frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 04 July 2023
DOI 10.3389/fmolb.2023.1218518

OPEN ACCESS

EDITED BY
Carolin Kosiol,
University of St Andrews, United Kingdom

REVIEWED BY
Ioanna Kotari,
University of Veterinary Medicine Vienna, Austria
Paul Martin Harrison,
McGill University, Canada

*CORRESPONDENCE
Sameh Magdeldin,
sameh.magdeldin@57357.org

RECEIVED 07 May 2023
ACCEPTED 22 June 2023
PUBLISHED 04 July 2023

CITATION
Anwar AM, Khodary SM, Ahmed EA, Osama A, Ezzeldin S, Tanios A, Mahgoub S and Magdeldin S (2023), gtAI: an improved species-specific tRNA adaptation index using the genetic algorithm.
Front. Mol. Biosci. 10:1218518.
doi: 10.3389/fmolb.2023.1218518

gtAI: an improved species-specific tRNA adaptation index using the genetic algorithm

Ali Mostafa Anwar[1], Saif M. Khodary[2], Eman Ali Ahmed[1,3], Aya Osama[1], Shahd Ezzeldin[1], Anthony Tanios[1], Sebaey Mahgoub[1] and Sameh Magdeldin[1,4]*

[1]Proteomics and Metabolomics Research Program, Basic Research Department, Children's Cancer Hospital 57357 (CCHE-57357), Cairo, Egypt, [2]Department of Genetics, Faculty of Agriculture, Cairo University, Giza, Egypt, [3]Department of Pharmacology, Faculty of Veterinary Medicine, Suez Canal University, Ismailia, Egypt, [4]Department of Physiology, Faculty of Veterinary Medicine, Suez Canal University, Ismailia, Egypt

The tRNA adaptation index (tAI) is a translation efficiency metric that considers weighted values (S_{ij} values) for codon–tRNA wobble interaction efficiencies. The initial implementation of the tAI had significant flaws. For instance, generated S_{ij} weights were optimized based on gene expression in *Saccharomyces cerevisiae*, which is expected to vary among different species. Consequently, a species-specific approach (stAI) was developed to overcome those limitations. However, the stAI method employed a hill climbing algorithm to optimize the S_{ij} weights, which is not ideal for obtaining the best set of S_{ij} weights because it could struggle to find the global maximum given a complex search space, even after using different starting positions. In addition, it did not perform well in computing the tAI of fungal genomes in comparison with the original implementation. We developed a novel approach named genetic tAI (gtAI) implemented as a Python package (https://github.com/AliYoussef96/gtAI), which employs a genetic algorithm to obtain the best set of S_{ij} weights and follows a new codon usage-based workflow that better computes the tAI of genomes from the three domains of life. The gtAI has significantly improved the correlation with the codon adaptation index (CAI) and the prediction of protein abundance (empirical data) compared to the stAI.

KEYWORDS

codon usage, tRNA adaptation index, molecular evolution, translational selection, codon usage analysis

1 Introduction

Grantham et al. (1980) highlighted the unequal usage of synonymous codons among different genes and genomes in a phenomenon currently known as codon usage bias (CUB). Thenceforward, scientists were investigating the effect of synonymous mutations on the efficiency/accuracy of protein translation and several biological processes ranging from RNA processing to protein folding and their potential consequences on the overall performance and evolution of living organisms (Chamary et al., 2006; Plotkin and Kudla, 2011). Codon usage is positively associated with the analogous tRNA in a species—the tRNA pool determines the available amino acid used during the protein extension process. Therefore, protein expression and translation efficiency are highly associated with CUB (Karlin et al., 2001; Gustafsson et al., 2004). Accordingly, codons with high occurrence in a gene (putative optimal codons) improve

the protein translation rate, and rare codons will cause a reduction in the translation and might cause translation errors (Ikemura, 1981).

In biotechnology studies, heterologous expression was applied to assemble vaccines and pharmaceuticals (Han et al., 2010; Liu et al., 2018). Codon optimization was proposed to increase heterologous gene expression (Quax et al., 2015; Brandis and Hughes, 2016; Fu et al., 2020). Many studies reported the success of the codon optimization approach to upregulate gene expression up to 1,000-fold (Quax et al., 2015). Several software tools are used for codon optimization and are patented to serve commercial purposes such as GenSmart Design (https://www.genscript.com/gene-and-plasmid-construct-design.html) and GENEWIZ (https://www.genewiz.com/en-GB/Public/Services/Gene-Synthesis/codon-optimization). A number of codon optimization algorithms are not open-source (Satya et al., 2003; Huang et al., 2021) or should be requested from the authors (Fuglsang, 2003). Regardless of their availability, many of those protein expression optimization software tools are based on the tRNA adaptation index (tAI) (Gould et al., 2014; Watts et al., 2021; Raguin et al., 2023) and codon adaptation index (CAI) (Fu et al., 2020). Many indices were developed to measure the degree of preference for the unbalanced use of codons. Some are codon-specific such as relative synonymous codon usage (RSCU), and others are gene-specific such as the effective number of codons (ENc) (Wright, 1990; Sun et al., 2013) and CAI (Sharp and Li, 1987). A relatively new index named tAI was introduced by dos Reis et al. (2004) to become a formal measure for CUB associated with translational selection. The tRNA presents a complementary anticodon for an amino acid to be incorporated into the growing polypeptide chain during the translation process. The codon–anticodon interactions at the first two codon positions are governed solely by canonical (Watson–Crick) base pairing rules, unlike the third codon position at which non-canonical (wobble) base pairing also occurs (Crick, 1966). The tAI considers weights for canonical and wobble interaction efficiencies between codons and tRNA molecules. To compute the tAI, first, the absolute adaptiveness value (W_i) for codon i is calculated by the following equation:

