Index Definition of GenRCA Rare Codon Analysis Tool

Version: 2.0.1

Update: September 15, 2025

If you find our tool useful, please cite our paper using the following:

Fan, K., Li, Y., Chen, Z. et al. GenRCA: a user-friendly rare codon analysis tool for comprehensive evaluation of codon usage preferences based on coding sequences in genomes. *BMC Bioinformatics* 25, 309 (2024). https://doi.org/10.1186/s12859-024-05934-z

Content

Commonly Used Index Descriptors

Indices based on non-uniform usage of synonymous codon

Indices in this section calculate the deviation of codon usage frequency from a "uniform" or expected "background" distribution. An increased CUB observed in a certain gene or genomic region suggests that selection has acted upon it, favoring certain codons that impact expression levels. Generally, these indices demonstrate a consistent relationship with the level of codon usage uniformity or frequency. At the extremes, these methods will identify cases where only one codon is used (maximal bias) or where all codons are utilized with equal frequency (minimum bias). It is worth noting that these indices are based solely on the coding sequence, thus offering an informative measurement of CUB without the need for prior knowledge.

Relative Synonymous Codon Use (RSCU)

The Relative Synonymous Codon Usage (RSCU)[1] value for a codon represents the ratio between the observed frequency of a particular codon and the expected frequency, assuming equal usage of all synonymous codons for the corresponding amino acid. The formula for calculating RSCU is as follows:

$$RSCU_{ij} = \frac{x_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}}$$

where x_{ij} denotes the number of occurrences of the j-th codon for the i-th amino acid, while n_i represents the degeneracy for the i-th amino acid. It is worth noting that if the codon is missing from the gene, we consider the RSCU value of this codon as 0.0001. The RSCU value of 1 indicates the average synonymous codon usage. A codon with an RSCU value greater than 1 is considered to be preferred or overrepresented, as it is used more frequently than expected. Conversely, a codon with an RSCU value less than 1 is considered to be unpreferred or underrepresented, as it is used less frequently than expected based on the assumption of equal usage of all synonymous codons for the corresponding amino acid. Therefore, the RSCU value is a useful tool for identifying codon bias and for gaining insight into the regulation of gene expression. Furthermore, the RSCU values also serve as weights for codons in various indices that require normalization of codon count into codon frequency and elimination of dependency on gene length, such as Codon Adaptation Index (CAI), Intrinsic Codon Deviation Index (ICDI), etc.

The RSCU of a gene on our website can be computed by taking the arithmetic mean of all codons. The calculation can be performed as follows:

$$RSCU(g) = \frac{\sum_{k=1}^{L} rscu_k}{L}$$

where L is the length of a gene measured in codons and $rscu_k$ is the RSCU value for the k-th codon in the gene.

Effective Number of Codons (ENC)

The mathematical formula utilized in the calculation of the Effective Number of Codons (ENC)[2,3] is based on principles derived from the field of population genetics. ENC takes into account the degeneracy level of amino acids and computes the total number of distinct codons used in a sequence. This index can assess the extent of bias or preference in the usage of synonymous codons within a particular gene or genome.

The theoretical value of ENC ranges from 61 (when there is no bias and all synonymous codons are used uniformly) to 20 (when there is maximal bias or preference observed in synonymous codon usage). A lower ENC value indicates a stronger inclination towards specific codons for a particular amino acid, while a higher ENC value suggests a more balanced or uniform utilization of synonymous codons.

For an amino acid (AA) with a degeneracy of k, meaning it has k synonymous codons, each with counts $n_1, n_2, \ldots, n_k$, $n = \sum_{i=1}^{k} n_i$ and $p_i = n_i/n$, the effective number of codons for each amino acid ($\widehat{N}_{cAA}$) is calculate as follows:

$$\widehat{N}_{cAA} = \frac{1}{F_{AA}}$$

where

$$F_{AA} = \frac{n\sum_{i=1}^{k} {p_i}^2 - 1}{(n-1)} \quad n > 1$$

It is important to mention that if there is a maximum bias in the usage of codons for an amino acid, meaning that only one synonymous codon is used, the F_{AA} for that amino acid is 1.

Finally, for the standard genetic code, the formula for calculating ENC ($\widehat{N}_c$) for a gene is as follows:

$$\widehat{N}_c = 2 + \frac{9}{\bar{F}_2} + \frac{1}{F_3} + \frac{5}{\bar{F}_4} + \frac{3}{\bar{F}_6}$$

where $\bar{F}_i$ (i = 2, 3, 4 and 6) represents the average values of F_{AA} for all amino acids with degeneracy i. It is worth noting that rarely used amino acids (indicated by a F_{AA} value of 0) should be regarded as absent and are not taken into consideration when calculating the averages. However, if isoleucine (Ile) is absent or rarely used (F_3 = 0), the value of F_3 should be calculated as the average of $\bar{F}_2$ and $\bar{F}_4$.

Relative Codon Bias Strength (RCBS)

The Relative Codon Bias (RCB)[4] is a metric used to assess the level of bias in codon usage. It is calculated by comparing the observed frequency of a specific codon to the expected frequency assuming random codon usage, taking into account the biased base composition at three sites as observed in the sequence being studied. This difference is then divided by the expected frequency, providing a normalized measure of codon usage bias that is independent of overall base composition variability. The RCB of the codon xyz is calculated as follows:

$$d_{xyz} = \frac{f(x,y,z) - f_1(x)\cdot f_2(y)\cdot f_3(z)}{f_1(x)\cdot f_2(y)\cdot f_3(z)}$$

Specifically, $f(x,y,z)$ represents the observed frequency of codon xyz, where x, y, and z denote the first, second, and third nucleotides of that codon, respectively. Additionally, $f_1(x)$, $f_2(y)$ and $f_3(z)$ represent the observed frequencies of the individual bases x, y, and z at positions 1, 2, and 3 of the codons. Rare codons are assigned lower d_{xyz} values, typically approaching -1. Conversely,

highly frequent codons are assigned higher d_{xyz} values, which can reach a value of 1.

The RCBS (Relative Codon Bias Strength)[4] is designed to consider base compositional bias, enabling a more robust estimation of highly favored codon frequencies while accounting for other characteristics of the coding sequence, including GC content bias. It provides a quantitative evaluation of the deviation from expected codon frequencies, taking into account the underlying RCB factors observed in the gene's sequence. To calculate the RCBS for a gene with L codons, the following formula is used:

$$RCBS = \left(\prod_{i=1}^{L}\left(1 + d_{xyz}^{i}\right)\right)^{1/L} - 1 = \exp\frac{1}{L}\sum_{i=1}^{L}\left(1 + d_{xyz}^{i}\right) - 1$$

A value of RCBS close to 0 suggests a lack of bias in codon usage. Conversely, a value greater than 0.5 indicates a favorable preference for specific codon usage, which signifies high gene expression.

Directional Codon Bias Score (DCBS)

According to the RCBS (Relative Codon Bias Strength)[4] formula, rare codons have negative RCB values, approaching -1, while very frequent codons have positive RCB values, approaching 1. As a result, the presence of very rare codons decreases the final RCBS score of a gene, while the presence of very frequent codons increases it. However, genes with a high Codon Usage Bias (CUB) should include both very frequent and very rare codons. To ensure that both positive and negative codon usage biases contribute to the same direction and increase the RCBS score, a modified version called The Directional Codon Bias Score (DCBS)[5] has been proposed.

DCBS considers both overrepresented and underrepresented codons, providing a more accurate evaluation of codon bias in a gene by considering the directionality of codon usage bias. To calculate DCBS, the Directional Codon Bias (DCB) of a codon triplet xyz is defined as follows:

$$d_{xyz} = max\left(\frac{f(x,y,z)}{f_1(x)\cdot f_2(y)\cdot f_3(z)}, \frac{f_1(x)\cdot f_2(y)\cdot f_3(z)}{f(x,y,z)}\right)$$

Identical to RCBS, $f(x,y,z)$ represents the observed frequency of codon xyz, and $f_1(x)$, $f_2(y)$ and $f_3(z)$ represent the observed frequencies of the individual bases x, y, and z at positions 1, 2, and 3 of the codons.

