higher values indicating more codons being used for each amino acid, i.e less bias and vice versa. This measures codon bias irrespective of gene length and can be an indicator of codon usage with reference to mutational bias.

## ❑ 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.

## ❑ 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.

## ❑ Translational selection index

Translational selection (P2) index measure the bias of anticodon-codon interactions, from which we can indicate the translation efficiency [24]. Generally, P2-index ranges from 0 to 1. Its values have been noted to be high for highly expressed genes and low for lowly expressed genes.

## ❑ 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 .

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.

## ❑ 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  $G = C$ . 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.

## ❑ Establish reference genes set

Genes with high expression level within the organism genomes are usually named reference gene sets. Reference gene sets are used to calculate the **CAI**. In BCAWT two options may be used:

- Reference genes set given by the users.
- 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 uses the correlation method described in to determine the putative optimal codons. Where each synonymous codon RSCU in one amino acid family correlated with all genes ENc, an 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  $0.05/n$  where  $n$  is equal to the number of synonymous codons in such amino acid family.

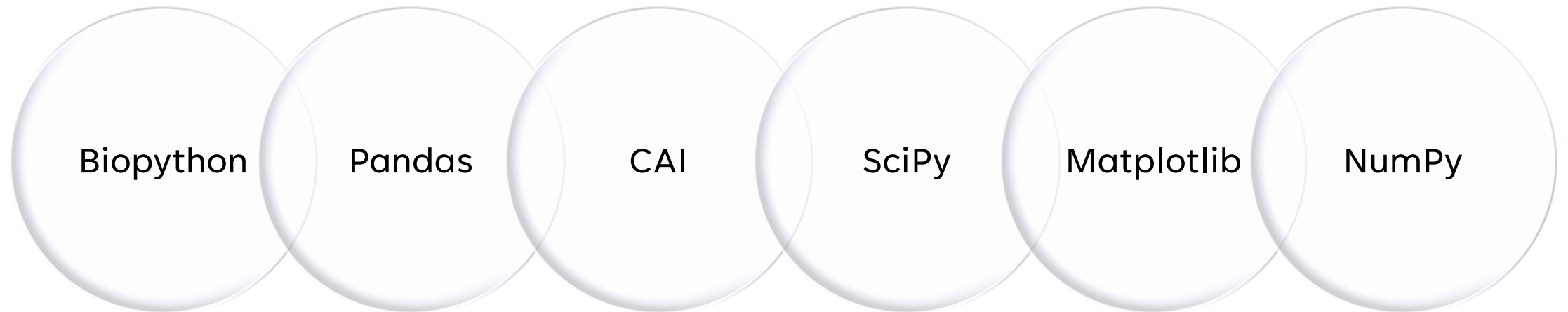
## ❑ Codon adaptation index

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 CDSs. One way to do that is correspondence analysis (COA by plotting groups of genes on continuous axes in multidimensional space according to the trends affecting the synonymous codon usage within the genes group.



DEPENDENCIES

# DEPENDENCIES



A series of white, overlapping geometric lines and polygons on a black background, located on the left side of the slide.

# THANK YOU

Huda Tariq AbdelKhalek

Mennatullah Mahmoud Mohamed

Hager Tariq Ahmed