$$W_i = \sum_{j=1}^{n_j} \left(1 - S_{ij}\right) tGNC_{ij}, \tag{1}$$

where n_j is the number of tRNA isoacceptors that can recognize the ith codon, S_{ij} is the codon–anticodon coupling efficiency having values ranging from 0 (perfect interaction) to 1 (weak interaction) (dos Reis et al., 2004), and $tGNC_{ij}$ is the gene copy number of the jth anticodon that can recognize the ith codon.

Then, each W_i is normalized to the maximum W_i value to obtain the relative adaptiveness value (w_i). Finally, the tAI of a gene can be defined as the geometric mean of the w_i values of its codons (dos Reis et al., 2004):

$$tAI_g = \exp\left(\frac{1}{O_{tot}} \sum_{i \in I} \log w_i\right), \tag{2}$$

where O_{tot} is the frequency of the total codons.

The S_{ij} weights inferred by the original tAI (otAI) implementation were based on optimizing the correlation between tAI (Eq. 1) and gene expression levels in *Saccharomyces cerevisiae* using the Nelder–Mead method under the assumption that highly expressed genes contain codons with higher adaptation to the tRNA pool (driven by the force of translational selection). In a study by Dana and Tuller (2014), two problems are associated with the original tAI implementation. First, it depends on gene expression information, often unavailable for many organisms (especially novel ones). Second, generated weights were specific for *Saccharomyces cerevisiae*. They suggested the possibility that wobble interaction efficiencies shall differ significantly among genomes from different domains. So, it would not be plausible to use the weights specifically for *Saccharomyces cerevisiae* to compute the tAI of other organisms. Consequently, they developed the species-specific tAI (stAI) (Sabi and Tuller, 2014) to solve these problems.

The inferred stAI weights are based on optimizing the correlation between the tAI (Eq. 1) and a CUB index, namely, directional codon bias score (a modified version of relative CUB (Oymondal et al., 2009)) using the hill climbing algorithm under the assumption that highly expressed genes have higher adaptation to the tRNA pool and higher CUB (Sabi and Tuller, 2014). This eliminates the need for additional gene expression data and generates weights specific to the tested organism, indicating the value of stAI in tAI computation, especially for non-fungal species. However, two limitations in the stAI are as follows: 1) using the hill climbing optimization method by which only local maxima can be reached and often gets stuck in ridges and plateau scenarios (Thengade and Dondal, 2012); hence, the best set of S_{ij} weights may not be obtained even after using different starting positions (random restart) (Russell and Norvig, 2010; Yang, 2014); 2) the outperformance of the original tAI over the stAI in predicting the protein abundance (PA) of fungal organisms (Sabi and Tuller, 2014).

Here, we introduce a novel approach for tAI computation, namely, genetic tAI (gtAI), to solve the problems associated with the stAI, which affect its performance. The gtAI uses a genetic algorithm to reach the global maximum (best set of S_{ij} weights), solving the issue of obtaining a meaningful set of S_{ij} weights for each organism. It also utilizes robust CUB indices (ENc and RSCU) different from the directional codon bias score (DCBS) employed by the stAI.

2 Materials and methods

2.1 Establishing a reference set of genes using the effective number of codons

A reference set of genes is defined as a set of genes with the highest expression levels in a genome, such as ribosomal genes and translation elongation factors (Duret, 2000; Ghaemmaghami et al., 2003; Goetz and Fuglsang, 2005). The ENc is a widely used measure of CUB at the gene level, and in theory, it negatively correlates with gene expression (Sun et al., 2013). Given the assumption that highly expressed genes are highly biased (Sabi and Tuller, 2014), a reference set is obtained by selecting genes with the lowest ENc values (highest expression) in the tested genome. The ENc is calculated using the equations of the improved ENc implementation by Sun et al. (2013):

$$EN_c = N_s + \frac{K_2 \sum_j^{k_2} nj}{\sum_{j=1}^{k_2} \left(njF_{CF.j}\right)} + \frac{K_3 \sum_j^{k_3} nj}{\sum_{j=1}^{k_3} \left(njF_{CF.j}\right)} + \frac{K_4 \sum_j^{k_4} nj}{\sum_{j=1}^{k_4} \left(njF_{CF.j}\right)}, \tag{3}$$

where N_S is the number of codon families with a single codon. K_i is the number of i-fold codon families. In addition, $F_{CF.\ j}$ is F_{CF} for family j obtained from the following equation:

$$F_{CF} = \sum_{i=1}^{m} \left(\frac{n_i + 1}{n + m}\right)^2, \tag{4}$$

where n_i is the count of codon i in the codon family of m synonymous codons.