The DCBS of a gene, which is composed of L codons, is calculated as follows:

$$\mathrm{DCBS} = \frac{\sum_{i=1}^{L} d_{xyz}}{L}$$

Similar to RCBS, a value close to 0 for DCBS indicates a lower degree of codon preference, while a higher value indicates a stronger bias in codon usage.

Codon Deviation Coefficient (CDC)

The Codon Deviation Coefficient (CDC)[6] is based on the cosine distance metric between the expected and the observed codon usage. This metric takes into account background nucleotide compositions (BNC) specific to codon positions, thereby allowing for a more precise evaluation of CUB in diverse genetic sequences.

In this context, the four nucleotides (adenine, thymine, guanine, and cytosine) are denoted as A, T, G, and C, while the GC content and purine content are represented as S and R, respectively. The expected

nucleotide contents (A, T, G, C) at codon position i (i = 1, 2, 3) is derived based on the observed positional GC and purine contents, which can be formulated as follows:

$$A_i = (1 - S_i)R_i$$
$$T_i = (1 - S_i)(1 - R_i)$$
$$G_i = S_i R_i$$
$$C_i = S_i(1 - R_i)$$

For any sense codon $xyz(x, y, z \in \mathrm{A, T, G, C})$, the expected usage πxyz is defined as the product of its constituent expected nucleotide contents $x_1 y_2 z_3$, normalized by the sum over all sense codons:

$$\pi xyz = \frac{x_1 y_2 z_3}{\sum_{abc} w_{abc} a_1 b_2 c_3}$$

where

$$w_{abc} = \begin{cases} 1, & if\ abc\ is\ a\ sense\ codon \\ 0, & otherwise \end{cases} \quad and\ a, b, c \in A, T, G, C$$

Also, the observed usage of the sense codon xyz is normalized by the length of gene:

$$\hat{\pi} xyz = \frac{F_{xyz}}{L}$$

where F_{xyz} represent the observed frequency of that codon in the gene, and L is the length of a gene measured in codons.

Then, CDC is calculated using the cosine distance metric, which measures the similarity between the expected (π) and observed ($\hat{\pi}$) codon usage patterns:

$$\mathrm{CDC} = 1 - \frac{\sum_{xyz} \pi xyz \times \hat{\pi} xyz}{\sqrt{\sum_{xyz} \pi xyz^2 \times \sum_{xyz} \hat{\pi} xyz^2}}$$

The CDC ranges from 0 to 1, where a value closer to 0 indicates a closer correspondence between the expected and observed codon usage patterns, suggesting a lower level of bias in codon usage. Conversely, a CDC value closer to 1 represents the greater level of codon usage bias.

Measure Independent of Length and Composition (MILC)

The Measure Independent of Length and Composition (MILC)[7,8] quantifies the variation in codon usage by calculating a log-likelihood ratio between the expected and observed counts of codons. MILC is a versatile measure that remains unaffected by alterations in gene length and overall nucleotide composition, and introducing little noise into measurements. This approach yields similar numerical results to the widely used χ^2 test, but may offer theoretical advantages in statistical analyses. Only sequences consisting of 80 codons or more are recommended to be analyzed using MILC. A higher value indicates a stronger bias in codon usage, which signifies higher gene expression.

The individual contribution (M_a) of each amino acid a to the MILC statistic is calculated as follows:

$$M_a = \sum_{c \in a} O_c In \frac{O_c}{E_c} = \sum_{c \in a} O_c In \frac{f_c}{g_c}$$

where O_c represents the observed count of codon c in a gene, E_c denotes the expected count of the same codon, f_c signifies the frequency of codon c in a gene, and g_c represents the expected frequency of the same codon. It is important to note that the sum of f_c or g_c over all codons for each amino acid should equal 1. In addition, the O_c/E_c ratio is mathematically equal to, and can be

replaced by f_c/g_c.

Then, MILC can be calculated by the following formula:

$$MILC = \frac{\sum_a M_a}{L} - C$$

where L represents the gene length measured in codons, aiming to account for the expected increase associated with the total number of codons. The correction factor C is employed to address the issue of overestimating the overall bias in shorter sequences. In the case where the codon usage in the gene aligns with the expected distribution and all amino acids are present, the occurrence of sampling errors leads to an increase in the χ^2 score by 41, and an increase in the 'scaled' χ^2 by $41/L$. Consequently, the correction factor C can be computed using the following formula:

$$C = \frac{\sum_a (r_a - 1)}{L} - 0.5$$

where r_a represents the number of possible codons for the amino acid a, which corresponds to its degeneracy class. Only the amino acids that are actually present at least once in the sequence contribute to the calculation of C. For example, if a gene lacks one of the amino acids with a four-fold degeneracy, C would be calculated as $38/L$. In cases where the observed frequencies closely match the expected codon distribution, MILC values may become negative. To compensate for this, a constant of 0.5 is subtracted to the correction factor C.

Intrinsic Codon Deviation Index (ICDI)

The Intrinsic Codon Deviation Index (ICDI)[9] utilizes the RSCU[1] and the degeneracy of amino acids in the sequence, giving equal weight to all included amino acids. It can be applied to evaluate codon bias in genes from species without knowledge of optimal codons, and has the potential to help predict gene functionality. A gene with strong bias, using only one codon per amino acid, would have an ICDI value of 1, while a gene utilizing all codons equally would have a value of 0.

The ICDI was calculated through the following steps. Firstly, for each of the 18 amino acids (methionine and tryptophan excluded) with k number of synonymous codons, the value of S_k is defined as follows:

$$S_k = \sum \frac{(rscu_i - 1)^2}{k(k-1)}$$

where $rscu_i$ is the RSCU[1] value of the i-th codon. It is worth noting that if the amino acid is absent in the gene sequence, the S value for that particular amino acid is considered as 0.0001. For the standard genetic code, the value of k corresponds to the degeneracy of the codons, which can be 2, 3, 4, or 6 triplets encoding for the same amino acid. Then the ICDI value for a gene is calculated as follows:

$$ICDI = \frac{\sum S_2 + S_3 + \sum S_4 + \sum S_6}{18}$$

Synonymous Codon Usage Order (SCUO)

The Synonymous Codon Usage Order (SCUO)[10,11] is based on an informatics method, using Shannon informational theory and maximum entropy theory to provide an estimate for the orderliness of synonymous codon usage. Entropy is a measure of uncertainty or information content in a probabilistic system. Shannon entropy quantifies the average amount of information needed to describe or predict the outcomes of a random variable. In the context of information theory, entropy represents the amount of "surprise" or "new information" associated with the occurrence of an event.

According to the information theory, the uncertainty in synonymous codon usage of the i-th amino acid can be quantified using a function known as the 'entropy' of the probability distribution:

$$H_i = -\sum_{j=1}^{k_i} p_{ij} log p_{ij}$$

where H_i represents the entropy of the i-th amino acid and k_i is its degree of codon degeneracy. Moreover, p_{ij} is the normalized frequency of j-th codon usage for the i-th amino acid, which can be calculated as follows:

$$p_{ij} = \frac{x_{ij}}{\sum_{j=1}^{n_i} x_{ij}}$$

It is noteworthy the case that the synonymous codons are chosen at random, without any bias or preference, each codon has an equal probability of occurrence for representing the i-th amino acid. This results in a uniform distribution of synonymous codons, maximizing the entropy associated with the i-th amino acid in each sequence. Consequently, the maximum entropy for the i-th amino acid within each sequence is as follows:

$$H_i^{max} = -log\frac{1}{k_i} = log k_i$$

Similarly, if only one of the synonymous codons is exclusively utilized for representing the i-th amino acid, it indicates an extreme bias in the codon usage. In this case, the entropy associated with the i-th amino acid in each sequence reaches the minimum value of 0.