2.2 Calculating the relative synonymous codon usage for the reference set

The RSCU is a codon-specific CUB measurement defined as the ratio of the observed to the expected frequency of codons, under the null hypothesis that all synonymous codons for a particular amino acid are used equally (Sharp and Li, 1986). It gives an accurate value for each amino acid codon ranging from 0 to the number of synonymous codons for that amino acid. The RSCU values for the reference set are calculated using the following equation:

$$RSCU = \frac{O_{ac}}{\frac{1}{k_a}\sum_{c\in C_a} O_{ac}}, \quad (5)$$

where O_{ac} is the count of codon c for the amino acid a and k_a is the number of synonymous codons in the amino acid a family.

2.3 S_{ij} weight inference by the genetic algorithm

Since highly expressed genes are influenced by translational selection to include more codons with higher adaptation to the intracellular tRNA pool (i.e., optimal codons) (Sabi and Tuller, 2014), we expect to find a correlation between RSCU (Eq. 5) and absolute adaptiveness (W_i) values (Eq. 1). Therefore, we inferred unique S_{ij} weights for each organism by optimizing the non-parametric (Spearman's rank) correlation between RSCU (of the reference set) and W_i values using a genetic algorithm (https://pypi.org/project/gaft/). It should be noted that the correlation between RSCU and W_i is at the level of codons.

The genetic algorithm is a metaheuristic search approach inspired by the Darwinian principle of survival of the fittest. It will search for the best S_{ij} weights that maximize the correlation between RSCU and W_i while operating in Algorithm 1

```
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
```

Algorithm 1. The genetic algorithm operates to optimize the Sij weights used to calculate the tAI values.

Then, the best set of S_{ij} weights will be used to compute the tAI values using Eqs. 1, 2 (Sabi and Tuller, 2014).

2.4 Genomic data collection

The coding sequences of 12 organisms (*Ferroglobus placidus*, *Halomicrobium mukohataei*, *Methanocaldococcus jannaschii*, *Escherichia coli*, *Neisseria meningitides*, *Vibrio cholera*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Aspergillus fumigatus*, *Aspergillus nidulans*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*) used in this study as representatives of different domains were retrieved from NCBI (https://www.ncbi.nlm.nih.gov/genome/) in the FASTA format. Their tRNA gene copy numbers were obtained from GtRNAdb (Chan and Lowe, 2009). All information about the used organisms can be found in Supplementary Table S1.

2.5 CAI and the original tAI indices' calculation

The CAI was calculated using a Python package (Lee, 2018). In addition, the original tAI was calculated using a Python code developed by the authors (the same used to calculate the tAI in the gtAI package) using the S_{ij} weights found in the original study (Sabi and Tuller, 2014).

2.6 Protein abundance data collection

To test to what extent gtAI correlates with empirical data such as PA compared to the otAI and stAI, the PA data of *E. coli*, *C. elegans*, *D. melanogaster*, *S. cerevisiae*, and *S. pombe* were retrieved from PaxDB. The integrated PaxDB version (highest coverage) was used for all the organisms (version 4.1) (Wang et al., 2015). These organisms were chosen due to the availability of their PA data.

2.7 The impact of generation time and population size parameter choice on gtAI result reproducibility

First, we investigated the effect of the population size parameter on the gtAI result. Three random organisms were selected from the 12 used in this study (*S. cerevisiae*, *E. coli*, and *H. mukohataei*). For each organism, the non-parametric (Spearman) correlation between RSCU (of the reference set) and W_i values were optimized by the genetic algorithm used in gtAI calculation. We chose a constant generation time equal to 100 and different population sizes (10, 20, 30, n + 10 . . . , 100). Hence, each organism was optimized 10 times, each with a different population size to compare the best solution between each population size (inter-variability). This experiment

TABLE 1 Spearman's rank correlation analysis between CAI and the three tAI measurements (original tAI, stAI, and gtAI) for the 12 model organisms and their average GC content and ENc values.

		gtAI-CAI (rho)	stAI-CAI (rho)	tAI-CAI (rho)	GC content (%)	ENc
Archaea						
	Ferroglobus placidus	0.46*	0.48*	-	44.1	44.81
	Halomicrobium mukohataei	0.61*	0.41*	-	65.5	34.26
	Methanocaldococcus jannaschii	0.23*	0.14*	-	31.4	37.92
Bacteria						
	Escherichia coli	0.83*	0.82*	-	50.8	39.09
	Neisseria meningitidis	0.9*	0.67*	-	51.8	36.89
	Vibrio cholera	0.78*	0.8*	-	48.1	41.76
Eukarya (non-fungal)						
	Caenorhabditis elegans	0.8*	0.82*	-	35.4	42.02
	Drosophila melanogaster	0.89*	0.74*	-	42.0	38.69
Eukarya (fungal)						
	Aspergillus fumigatus	0.91*	0.78*	0.82*	49.5	40.64
	Aspergillus nidulans	0.94*	0.29*	0.91*	50.1	41.57
	Saccharomyces cerevisiae	0.94*	0.56*	0.87*	38.2	36.42
	Schizosaccharomyces pombe	0.88*	0.56*	0.84*	36	40.32

* represents p value <0.001.

was performed for each organism five times to reach the best solution within the same population size and for the same organism between different experiments (intra-variability). Then, we tested the best solution for selecting the suitable generation time by applying 1,000 generations on the same three organisms, with a constant population size equal to 60. Finally, we plotted the solutions from generation 1 to 1,000.