The SUCO (O_i) can be regarded as a quantitative measure of the bias in synonymous codon usage for the i-th amino acid within each sequence, which is defined as the normalized difference between the maximum entropy and the observed entropy for the i-th amino acid in each sequence:

$$O_i = \frac{H_i^{max} - H_i}{H_i^{max}} = 1 - \frac{H_i}{H_i^{max}}$$

Clearly, the value of O_i falls within the range of 0 to 1, where a value of 0 indicates that the synonymous codon usage for the i-th amino acid is completely random. Conversely, a value of 1 suggests an extreme bias in the synonymous codon usage. Therefore, O_i can be interpreted as a measure of the degree of bias in the synonymous codon usage for the i-th amino acid within each sequence.

The weighted sum of SCUO (O_w) each sequence can be represented as:

$$O_w = \sum_{i=1}^{n} w_i O_i$$

where n is the number of distinct amino acids (methionine and tryptophan excluded) in the sequence and w_i represents the relative abundance or composition ratio of the i-th amino acid in each sequence:

$$w_i = \frac{\sum_{j=1}^{k_i} x_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{k_i} x_{ij}}$$

Obviously, the value of O_w also ranges from 0 to 1. A value of 1 indicates maximum bias, while a value of 0 signifies no bias.

Weighted Sum of Relative Entropy (Ew)

Similar to Synonymous Codon Usage Order (SCUO)[10,11], the Weighted Sum of Relative Entropy (E_w)[12] is is also based on Shannon's information theory and maximum entropy theory to evaluate the synonymous codon usage bias. E_w is defined as the weighted sum of Relative Entropy (E_i):

$$E_w = \sum_{i=1}^{n} w_i E_i$$

where *n* is the number of distinct amino acids in the sequence and the formula for calculating w_i for each amino acid is provided in SUCO. Furthermore, the Relative Entropy[13] for the i-th amino acid (E_i) represents the evenness component of uncertainty, quantifying the ratio of the observed uncertainty in amino acid usage to the maximum possible uncertainty. This measure is calculated using the following function:

$$E_i = \frac{H_i}{H_i^{max}} = \frac{H_i}{log k_i}$$

It is worth noting that if degeneracy for amino acids is equal to 1, which means that the amino acids are encoded by only one codon, the denominator becomes zero, leading to an undefined value for E_i. Due to such nondegenerate amino acids cannot exhibit synonymous codon usage bias, they are typically excluded from consideration. Clearly, the E_w value range from 0 to 1. In contrast to O_w, a value of 0 indicates maximum bias, while a value of 1 signifies no bias.

Codon Preference (P)

The Codon Preference (P)[14] assesses the propensity of a specific set of codons to align with a predefined preferred usage. P is computed for all three reading frames, making it valuable for gene identification in sequenced DNA, predicting the relative expression level of genes, and detecting DNA sequencing errors that may lead to base insertions or deletions within coding sequences. A higher P value indicates a more frequent usage of preferred codons.

The preference parameter, for a codon xyz is calculated as follows:

$$P_{xyz} = \frac{f_{xyz}/F_{xyz}}{r_{xyz}/R_{xyz}}$$

where f_{xyz} represents the observed frequency of codon xyz. F_{xyz} is the frequency of all the codons for an amino acid and can be calculated as follows:

$$F_{xyz} = \sum_{xyz \in a} f_{xyz}$$

Then, r_{xyz} represents the frequency of codon in a random sequence, and R_{xyz} is the sum of r_{xyz} within the synonymous family which includes codon xyz. The formulae for their calculation are as follows:

$$R_{xyz} = \sum_{xyz \in a} r_{xyz}$$

$$r_{abc} = \frac{N_x N_y N_z}{N^3}$$

where N_x, N_y and N_z represent the observed frequencies of the individual bases x, y, and z at positions 1, 2, and 3 of the codon. N is the total number of bases in the sequence.

The preference of a gene P is defined as the geometric mean of P_{xyz}:

$$P = \left(\prod_{K=1}^{L} P_k \right)^{\frac{1}{L}} = \exp\left(\frac{1}{L} \sum_{k=1}^{L} InP_k \right)$$

where L is the length of a gene measured in codons and P_k is the preference for the k-th codon in the gene.

Maximum-likelihood Codon Bias (MCB)

The Maximum-likelihood Codon Bias (MCB)[15] is designed to account for background nucleotide composition and can be further adapted to correct for di-nucleotide biases. This method proves valuable for estimating ancestral codon usage bias and conducting genetic population analysis. A higher value of MCB indicates a stronger bias in codon usage. It estimates the bias in codon usage by assigning weights to each amino acid. These weights are derived from the likelihood of occurrence of each amino acid, considering its frequency and codon degeneracy. Nevertheless, MCB is not a maximum-likelihood method in the strictest sense.

The calculation can be expressed as follows:

$$B_g = \frac{\sum B_a \cdot logN_a}{A}$$

where N_a is the observed frequency of amino acid a, and A is the number of amino acids contributing to the index. The bias for an individual amino acid, B_a, can be obtained using the following formula:

$$B_a = \sum_{c \in a} \frac{(O_c - E_c)^2}{E_c}$$

Where O_c represents the observed count of codon c in a gene, E_c denotes the expected count of the same codon.

Indices based on codon frequency in a reference set of genes

Indices in this section rely on the comparison to the codon frequency in a reference set of genes. These indices necessitate knowledge of the reference genome of the host organism and compare the variation in codon usage between the reference genes and the host organism. Coding sequences containing codons that closely resemble those in the reference set will receive a higher score, indicating higher gene expression. The differences among the indices lie in the methodology employed to compute the similarity score.

Codon Adaptation Index (CAI)

The Codon Adaptation Index (CAI)[16] serves as a measure to quantify the degree of codon adaptation within a gene by comparing its codon usage with the preferred or optimal codons, which ranges from 0 to 1. A higher CAI score suggests that the codon usage in the gene is more adapted towards the preferred codons observed in highly expressed a reference set, and a value of 1 is considered ideal. Genes with higher scores are expected to exhibit greater translation efficiency and higher protein expression levels.

To calculate CAI, the first step is to construct a reference table of RSCU values specifically for highly expressed genes in the organism under investigation. The relative adaptiveness (w_{ij}) of a codon is calculated as the ratio of the frequency of use of that codon to the frequency of the optimal codon for the same amino acid:

$$w_{ij} = \frac{RSCU_{ij}}{RSCU_{imax}} = \frac{x_{ij}}{x_{imax}}$$

where $RSCU_{imax}$ refers to the RSCU value of the codon that is most frequently used among all synonymous codons for the i-th amino acid, while x_{imax} represents the count or frequency of occurrences of that codon.

The CAI for a gene is then calculated as the geometric mean of the w_k values corresponding to each of the codon used in that gene:

$$CAI = \exp\left(\frac{1}{L}\sum_{k=1}^{L} lnw_k\right)$$

where L is the length of the gene measured in codons and w_k is the w value for the *k*-th codon in that gene. L does not have an inherent impact on CAI. However, it is worth noting that CAI values for shorter genes may exhibit increased variability due to sampling effects.

Codon Frequency Distribution (CFD)

The Codon Frequency Distribution (CFD)[17] is a factor that shows percentage of rare codon (defined as <30% usage frequency), which can be calculated as follows:

$$w_{ij} = \begin{cases} 1, if\ x_{ij}/x_{imax} < 0.3 \\ 0, else \end{cases}$$

where x_{ij} is the frequency of j-th codon for the i-th amino acid, and x_{imax} is the frequency of the optimal codon for the same amino acid. The ratio of x_{ij} to x_{imax} ranges from 0 to 1, and a value of

1 is set for the codon with the highest usage for a given amino acid in the desired expression organism. Then, the CFD for a gene is then calculated as the arithmetic mean of the w_k values corresponding to each of the codon used in that gene:

$$CFD = \frac{1}{L}\sum_{k=1}^{L} w_k$$

where L is the length of the gene measured in codons and w_k is the w value for the *k*-th codon in that gene.

Frequency of Optimal Codons (FOP)

The Frequency of Optimal Codons (FOP)[18,19] is a simplified version of CAI that provides a lower-resolution assessment of codon usage bias, which is derived by calculating the ratio of the number of optimal codons (ATG, TGG and stop codons excluded) to the total number of codons present in the gene:

$$FOP = \frac{\#\ optimal\ codons\ in\ sequence}{\#\ total\ codons\ in\ sequence}$$

The value of FOP ranges from 0 to 1, representing the occurrence of the optimal codon within a gene sequence. A value of 0 indicates that the optimal codons never appear, while a value of 1 indicates that the optimal codons always appear in the gene sequence. It is noteworthy that the optimal codons can be defined based on several factors, including nucleotide chemistry, codon usage bias, and tRNA availability. In our work, optimal codons were chosen to be the codons with the highest frequency of occurrence for each amino acid in reference.

Codon Usage Similarity Index (COUSIN)

The Codon Usage Similarity Index (COUSIN)[20] compares the CUB of a query against that of a reference and normalizes the output using a Null Hypothesis of random codon usage. This index is valuable in identifying differential heterogeneity both between and within genomic data sets. A COUSIN score of 1 indicates that the codon usage preferences in the query are similar to those in the reference dataset. A score of 0 suggests that the codon usage preferences in the query are similar to those in the null hypothesis, where there is an equal usage of synonymous codons. Scores between 0 and 1 imply that the codon usage preferences in the query are similar to those in the reference, but with a smaller magnitude. Scores above 1 indicate that the codon usage preferences in the query are similar to those in the reference, but with a larger magnitude.

To calculate COUSIN, it is necessary to first calculate the deviation scores $dev_{c,a}$:

$$dev_{c,a} = f_{c,a}^{ref} - f_{c,a}^{H_0}$$

where $f_{c,a}^{ref}$ is the frequency of the codon c among its synonymous codons in the reference, and $f_{c,a}^{H_0}$ is the observed frequency of that codon among its synonymous codons in the query sequence.

Then, the weight for each codon in the reference and query gene is defined by multiplying the codon frequency in the reference by its deviation score:

$$W_{c,a}^{ref} = f_{c,a}^{ref} \times dev_{c,a}$$

$$W_{c,a}^{que} = f_{c,a}^{que} \times dev_{c,a}$$

Using the identical deviation score to calculate the weights enables a direct comparison of the scores between the query and the reference.

The $COUSIN_{18}^{a}$ for each amino acid a is determined by the ratio of the sum of weights of all synonymous codons associated with the amino acid in the query sequence, to the corresponding cumulative weights in the reference sequence. The $COUSIN_{18}$ for query sequence is obtained by summing the individual $COUSIN_{18}^{a}$ scores of all amino acids.

$$COUSIN_{18}^{a} = \frac{1}{N} \times \frac{\sum_{c \in a} W_{c,a}^{que}}{\sum_{c \in a} W_{c,a}^{ref}}$$

$$COUSIN_{18} = \sum_{a \in A} COUSIN_{18}^{a}$$

where N represents the count of amino acids present in both the query and the reference, and A denotes the set of these amino acids.

The $COUSIN_{59}^{a}$ for each amino acid a is a weighted average of the ratio of query and reference. The $COUSIN_{59}$ for query sequence is obtained by summing the individual $COUSIN_{59}^{a}$ scores of all amino acids:

$$COUSIN_{59}^{a} = f_{a}^{que} \times \frac{\sum_{c \in a} W_{c,a}^{que}}{\sum_{c \in a} W_{c,a}^{ref}}$$

$$COUSIN_{59} = \sum_{a \in A} COUSIN_{59}^{a}$$

Codon Bias Index (CBI)

The Codon Bias Index (CBI)[21] provides insights into the presence of components with high CUB within a specific gene. It can be employed to characterize the expression of foreign genes in a host organism. CBI quantifies the usage of optimal codons by calculating the ratio of optimal codons to the total number of codons in a gene, with the expected usage serving as a scaling factor. A value of 1 denotes the exclusive utilization of preferred codons, whereas a value of 0 signifies random codon selection. Negative values indicate a higher frequency of nonpreferred codons being employed.

The formula for calculating CBI is as follows:

$$CBI = \frac{N_{pfr} - N_{rand}}{N_{tot} - N_{rand}}$$

where N_{pfr} represents the total number of occurrences of preferred codons, N_{rand} denotes the expected number of preferred codons if all synonymous codons were used equally, and N_{tot} corresponds to the total number of codons in the sequence. These quantities can be calculated as follows:

$$N_{pfr} = \sum_{c \in C_{prf}} N_c$$

$$N_{rand} = \sum_{a \in A} N_a \frac{O_a^{prf}}{K_a}$$

$$N_{tot} = \sum N_c$$

where C_{prf} represents the subset of optimal codons selected from all codons C included in the analysis. N_c refers to the number of occurrences of codon c in the sequence, while N_a represents the number of occurrences of amino acid a. O_a^{prf} quantifies the number of instances where optimal codons are used for amino acid a, and K_a reflects the redundancy of amino acid a. On our website, we defined optimal codons as the codons with the highest frequency of occurrence for each amino acid in the reference. Moreover, methionine (M) and tryptophan (W) are not involved in the calculation.

Mean Dissimilarity-based Index (Dmean)

The Mean Dissimilarity-based Index (Dmean)[22] is a measure that quantifies the level of diversity in synonymous codon usage across different gene sets or genomes. Dmean can also be used to analyze other compositional features, such as amino acid usage and dinucleotide relative abundance, as a genomic signature. By utilizing Dmean, we can gain a better understanding of the diversity in composition among genes. The value of Dmean ranges from 0 to 2. When all genes exhibit identical synonymous codon preferences for all amino acids, Dmean will be at its minimum value of 0. In simpler terms, lower Dmean values indicate a lower level of diversity in the usage of codons. This measure allows researchers to compare and rank different genomes based on their overall codon usage diversity, which can be calculated as follows:

$$Dmean = \frac{\sum D(X_i, X_j)}{N(N-1)/2}$$

where N is the total number of genes, and $D(X_i, X_j)$ is the pearson correlation distance, which can be employed to measure the dissimilarity in synonymous codon usage between two genes denoted as X_i and X_j. This distance metric is computed as one minus the Pearson's product moment correlation coefficient $cor(X_i, X_j)$:

$$D(X_i, X_j) = 1 - cor(X_i, X_j)$$

The $cor(X_i, X_j)$ is a widely used measure that quantifies the linear relationship between two genes, and can be calculated as the ratio of the covariance between X_i and X_j to the product of their standard deviations:

$$cor(X_i, X_j) = \frac{cov(X_i, X_j)}{\sigma(X_i) \cdot \sigma(X_j)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X_j})(X_i - \bar{X_j})}{|X_i - \bar{X_j}| \cdot |X_i - \bar{X_j}|}$$

This coefficient ranges from -1 to 1. A positive value indicates a direct or positive relationship, whereas a negative value indicates an inverse or negative relationship. A value close to 0 suggests a weak or negligible linear correlation. Therefore, the $D(X_i, X_j)$ can range from 0 to 2. A value of 0 indicates the maximum similarity between two genes in terms of synonymous codon usage, while a value of 2

represents the maximum diversity.

Moreover, X_i and X_j in the i-th and j-th coding sequence is two collections of x_{ac}, which consists of 59 codons (methionine, tryptophan, and stop codons were excluded). The value of x_{ac} can be calculated as follows:

$$x_{ac} = \frac{n_{ac}}{max(n_{ac})}$$

where n_{ac} represents the number of occurrences of c-th codon for the a-th amino acid, and $max(n_{ac})$ represents the number of occurrences of the most frequently used synonymous codon for the same amino acid.

On our website, the Dmean is calculated as the Pearson correlation distance between the reference gene and the query gene. Lower values indicate a greater similarity in the pattern of codon usage between the reference and the query gene, as well as higher levels of gene expression.

Relative Codon Adaptation (RCA)

For any given reference set, the Relative Codon Adaptation (RCA)[23] first calculates the expected frequency of a codon based on its positional base frequencies and then quantifies codon adaptation by comparing the observed codon frequency to the expected codon frequency. Consequently, RCA explicitly takes into account the genomic base composition, enabling more reliable and precise estimates of gene expression, and a higher RCA score for genes that tend to include codons that are more frequent in highly expressed genes. This enhanced accuracy is particularly noticeable in scenarios with high mutational bias or reduced selection for translational efficiency, where CAI might be susceptible to misinterpretation due to mutational bias artifacts in the reference set.