3 Results

3.1 CAI correlations with gtAI, stAI, and otAI

The Williams' test was used to compare the rho values at an alpha score of 0.01 (two-sided test). The gtAI values for *H. mukohataei*, *M. jannaschii*, *E. coli*, *N. meningitidis*, *D. melanogaster*, *A. nidulans*, *S. pombe*, *A. fumigatus*, and *S. cerevisiae* (9 out of 12) revealed a statistically significant (Williams' test p value <0.01) higher correlation with CAI than stAI (Table 1). Moreover, the gtAI in all the fungal organisms showed a higher considerable correlation (*A. nidulans*, *S. pombe*, *A. fumigatus*, and *S. cerevisiae*) with the CAI than the original tAI (Supplementary Tables S1–S4).

3.2 Repeated random sampling and correlations with the CAI

First, we calculated the gtAI, stAI, and CAI values for the genes of all tested organisms. Then, for each replicate (1,000 replicates with replacement), we sampled a 25% random sample size from the values of these measures for the Spearman's rank correlation analyses. This is to make sure that the reference set of genes present among other genes is not causing inflated gtAI-CAI correlations. The rep_sample_n R function from the infer package was used in the random sampling (https://www.rdocumentation.org/packages/infer/versions/1.0.4/topics/rep_sample_n). It does not specify a particular distribution type to be used but rather allows for repeated sampling of data from a specified data frame. The script of this random sampling method could be found here (https://github.com/AliYoussef96/gtAI/blob/master/random%20sampling.r).

The result showed that the gtAI exhibited stronger correlations for the same nine organisms compared to the stAI. Furthermore, a stronger correlation with the CAI in the four fungal organisms compared to the otAI is shown (Figure 1).

3.3 SCUO correlations with gtAI, stAI, and otAI

The SCUO is a codon usage index that does not involve the use of a reference set in its calculation (Wan et al., 2004). The gtAI, stAI, otAI, and SCUO values were calculated for all 12 organisms. The gtAI has outperformed both stAI and otAI by exhibiting a stronger statistically significant correlation with SCUO in eight organisms consistent with CAI association results except in *E. coli*, and the gtAI-SCUO and stAI-SCUO correlations are 0.26 and 0.27, respectively (Table 2). A two-sided Williams' test was used to compare the rho values at an alpha score of 0.01.