Similar to CAI and RCBS, the score of RCA in the gene is calculated as the geometric mean of the RCA_{xyz} values for each codon xyz:

$$RCA = \left(\prod_{k=1}^{L} RCA_{xyz(k)}\right) = \exp\left(\frac{1}{L}\sum_{k=1}^{L} RCA_{xyz(k)}\right)$$

where L represents the length of the query sequence measured in codons and $RCA_{xyz(k)}$ is the RCA value for the k-th codon in the gene, which can be calculated as follows:

$$RCA_{xyz} = \frac{f(x,y,z)}{f_1(x)f_2(y)f_3(z)}$$

where $f(x,y,z)$ represents the observed frequency of codon xyz in the reference, and x, y, and z denote the first, second, and third nucleotides of that codon, respectively. Additionally, $f_1(x)$, $f_2(y)$ and $f_3(z)$ represent the observed frequencies of the individual bases x, y, and z at positions 1, 2, and 3 of the codons in the same reference set.

Codon Usage Frequency Similarity (CUFS)

The Codon Usage Frequency Similarity (CUFS)[24] is a novel measure of functional and expression similarity between genes that incorporates the frequency of codons and reflects similarities in amino acid usage. By examining the codon bias and its connection to gene expression regulation, CUFS

becomes a valuable tool for understanding gene function. Despite considering post-transcriptional aspects and genomic organization, CUFS demonstrates a strong correlation with 3D genomic distance, highlighting its significance in assessing functional similarity between genes.

The CUFS offers a continuous metric based on Endres–Schindelin (ES) metric[25] to evaluate the degree of similarity between genes and returns a distance estimation. Specifically, as genes demonstrate greater resemblance in terms of codon usage frequencies, their inferred distance is substantially reduced, resulting in lower CUFS values. In other words, a lower CUFS value indicates greater resemblance in terms of codon usage frequencies. A value of 0 indicates that the codon usage frequency between two genes is completely identical. Given the frequency vectors of a gene pair p and q, the calculation of Codon Usage Frequency (CUF) distance(similarity) between them, utilizing the ES metric, is as follows:

$$m = \frac{1}{2}(p + q)$$

$$d_{ES}(p, q) = \sqrt{d_{KL}(p, m) + d_{KL}(q, m)}$$

where d_{KL} is Kullback-Leibler (KL)[26] divergence, also known as relative entropy, which is a measure used to quantify the amount of information lost when one probability distribution is used to approximate another. It calculates the average difference in information content between corresponding elements of the two distributions. When the KL divergence is 0, it signifies that the two distributions are completely identical. Mathematically, KL divergence is defined as follows:

$$d_{KL}(p, q) = \sum p_i log \frac{p_i}{q_i}$$

The frequency vectors of codon usage for the same amino acid can be computed as follows:

$$c_i = \frac{n_i}{\sum_{j \in AA} n_j}$$

$$\sum_{i=1}^{61} c_i = 20$$

where the number of observed codons (n_i) is normalized by dividing it by the sum of all synonymous codons encoding the same amino acid.

Codon Usage Bias (B)

The Codon Usage Bias (B)[27] is a metric derived from the frequency-weighted sum of distances between the relative codon usage frequencies of two sets of genes. It is utilized to estimate the expression level by comparing the fraction of the distance of the query set concerning all genes to the distance from a reference set or a linear combination of reference sets.

When considering F as a particular gene and R as the set of all genes (reference), the codon bias of F with respect to R can be calculated using the following formula:

$$B(F|R) = \sum_a P_a(F) \sum_{xyz \in a} |f_{xyz} - r_{xyz}|$$

where f_{xyz} and r_{xyz} are the observed frequency of codon xyz, normalized by dividing each by the

sum of all synonymous codons encoding the same amino acid. For each amino acid a, $\sum_{xyz \in a} f_{xyz} = 1$ and $\sum_{xyz \in a} r_{xyz} = 1$. $P_a(F)$ represents the set of normalized amino acid frequencies of the gene F, and $\sum_a P_a(F) = 1$.

Similarly, the codon bias of R with respect to F is calculated as follows:

$$B(R|F) = \sum_a P_a(R) \sum_{xyz \in a} |r_{xyz} - f_{xyz}|$$

where $P_a(R)$ represents the set of normalized amino acid frequencies of the gene R, and $\sum_a P_a(R) = 1$.

For $B(F|R)$ and $B(R|F)$, the maximum possible value is 2.00 and uncommon to exceed 0.5. Codon usage differences between two gene families generally range from 0.05 to 0.300. A higher value indicates significant codon usage differences between the two sets of genes. A symmetric version of B is obtained by averaging $B(F|R)$ and $B(R|F)$:

$$B = \big(B(F|R) + B(R|F)\big)/2$$

In general, the value of $B(F|R)$ and $B(R|F)$ differ little, indicating that differences in amino acid usages between F and R have little influence on the calculated relative codon biases. Consequently, we have opted to adopt the $B(F|R)$ version as the ultimate outcome.

Indices based on adaptation to the tRNA levels and their supply

Indices in this section are based on the adaptation of codons to the levels of tRNA in the cell. tRNA molecules are considered as major factors that influence translation elongation at the genomic level. Previous research has indicated that intracellular tRNA levels are correlated with codon usage and amino acid composition in various prokaryotic and eukaryotic species[28,29]. It is widely accepted that the preference for certain codons is due to differences in the abundance of corresponding tRNAs within the cell. Generally, codons with higher tRNA abundance are utilized more frequently. Genes that employ these preferred codons tend to exhibit enhanced translation efficiency and accuracy.

tRNA Adaptation Index (tAI)

The tRNA Adaptation Index (tAI)[30] is a quantifiable metric used to assess translational efficiency by considering the intracellular abundance of tRNA molecules and the efficacy of each codon-anticodon interaction. The tAI value ranges from 0 to 1, with higher values indicating that genes tend to incorporate codons that are more adapted to the tRNA pool. Specifically, the index incorporates weights that have been initially derived from gene expression data in Saccharomyces cerevisiae. These weights assign values to each wobble interaction, reflecting the efficiency of the codon-anticodon pairing during translation, which can be calculated as follows:

$$W_i = \sum_{j=1}^{n_i} \big(1 - s_{ij}\big) tGCN_{ij}$$

where n_i is the number of tRNA types/anticodons that pair with the ith codon, $tGCN_{ij}$ is the gene

copy number of the j-th tRNA molecule that recognizes the i-th codon. The copy numbers of tRNA can be retrieved from Genomic tRNA Database (http://gtrnadb.ucsc.edu/). The s_{ij} represents the selective constraint governing the efficiency of the codon-anticodon coupling. This value falls within the range of 0 to 1, with values closer to 0 indicating a more efficient wobble interaction between the codon and anticodon.

The table below displays the optimized s-values values of the codon-anticodon coupling:

S	dosReis[30]	Tuller[31]
$S_{G:U}$	0.41	0.561
$S_{I:C}$	0.28	0.28
$S_{I:A}$	0.9999	0.9999
$S_{U:G}$	0.68	0.68
$S_{L:A}$	0.89	0.89

L, lysidine; I, inosine. The values of dosReis and Tuller can refer to literature[30], [31]. In this work, we use the s-values measured by dosReis et al.

To determine the classical translational efficiency cTE_i of a codon *i*, normalized weights can be utilized. These normalized weights can be obtained by the following formula to have a maximum value of 1:

$$cTE_i = \begin{cases} W_i/W_{max} & if\ W_i \neq 0 \\ W_{geomean} & if\ W_i = 0 \end{cases}$$

where W_{max} is the maximum W_i value and $W_{geomean}$ is the geometric mean of all cTE_i with $W_i \neq 0$.

The tAI of a gene (tAI_g) is the geometric mean of these relative adaptiveness values, and ranges from 0 (low efficiency) to 1 (high efficiency):

$$tAI_g = \left(\prod_{k=1}^{L} cTE_k\right)^{1/L} = \exp\left(\frac{1}{L}\sum_{k=1}^{L} ln(cTE_k)\right)$$

where L is the length of a gene measured in codons and cTE_k is the translational effiency for the k-th codon in the gene.

Genetic tRNA Adaptation Index (gtAI)

The initial implementation of the tRNA adaptation index (tAI)[30] exhibited notable shortcomings. The s_{ij} weights generated were optimized based on gene expression patterns in Saccharomyces cerevisiae, potentially leading to variations across different species. The subsequent stAI[32,33] method utilized a hill

climbing algorithm for s_{ij} weight optimization, but this approach wasn't optimal for complex search spaces, potentially struggling to locate the global maximum even with varied starting points. To overcome these limitations, a species-specific approach called the Genetic tRNA Adaptation Index (gtAI)[34] was developed.

The gtAI employs a genetic algorithm to obtain the best set of s_{ij} weights, addressing the issue of obtaining meaningful weights for each organism. It also uses robust CUB indices, namely ENC[2,3] and RSCU[1], instead of the directional codon bias score (DCBS) [5] used in stAI. Highly expressed genes are influenced by translational selection, resulting in the incorporation of codons that better adapt to the intracellular tRNA pool. Therefore, a correlation is expected between RSCU and absolute adaptiveness (W_i) values. In the stAI, unique s_{ij} weights are optimized for each organism by maximizing the nonparametric correlation between RSCU and W_i using a genetic algorithm. This algorithm is a metaheuristic search approach inspired by the concept of survival of the fittest, which algorithm is as follows:

Input: Genome coding sequences
Initialize S, vector of the initial population as chromosomes (Sij sets) with random Sij values (genes) Generation time = n;
For s in S **do**
Evaluate fitness function(s);
n += 1
IntialLabel;
Test:
Selection(s) where Sij sets that exhibit higher correlation between RSCU and Wi;
Do:
Crossover(s);
Mutation(s);
Evaluate fitness function(s);
If n = Generation time, **then**
Output = Best fitness(s);
Else
Go to IntialLabel
Output: the best set of Sij weights + tAI values

In this work, the genetic algorithm operates with a population size of 60 and runs for 1000 iterations to search for the best s_{ij} weights that maximize the correlation between RSCU and W_i. Then, the best set of s_{ij} weights will be used to calculate the genetic tRNA adaptation index (gtAI) using the following equations:

$$W_i = \sum_{j=1}^{n_i} (1 - s_{ij}) tGCN_{ij}$$

$$gtAI = \left(\prod_{k=1}^{L} W_k\right)^{1/L} = \exp\left(\frac{1}{L}\sum_{k=1}^{L} lnW_k\right)$$

where L is the length of a gene measured in codons and W_k is the weight for the k-th codon in the gene. Same as tAI, the value of gtAI ranges from 0 to 1, with higher values indicating that genes tend to incorporate codons that are more adapted to the tRNA pool.

P2 Index

The P2 Index[35] quantifies the efficiency of interactions between codons and their corresponding anticodons, offering insights into translation efficiency in cases where information regarding preferred codon sets is unavailable. A higher P2 value is typically observed in highly expressed genes, whereas genes with lower expression levels tend to exhibit lower P2 values. Furthermore, a P2 value exceeding 0.5 indicates the presence of translational selection influencing the coding sequence. P2 Index is derived from the proportion of pyrimidine-ending codons that exhibit intermediate strength, which can be calculated according to the following equation:

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

where W = A or U, S = C or G, and Y = C or U. WWC, SSU, WWY and SSY represent the frequency of their corresponding codons, respectively.

Indices based on complex patterns of codon usage

Indices in this section are based on measures of complex patterns of codon usage. In the coding region, sequences longer than a single codon may contain regulatory codes that pertain to various aspects of gene expression and intracellular processes. Therefore, it becomes necessary to develop indices that can capture these aspects, which cannot be adequately computed based solely on single codon distributions. Changes in the composition of codons can have an impact on these longer patterns, inducing a selection pressure on codon composition. Consequently, the codon-based indices should encompass measures that go beyond analyzing single codons and instead employ more advanced statistical methods. Indices in this group capture complex signals that are influenced by and, in turn, influence codons.

GC Content at the First Position of Synonymous Codons (GC1)

The GC Content at the First Position of Synonymous Codons (GC1) represents the frequency of G+C usage at the first position of synonymous codons, which can be calculated as follows:

$$GC1 = \frac{GNN + CNN}{ANN + TNN + GNN + CNN}$$

where N represent an arbitrary base, and ANN, TNN, GNN and CNN represent the frequency of their corresponding codons, respectively.

GC Content at the Second Position of Synonymous Codons (GC2)

The GC Content at the Second Position of Synonymous Codons (GC2) represents the frequency of G+C usage at the second position of synonymous codons, which can be calculated as follows:

$$GC2 = \frac{NGN + NCN}{NAN + NTN + NGN + NCN}$$

where N represent an arbitrary base, and NAN, NTN, NGN and NCN represent the frequency of their corresponding codons, respectively.

GC Content at the Third Position of Synonymous Codons (GC3)

The GC Content at the Third Position of Synonymous Codons (GC3)[36] represents the frequency of G+C usage at the third position of synonymous codons, which exhibits variability in nucleotide composition. Changes in the third position of codons typically do not result in alterations to the encoded amino acids. Furthermore, base mutations occurring at this position are often subjected to less selection pressure. Consequently, the investigation of the base composition at position 3 holds significance in the study of codon preference. GC3 ranges from 0 to 1, and a higher value is correlated with highly expressed genes. The formula for GC3 is as follows:

$$GC3 = \frac{NNG + NNC}{NNA + NNT + NNG + NNC}$$

where N represent an arbitrary base, and NNA, NNT, NNG and NNC represent the frequency of their corresponding codons, respectively.

GC Content (GC)

The GC content, representing the percentage of guanine (G) and cytosine (C) nitrogenous bases in a gene, is a valuable measure for assessing the base composition. This measure provides insights into the relative proportion of G and C bases compared to the other two bases, adenine (A) and thymine (T) in DNA, or adenine (A) and uracil (U) in RNA. The formula for GC3 is as follows:

$$GC = \frac{G + C}{A + T + G + C}$$

where A, T, G, C represent the number of corresponding bases in the gene, respectively.

Although AT-rich intergenic regions can reduce the overall GC content of a gene or genome, there exists a strong correlation between the local GC composition (GC1, GC2, and GC3) and the overall GC composition of the sequence. In fact, the higher the overall GC content bias, the higher the local GC composition.

Effective Number of Codon Pairs (ENcp)

The Similar to ENC, Effective Number of Codon Pairs (ENcp)[37] is another metric used to quantify codon usage bias in a genome. By comparing the observed frequencies of codon pairs to the expected frequencies under the assumption of equal codon pair usage, ENcp provides a numerical value that reflects the level of bias or preference in codon pair usage. ENcp can range from 20 (when no bias or all the synonymous codon pairs are used uniformly) to 61 (when the maximal bias or preference observed in synonymous codon pair usage). The calculation of ENcp is analogous to that of ENC, with

the additional incorporation of a square root:

$$\widehat{N}_{cp} = \sqrt{\sum_m \frac{k_m}{\bar{F}_m}}$$

where $\bar{F}_m$ represents the average value of F_{AA} for all amino acid pairs with degeneracy m, and k_m denotes the count of amino acid pairs that share m synonymous representations.