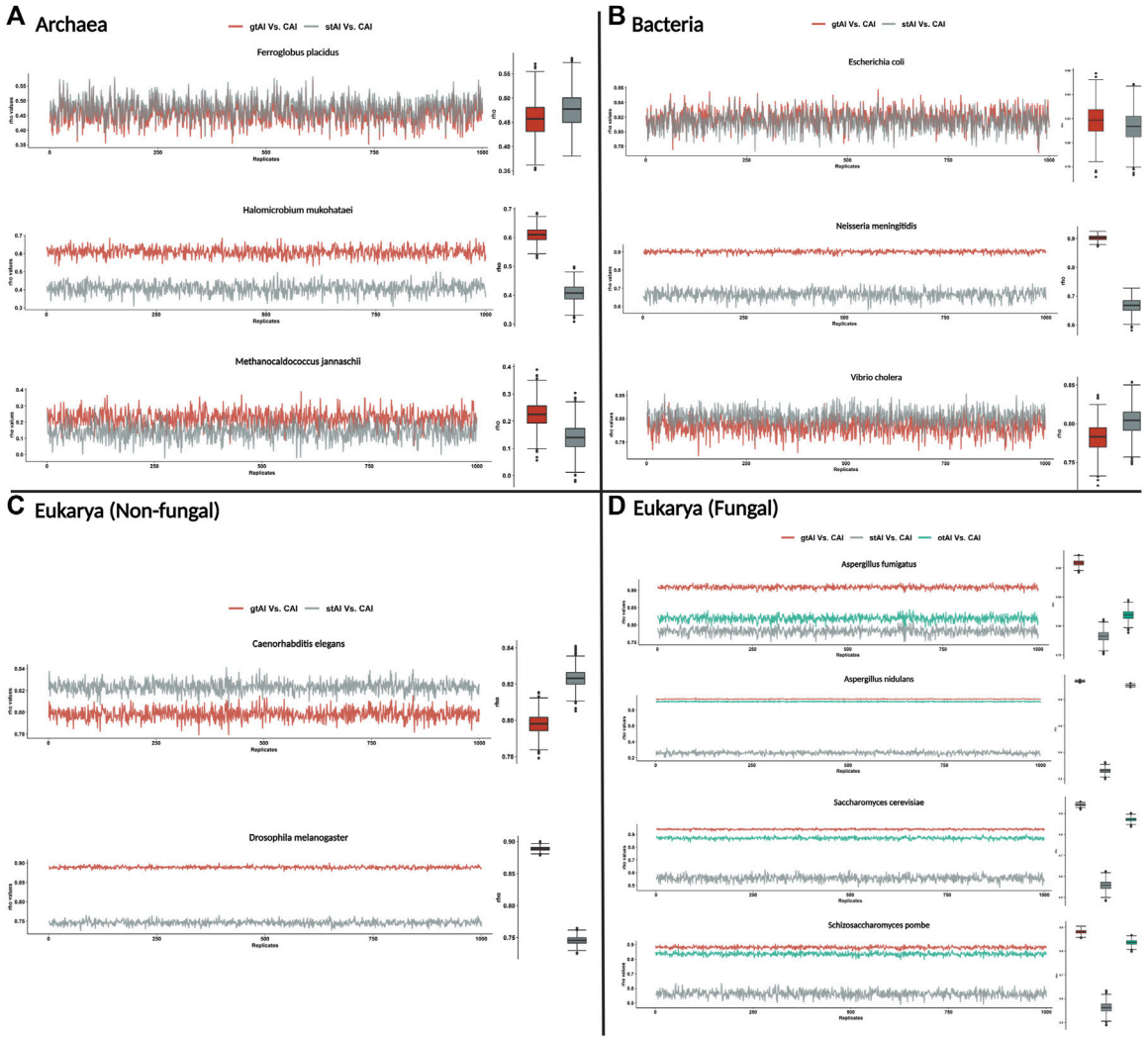

FIGURE 1
Repeated Spearman's rank correlation analysis between gtAI-CAI, stAI-CAI, and otAI-CAI for randomly sampled values. The gtAI, stAI, and CAI values for each organism (the 12 used in the study) were randomly sampled with size 25%, and 1,000 replicates with replacement have been taken. Spearman's rank correlation analysis between gtAI-CAI, stAI-CAI, and otAI-CAI for each replicate was applied. The line plot shows the rho values for each replicate, and the box plots show the distributions. **(A)** Three archaea organisms being tested in this study, **(B)** the three bacterial organisms, **(C)** two eukaryotic (non-fungal) organisms, and **(D)** four eukarya (fungal) organisms. The color code is red for gtAI-CAI correlation results, gray for stAI-CAI correlation results, and green for the original tAI (otAI)-CAI correlation results.

3.4 The gtAI correlates better with PA data than stAI and CAI in both fungal and non-fungal organisms

The Williams' test was used to compare the rho values at an alpha score of 0.01 (two-sided test). For *C. elegans*, *D. melanogaster*, *S. pombe*, *S. cerevisiae*, and *E. coli*, the gtAI showed a higher statistically significant correlation with PA than the stAI and CAI (Williams' test p value <0.01). Furthermore, the gtAI exhibits a higher statistically significant correlation with PA in *E. coli* than the original tAI (Williams' test p value <0.01). On the other hand, the original tAI predicted the PA of fungal organisms better than the gtAI and stAI, which is expected as it used experimental microarray data from yeast to obtain an optimal set of S_{ij} values that maximizes the correlation between expression levels and tAI values (dos Reis et al., 2004). Consequently, the gtAI is a valuable tool as it improves the prediction of PA in many organisms (Table 3).

3.5 The absolute adaptiveness values generated by the gtAI reflect the evolutionary proximity

The absolute adaptiveness (W_i) values of a codon depend on both the efficacy of codon–anticodon interaction (S_{ij} values) and the abundance of tRNA available for that codon. The number of tRNA genes and their abundance are diverse among the three domains of life (Fujishima and Kanai, 2014). Therefore, in theory, W_i should explain the divergence of organisms from different domains. To examine whether the W_i calculated using S_{ij} values generated by the

TABLE 2 Spearman's rank correlation analysis between SCUO and the three tAI measurements (original tAI, stAI, and gtAI) for the 12 model organisms.

		gtAI-SCUO (rho)	stAI-SCUO (rho)	tAI-SCUO (rho)
Archaea				
	Ferroglobus placidus	0.22*	0.24*	-
	Halomicrobium mukohataei	0.45*	0.27*	-
	Methanocaldococcus jannaschii	0.2*	−0.03	-
Bacteria				
	Escherichia coli	0.26*	0.27*	-
	Neisseria meningitidis	0.4*	0.22*	-
	Vibrio cholera	0.27*	0.28*	-
Eukarya (non-fungal)				
	Caenorhabditis elegans	0.26*	0.23*	-
	Drosophila melanogaster	0.6*	0.5*	-
Eukarya (fungal)				
	Aspergillus fumigatus	0.59*	0.49*	0.52*
	Aspergillus nidulans	0.48*	0.25*	0.45*
	Saccharomyces cerevisiae	0.41*	0.27*	0.36*
	Schizosaccharomyces pombe	0.22*	0.22*	0.20*

* represents p value <0.001.

TABLE 3 Spearman's rank correlation analysis between PA and the three tAI measurements.

	gtAI-PA (rho)	stAI-PA (rho)	tAI-PA (rho)	CAI-PA (rho)
Caenorhabditis elegans	0.38*	0.36*	-	0.28*
Drosophila melanogaster	0.48*	0.44*	-	0.33*
Escherichia coli	0.54*	0.53*	0.5*	0.52*
Saccharomyces cerevisiae	0.50*	0.49*	0.56*	0.49*
Schizosaccharomyces pombe	0.61*	0.54*	0.62*	0.53*

* represents p value <0.001.

gtAI are biologically meaningful, a hierarchical clustering on principal component (HCPC) analysis of W_i values was performed. The clustering classified all 12 model organisms used in the study into three distinct groups (Figure 2).

3.6 The effect of population size on gtAI result reproducibility

The inter-variability resulting from changing the population size in the three organisms was extremely low. For *S. cerevisiae*, *E. coli*, and *H. mukohataei*, the average best solution in the five experiments (the optimization of the non-parametric Spearman correlation between RSCU and W_i values) ranged from 0.8109 to 0.8134 (SD = from 0.0016 to 0.0031), from 0.6017 to 0.6022 (SD = from 0.001 to 0.0023), and from 0.4287 to 0.4308 (SD = from 0 to 0.0067), respectively. Then, the coefficient of variation (CV) for the same population size (10, 20, 30, n + 10, . . . , 100) was computed from the results of the five performed experiments for each organism. *S. cerevisiae* showed CV ranging from 0.00057 to 0.0042, *E. coli* ranging from 0.0012 to 0.0036, and *H. mukohataei* ranging from 0 to 0.0223. The coefficient of variation (CV) shows the extent of variability to the population's mean. Therefore, as the variability decreases, the CV approaches zero. The CV values for the tested genomes showed an extremely low intra-variability, approximately equal to zero CV. Therefore, the inter-variability and intra-variability resulting from choosing different population sizes in the gtAI algorithm will not influence the reproducibility of the results. However, we recommend choosing a constant population size for the whole analysis.

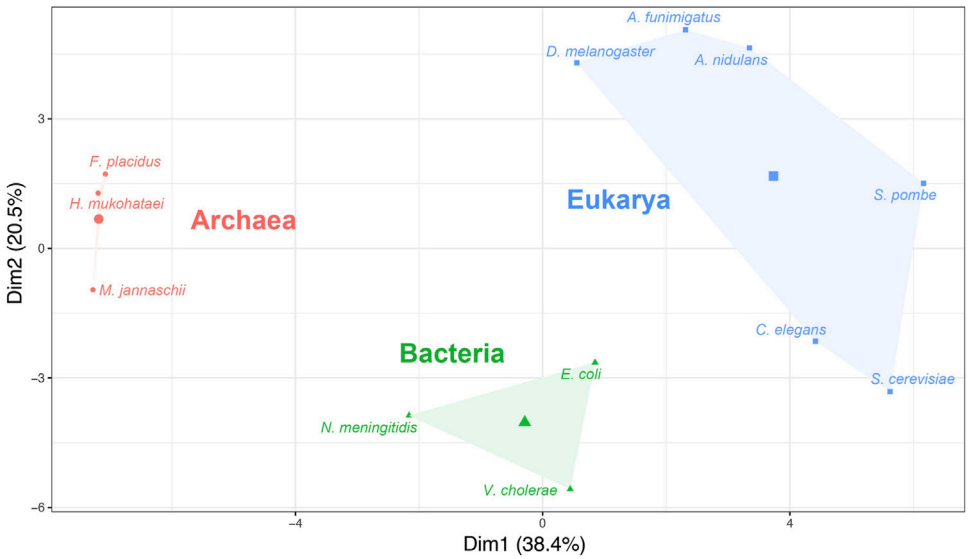

FIGURE 2
Hierarchical clustering on principal component (HCPC) analysis based on the absolute adaptiveness values (W_i) of the 12 model organisms. The *x*-axis and *y*-axis represent the first and second principal components (Dim1 and Dim2), respectively (the clustering was performed using Ward's method).

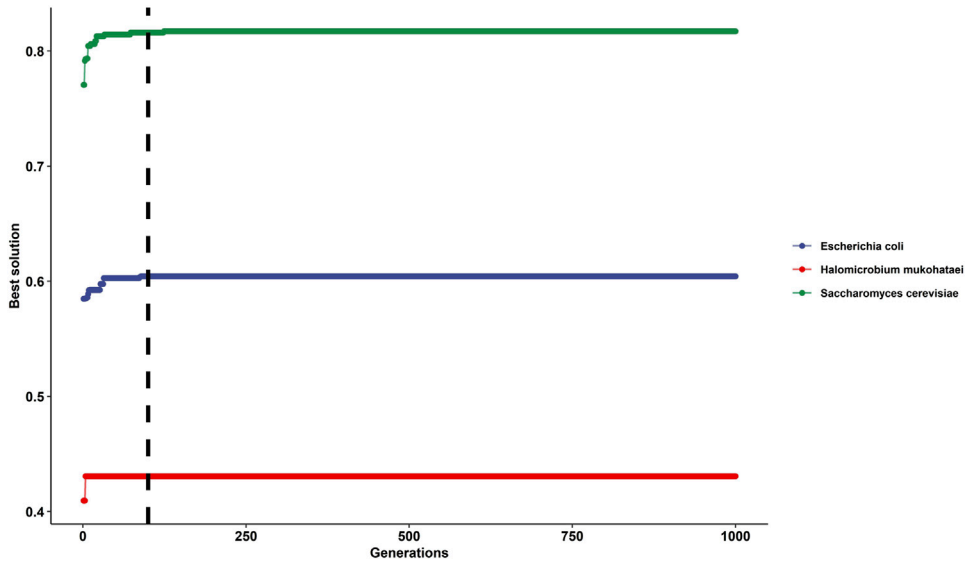

FIGURE 3
Impact of selecting different generation times for the genetic algorithm on gtAI results. The figure shows the best solutions after applying 1,000 generations in *Saccharomyces cerevisiae* (green line), *Escherichia coli* (blue line), and *Halomicrobium mukohataei* (red line). The *y*-axis represents the best solutions for optimizing the non-parametric (Spearman) correlation between RSCU (of the reference set) and W_i values at each generation. The *x*-axis represents the generation number—the vertical black line represents the generation time number 100 (the default generation time in the gtAI package).