For standard genetic code, ENcp can be calculated as follows:

$$\widehat{N}_{cp} = \sqrt{4 + \frac{36}{\bar{F}_2} + \frac{4}{\bar{F}_3} + \frac{101}{\bar{F}_4} + \frac{30}{\bar{F}_6} + \frac{90}{\bar{F}_8} + \frac{1}{F_9} + \frac{64}{\bar{F}_{12}} + \frac{25}{\bar{F}_{16}} + \frac{6}{\bar{F}_{18}} + \frac{30}{\bar{F}_{24}} + \frac{9}{\bar{F}_{36}}}$$

Codon Pair Score (CPS)

The Codon Pair Score (CPS)[38,39] is a metric used to assess the similarity in codon pair preferences between viruses and their host species. It is calculated by taking the natural logarithm of the observed ratio of a specific codon pair's occurrence to the expected number of occurrences of that codon pair in all protein-coding sequences of a species:

$$CPS_i = ln\left(\frac{F(AB)}{\frac{F(A)\cdot F(B)}{F(X)\cdot F(Y)}\cdot F(XY)}\right)$$

where $F(AB)$ is the frequency of a specific codon pair AB, $F(A)$ and $F(B)$ represent the frequencies of individual codons A and codon B, respectively. Similarly, $F(XY)$ is the frequency of an amino acid pair XY, $F(X)$ and $F(Y)$ represent the frequencies of individual amino acids X and amino acid Y, respectively. A positive CPS value signifies that the given codon pair is statistically over-represented, while a negative CPS indicates that the pair is statistically under-represented. In other words, a higher positive CPS score signifies a stronger preference for a specific codon pair. In addition, Codon pairs that are equally under or overrepresented have a CPS value equidistant from 0.

The average CPS values of all codon pairs present in a gene are collectively referred to as Codon Pair Bias (CPB), which can be calculated as follows:

$$CPB = \sum_{k=1}^{L} \frac{CPS_k}{L-1}$$

where L is the length of a gene and CPS_k is the CPS value for the k-th codon in the gene.

Codon Volatility

The Codon Volatility[40] refers to the probability of a nonsynonymous change occurring through a random point mutation in the codon. The volatility of a codon is influenced by mutational patterns, including the ratio of transitional to transversional changes. In the simplest mutation model (SMM),

where all nucleotides have equal mutation rates and are equally exchangeable, codon volatility is determined by the proportion of neighboring codons that encode different amino acids due to point mutations. For example, a codon such as TTG (encoding Leucine) has a volatility of 6/8, as 6 out of its 8 neighboring codons (excluding stop codons) lead to nonsynonymous changes. The range of codon volatility spans from 0.5 (e.g., CGA for Arginine) to 1 (e.g., TGG for Tryptophan or ATG for Methionine). The volatility of a codon c is employed as a measure to quantify the likelihood that the most recent nucleotide mutation in that codon resulted in an amino acid substitution. If a gene contains numerous residues that are under selective pressure for amino acid replacements, the resulting codons within that gene will, on average, display increased volatility. Conversely, if a gene is subjected to strong purifying selection to maintain its amino acids, the resulting sequence will, on average, exhibit lower volatility.

The codon volatility is calculated as follows:

$$v(c) = \frac{1}{no.\ of\ neighbours} \sum_{neighbour\ \ c_i} D\big(acid(c), acid(c_i)\big)$$

where $no.\ of\ neighbours$ is for the sum of sense codon c_i that can mutate into c by a single point mutation. D, representing the Hamming distance, is defined as zero when two amino acids are identical and one otherwise.

The volatility of a gene can be computed as the arithmetic of codon volatility, which can be calculated as follows:

$$v(g) = \frac{\sum_{k=1}^{L} v_k}{L}$$

where L is the length of a gene measured in codons and v_k is the volatility for the k-th codon in the gene.

Negative CIS Elements

Negative cis elements[41] are DNA sequences that can repress the transcription of a gene by binding to specific transcription factors or other regulatory proteins. They can affect gene expression in various ways, such as reducing the amount of mRNA produced, altering the splicing or stability of mRNA, or interfering with the translation of mRNA. Here the number of negative CIS elements within the sequence is counted.

Negative Repeat Elements

Negative repeat elements[42] can reduce protein expression by creating frameshift mutations or premature stop codons. This can happen when the repeated sequence overlaps with a coding region or a splice site. The repeated sequence can cause errors in DNA replication or repair, leading to changes in the amino acid sequence of the protein. This can result in loss of function or toxicity of the protein. Here the number of negative repeat elements within the sequence is counted.

References

(1) Sharp, P. M.; Li, W.-H. An Evolutionary Perspective on Synonymous Codon Usage in Unicellular Organisms. *J Mol Evol* **1986**, *24* (1–2), 28–38. https://doi.org/10.1007/BF02099948.

(2) Wright, F. The 'Effective Number of Codons' Used in a Gene. *Gene* **1990**, *87* (1), 23–29. https://doi.org/10.1016/0378-1119(90)90491-9.

(3) Satapathy, S. S.; Sahoo, A. K.; Ray, S. K.; Ghosh, T. C. Codon Degeneracy and Amino Acid Abundance Influence the Measures of Codon Usage Bias: Improved Nc ($\hat{N}_c$) and ENCprime ($\hat{N}'_c$) Measures. *Genes Cells* **2017**, *22* (3), 277–283. https://doi.org/10.1111/gtc.12474.

(4) Roymondal, U.; Das, S.; Sahoo, S. Predicting Gene Expression Level from Relative Codon Usage Bias: An Application to Escherichia Coli Genome. *DNA Research* **2009**, *16* (1), 13–30. https://doi.org/10.1093/dnares/dsn029.

(5) Sabi, R.; Tuller, T. Modelling the Efficiency of Codon–TRNA Interactions Based on Codon Usage Bias. *DNA Research* **2014**, *21* (5), 511–526. https://doi.org/10.1093/dnares/dsu017.

(6) Zhang, Z.; Li, J.; Cui, P.; Ding, F.; Li, A.; Townsend, J. P.; Yu, J. Codon Deviation Coefficient: A Novel Measure for Estimating Codon Usage Bias and Its Statistical Significance. *BMC Bioinformatics* **2012**, *13* (1), 43. https://doi.org/10.1186/1471-2105-13-43.

(7) Supek, F.; Vlahoviček, K. Comparison of Codon Usage Measures and Their Applicability in Prediction of Microbial Gene Expressivity. *BMC Bioinformatics* **2005**, *6* (1), 182. https://doi.org/10.1186/1471-2105-6-182.

(8) Supek, F.; Vlahoviček, K. Correction: Comparison of Codon Usage Measures and Their Applicability in Prediction of Microbial Gene Expressivity. *BMC Bioinformatics* **2010**, *11*, 463. https://doi.org/10.1186/1471-2105-11-463.

(9) Freire-Picos, M. A.; Gonzalez-Siso, M. I.; Rodríguez-Belmonte, E.; Rodríguez-Torres, A. M.; Ramil, E.; Cerdan, M. E. Codon Usage in Kluyveromyces Lactis and in Yeast Cytochrome C-Encoding Genes. *Gene* **1994**, *139* (1), 43–49. https://doi.org/10.1016/0378-1119(94)90521-5.

(10) Wan, X.-F.; Xu, D.; Kleinhofs, A.; Zhou, J. Quantitative Relationship between Synonymous Codon Usage Bias and GC Composition across Unicellular Genomes. *BMC Evol Biol* **2004**, *4* (1), 19. https://doi.org/10.1186/1471-2148-4-19.

(11) Wan, X.-F.; Zhou, J.; Xu, D. CodonO: A New Informatics Method for Measuring Synonymous Codon Usage Bias within and across Genomes. *International Journal of General Systems* **2006**, *35* (1), 109–125. https://doi.org/10.1080/03081070500502967.

(12) Suzuki, H.; Saito, R.; Tomita, M. The 'Weighted Sum of Relative Entropy': A New Index for Synonymous Codon Usage Bias. *Gene* **2004**, *335*, 19–23. https://doi.org/10.1016/j.gene.2004.03.001.

(13) Wang, H.-C.; Badger, J.; Kearney, P.; Li, M. Analysis of Codon Usage Patterns of Bacterial Genomes Using the Self-Organizing Map. *Molecular Biology and Evolution* **2001**, *18* (5), 792–800. https://doi.org/10.1093/oxfordjournals.molbev.a003861.

(14) Gribskov, M.; Devereux, J.; Burgess, R. R. The Codon Preference Plot: Graphic Analysis of Protein Coding Sequences and Prediction of Gene Expression. *Nucl Acids Res* **1984**, *12* (1Part2), 539–549. https://doi.org/10.1093/nar/12.1Part2.539.