3.7 The effect of generation time on gtAI result reproducibility

At generation time 100 and higher, the solution was constant or had shallow changes (Figure 3). Therefore, we recommend using a generation time of 100 or higher, but a constant generation time must be selected for the whole analysis.

4 Discussion

The tAI is a formal measure of the force of translational selection. It has been widely employed to investigate fundamental questions related to gene expression, molecular evolution, and virus–host adaptation (Sharp et al., 2005; Man and Pilpel, 2007; Goodman et al., 2013; Pechmann and Frydman, 2013; Sabi and

Tuller, 2014). Due to its importance, we were motivated to improve its performance by solving the issues of previous methods used for its calculation. We evaluated our proposed method mainly by examining whether it correlates better with other well-established codon usage indices. In addition, it shows a better association with empirical PA data.

The CAI is a gene-specific CUB index (Sharp and Li, 1987). Many studies suggested that the CAI is a good predictor of gene expression at mRNA and protein levels and has been used in many studies as a reference index to compare new indices and methods (Carbone et al., 2003; Sun et al., 2013; Fu et al., 2020). Accordingly, we conducted a comparative analysis on 12 model organisms (Supplementary Table S1) to evaluate the performance of the gtAI method compared to the original tAI and stAI (Supplementary Tables S2–S5) by examining whether it correlates better with the CAI using Spearman's rank correlation analysis. The gtAI managed to outperform both methods by exhibiting a stronger significant correlation in 9 out of 12 model organisms with the CAI.

Attempting to explain the reason behind the better association of the stAI with the CAI than the gtAI in *F. placidus*, *V. cholera*, and *C. elegans* revealed a notable conclusion. These three organisms showed the highest ENc (low CUB) average value within their domains. For example, the average ENc value of the reference set for *F. placidus* was 44.8, while in *H. mukohataei* and *M. jannaschii*, the ENc values were 34.25 and 37.91, respectively. The same trend was observed in the bacterial group, as the average ENc value for *V. cholera* was 41.83, 39.08 for *E. coli*, and 37.08 for *N. meningitidis*. For non-fungal eukaryotes, *C. elegans* exhibited a 42.02 average ENc value, while 38.69 for *D. melanogaster*. This shows one limitation in our approach which can be attributed to organisms with overall weak CUB. It slightly influenced the result of the gtAI leading to the better correlation of stAI in these organisms. Furthermore, insights into the relation between GC content and gtAI performance have revealed that the change in GC content could not explain the slight outperformance of the stAI over the gtAI in terms of correlation with the CAI except in archaeal genomes. To embark on, in non-fungal eukaryotes, though *C. elegans* has an average GC content of 35.4%, indicating a possible strong bias against GC-rich codons, it showed an overall relatively weak bias (ENc = 42.02) which resulted in the slight underperformance of gtAI compared to stAI. Additionally, though *D. melanogaster* has a relatively higher GC content of 42.0% (closer to 50%), indicating relatively weaker bias, it showed an overall relatively stronger bias (ENc = 38.69) than *C. elegans*. Meanwhile, in Archaea, it is notable that the change in their overall bias is consistent with their deviation from the 50% GC content. In other words, the archaeal genome that showed an underperformance of gtAI (*F. placidus*) has an average GC content of 44.1% which is the closest to 50% compared to the other two archaeal genomes of 31.4% and 65.5%, as well as the highest ENc value of 44.8 compared to the other two of 34.25 and 37.91. Without regard to this limitation, the results (Table 1) suggest that the gtAI method performance was better for S_{ij} value optimization, giving a more reliable tAI value that better demonstrates the effect of translational selection.

One can argue that the correlation between the gtAI and CAI might be inflated due to using the same reference set of genes. Hence, we conducted two more analyses to test it. In the first one, we obtained multiple random samples of CAI and the three tAI indices' values and performed Spearman's rank correlation analysis between each of them with the CAI. The results remained the same and agreed with our conclusion as the gtAI showed a stronger correlation with the CAI than the stAI in the same nine organisms. In addition, a higher correlation with the CAI than the original tAI in the four fungal organisms was shown. In the second analysis, we investigated whether gtAI correlates better with another CUB index that is independent of using a reference set of genes in its calculation, namely, SCUO (to exclude the reference set parameter). The results also revealed a stronger gtAI association with SCUO, further confirming that the correlation between the gtAI and CAI is not inflated nor affected by the reference set of genes.

In conclusion, our gtAI method can solve the query of S_{ij} value optimization and effectively estimate the tAI values while overcoming the limitations observed in other implementations. Performance evaluation showed that the gtAI method performed better than the original tAI and stAI by exhibiting a stronger correlation with the CAI and SCUO. It has also improved the prediction of PA compared to the stAI and CAI. The reproducibility of the genetic algorithm employed by the gtAI was tested and revealed its reliability in reaching the best solution in complex optimization problems. The *Wi* values generated by the gtAI correctly reflect the evolutionary proximity between organisms from different domains of life. Indeed, one significant advantage of CUB-dependent tAI computation methods (i.e., gtAI and stAI) over the original tAI is the lack of neediness for external information such as gene expression data or mRNA levels (which are often unavailable for most genomes). We believe that the gtAI will allow for obtaining higher quality tAI results used to draw conclusions about the force of translational selection acting on genes in related studies.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