(15) Urrutia, A. O.; Hurst, L. D. Codon Usage Bias Covaries With Expression Breadth and the Rate of Synonymous Evolution in Humans, but This Is Not Evidence for Selection. *Genetics* **2001**, *159* (3), 1191–1199. https://doi.org/10.1093/genetics/159.3.1191.

(16) Sharp, P. M.; Li, W.-H. The Codon Adaptation Index-a Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications. *Nucl Acids Res* **1987**, *15* (3), 1281–1295. https://doi.org/10.1093/nar/15.3.1281.

(17) Ahmadpour, F.; Yakhchal, B.; Fatemi, S. S.-A.; Karkhane, A. A.; Talebi, S. Cloning and Expression of an Indigenous Mesophile Lipase and Evaluation of Bacillus Codon Translation in Pichia Pastoris under Control of Two Different Promoters. *JABR* **2016**, *3* (2), 413–418.

(18) Ikemura, T. Correlation between the Abundance of Escherichia Coli Transfer RNAs and the Occurrence of the Respective Codons in Its Protein Genes: A Proposal for a Synonymous Codon Choice That Is Optimal for the E. Coli Translational System. *Journal of Molecular Biology* **1981**, *151* (3), 389–409. https://doi.org/10.1016/0022-2836(81)90003-6.

(19) Ikemura, T. Correlation between the Abundance of Yeast Transfer RNAs and the Occurrence of the Respective Codons in Protein Genes. *Journal of Molecular Biology* **1982**, *158* (4), 573–597. https://doi.org/10.1016/0022-2836(82)90250-9.
(20) Bourret, J.; Alizon, S.; Bravo, I. G. COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. *Genome Biology and Evolution* **2019**, *11* (12), 3523–3528. https://doi.org/10.1093/gbe/evz262.
(21) Bennetzen, J. L.; Hall, B. D. Codon Selection in Yeast. *Journal of Biological Chemistry* **1982**, *257* (6), 3026–3031. https://doi.org/10.1016/S0021-9258(19)81068-2.
(22) Suzuki, H.; Saito, R.; Tomita, M. Measure of Synonymous Codon Usage Diversity among Genes in Bacteria. *BMC Bioinformatics* **2009**, *10* (1), 167. https://doi.org/10.1186/1471-2105-10-167.
(23) Fox, J. M.; Erill, I. Relative Codon Adaptation: A Generic Codon Bias Index for Prediction of Gene Expression. *DNA Research* **2010**, *17* (3), 185–196. https://doi.org/10.1093/dnares/dsq012.
(24) Diament, A.; Pinter, R. Y.; Tuller, T. Three-Dimensional Eukaryotic Genomic Organization Is Strongly Correlated with Codon Usage Expression and Function. *Nat Commun* **2014**, *5* (1), 5876. https://doi.org/10.1038/ncomms6876.
(25) Endres, D. M.; Schindelin, J. E. A New Metric for Probability Distributions. *IEEE Trans. Inform. Theory* **2003**, *49* (7), 1858–1860. https://doi.org/10.1109/TIT.2003.813506.
(26) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Schilling, D. L., Series Ed.; Wiley Series in Telecommunications; John Wiley & Sons, Inc.: New York, USA, 1991. https://doi.org/10.1002/0471200611.
(27) Karlin, S.; Mrázek, J.; Campbell, A. M. Codon Usages in Different Gene Classes of the *Escherichia Coli* Genome. *Molecular Microbiology* **1998**, *29* (6), 1341–1355. https://doi.org/10.1046/j.1365-2958.1998.01008.x.
(28) Novoa, E. M.; Pavon-Eternod, M.; Pan, T.; Ribas de Pouplana, L. A Role for TRNA Modifications in Genome Structure and Codon Usage. *Cell* **2012**, *149* (1), 202–213. https://doi.org/10.1016/j.cell.2012.01.050.
(29) Rocha, E. P. C. Codon Usage Bias from TRNA's Point of View: Redundancy, Specialization, and Efficient Decoding for Translation Optimization. *Genome Res.* **2004**, *14* (11), 2279–2286. https://doi.org/10.1101/gr.2896904.
(30) Reis, M. d. Solving the Riddle of Codon Usage Preferences: A Test for Translational Selection. *Nucleic Acids Research* **2004**, *32* (17), 5036–5044. https://doi.org/10.1093/nar/gkh834.
(31) Tuller, T.; Veksler-Lublinsky, I.; Gazit, N.; Kupiec, M.; Ruppin, E.; Ziv-Ukelson, M. Composite Effects of Gene Determinants on the Translation Speed and Density of Ribosomes. *Genome Biol* **2011**, *12* (11), R110. https://doi.org/10.1186/gb-2011-12-11-r110.
(32) Sabi, R.; Tuller, T. Modelling the Efficiency of Codon–TRNA Interactions Based on Codon Usage Bias. *DNA Research* **2014**, *21* (5), 511–526. https://doi.org/10.1093/dnares/dsu017.
(33) Sabi, R.; Volvovitch Daniel, R.; Tuller, T. StAIcalc: TRNA Adaptation Index Calculator Based on Species-Specific Weights. *Bioinformatics* **2017**, *33* (4), 589–591. https://doi.org/10.1093/bioinformatics/btw647.
(34) Anwar, A. M.; Khodary, S. M.; Ahmed, E. A.; Osama, A.; Ezzeldin, S.; Tanios, A.; Mahgoub, S.; Magdeldin, S. GtAI: An Improved Species-Specific TRNA Adaptation Index Using the Genetic Algorithm. *Front. Mol. Biosci.* **2023**, *10*, 1218518. https://doi.org/10.3389/fmolb.2023.1218518.
(35) Gouy, M.; Gautier, C. Codon Usage in Bacteria: Correlation with Gene Expressivity. *Nucl Acids Res* **1982**, *10* (22), 7055–7074. https://doi.org/10.1093/nar/10.22.7055.
(36) Stenico, M.; Lloyd, A. T.; Sharp, P. M. Codon Usage in *Caenorhabditis Elegans* : Delineation of Translational Selection and Mutational Biases. *Nucl Acids Res* **1994**, *22* (13), 2437–2446. https://doi.org/10.1093/nar/22.13.2437.
(37) Alexaki, A.; Kames, J.; Holcomb, D. D.; Athey, J.; Santana-Quintero, L. V.; Lam, P. V. N.; Hamasaki-Katagiri, N.; Osipova, E.; Simonyan, V.; Bar, H.; Komar, A. A.; Kimchi-Sarfaty, C. Codon and Codon-Pair Usage Tables (CoCoPUTs): Facilitating Genetic Variation Analyses and Recombinant Gene Design. *Journal of Molecular Biology* **2019**, *431* (13), 2434–2441. https://doi.org/10.1016/j.jmb.2019.04.021.
(38) Kunec, D.; Osterrieder, N. Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. *Cell Reports* **2016**, *14* (1), 55–67. https://doi.org/10.1016/j.celrep.2015.12.011.

(39) Coleman, J. R.; Papamichail, D.; Skiena, S.; Futcher, B.; Wimmer, E.; Mueller, S. Virus Attenuation by Genome-Scale Changes in Codon Pair Bias. *Science* **2008**, *320* (5884), 1784–1787. https://doi.org/10.1126/science.1155761.
(40) Plotkin, J. B.; Dushoff, J.; Fraser, H. B. Detecting Selection Using a Single Genome Sequence of M. Tuberculosis and P. Falciparum. *Nature* **2004**, *428* (6986), 942–945. https://doi.org/10.1038/nature02458.
(41) Cheng, J.; Maier, K. C.; Avsec, Ž.; Rus, P.; Gagneur, J. Cis-Regulatory Elements Explain Most of the MRNA Stability Variation across Genes in Yeast. *RNA* **2017**, *23* (11), 1648–1659. https://doi.org/10.1261/rna.062224.117.
(42) Pearson, C. E.; Nichol Edamura, K.; Cleary, J. D. Repeat Instability: Mechanisms of Dynamic Mutations. *Nat Rev Genet* **2005**, *6* (10), 729–742. https://doi.org/10.1038/nrg1689.