AA: conceptualization, methodology, software, formal analysis, and visualization. SK: validation, investigation, writing—original draft. EA: writing—review and editing. AO: data curation. SE: data curation. AT: investigation. MS: investigation. MS: project administration, writing—review and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Egyptian Cancer Network (ECN), United States of America, and the Children's Cancer Hospital, Egypt 57357.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1218518/full#supplementary-material

References

Brandis, G., and Hughes, D. (2016). The selective advantage of synonymous codon usage bias in Salmonella. *PLoS Genet.* 12, 1005926. doi:10.1371/journal.pgen.1005926

Carbone, A., Zinovyev, A., and Képès, F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 19, 2005–2015. doi:10.1093/bioinformatics/btg272

Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006). Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* 7, 98–108. doi:10.1038/nrg1770

Chan, P. P., and Lowe, T. M. (2009). GtRNAdb: A database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* 37, 93–97. doi:10.1093/nar/gkn787

Crick, F. H. C. (1966). Codon—Anticodon pairing: The wobble hypothesis. *J. Mol. Biol.* 19, 548–555. doi:10.1016/S0022-2836(66)80022-0

Dana, A., and Tuller, T. (2014). The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res.* 42, 9171–9181. doi:10.1093/nar/gku646

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Res.* 32, 5036–5044. doi:10.1093/nar/gkh834

Duret, L. (2000). tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* 16, 287–289. doi:10.1016/S0168-9525(00)02041-2

Fu, H., Liang, Y., Zhong, X., Pan, Z., Huang, L., Zhang, H., et al. (2020). Codon optimization with deep learning to enhance protein expression. *Sci. Rep.* 10, 17617. doi:10.1038/s41598-020-74091-z

Fuglsang, A. (2003). Codon optimizer: A freeware tool for codon optimization. *Protein Expr. Purif.* 31, 247–249. doi:10.1016/s1046-5928(03)00213-4

Fujishima, K., and Kanai, A. (2014). tRNA gene diversity in the three domains of life. *Front. Genet.* 5, 142. doi:10.3389/fgene.2014.00142

Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., et al. (2003). Global analysis of protein expression in yeast. *Nature* 425, 737–741. doi:10.1038/nature02046

Goetz, R. M., and Fuglsang, A. (2005). Correlation of codon bias measures with mRNA levels: Analysis of transcriptome data from *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 327, 4–7. doi:10.1016/j.bbrc.2004.11.134

Goodman, D. B., Church, G. M., and Kosuri, S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. *Science* 342, 475–479. doi:10.1126/science.1241934

Gould, N., Hendy, O., and Papamichail, D. (2014). Computational tools and algorithms for designing customized synthetic genes. *Front. Bioeng. Biotechnol.* 2, 41. doi:10.3389/fbioe.2014.00041

Grantham, R., Gautier, C., Gouy, M., Moleculaire, E. E., Biome, L., De, I. U. L., et al. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* 8, 197. doi:10.1093/nar/8.1.197-c

Gustafsson, C., Govindarajan, S., and Minshull, J. (2004). Codon bias and heterologous protein expression. *Trends Biotechnol.* 22, 346–353. doi:10.1016/j.tibtech.2004.04.006

Han, J. H., Choi, Y. S., Kim, W. J., Jeon, Y. H., Lee, S. K., Lee, B. J., et al. (2010). Codon optimization enhances protein expression of human peptide deformylase in *E. coli*. *Protein Expr. Purif.* 70, 224–230. doi:10.1016/j.pep.2009.10.005

Huang, Y., Lin, T., Lu, L., Cai, F., Lin, J., Jiang, Y. E., et al. (2021). Codon pair optimization (CPO): A software tool for synthetic gene design based on codon pair bias to improve the expression of recombinant proteins in Pichia pastoris. *Microb. Cell. Fact.* 20, 209–210. doi:10.1186/s12934-021-01696-y

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.* 151, 389–409. doi:10.1016/0022-2836(81)90003-6

Karlin, S., Mrázek, J., Campbell, A., and Kaiser, D. (2001). Characterizations of highly expressed genes of four fast-growing bacteria. *J. Bacteriol.* 183, 5025–5040. doi:10.1128/JB.183.17.5025-5040.2001

Lee, S. (2018). Python implementation of codon adaptation index. *J. Open Source Softw.* 3, 905. doi:10.21105/joss.00905

Liu, B., Kong, Q., Zhang, D., and Yan, L. (2018). Codon optimization significantly enhanced the expression of human 37-kDa iLRP in *Escherichia coli*. *3 Biotech.* 8, 210. doi:10.1007/s13205-018-1234-y

Man, O., and Pilpel, Y. (2007). Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat. Genet.* 39, 415–421. doi:10.1038/ng1967

Oymondal, U. R., As, S. D., and Ahoo, S. S. (2009). Predicting gene expression level from relative codon usage bias: An application to *Escherichia coli* genome. 13–30.

Pechmann, S., and Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* 20, 237–243. doi:10.1038/nsmb.2466

Plotkin, J. B., and Kudla, G. (2011). Synonymous but not the same: The causes and consequences of codon bias. *Nat. Rev. Genet.* 12, 32–42. doi:10.1038/nrg2899

Quax, T. E. F., Claassens, N. J., Söll, D., and van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. *Mol. Cell.* 59, 149–161. doi:10.1016/j.molcel.2015.05.035

Raguin, A., Stansfield, I., and Romano, M. C. (2023). ExpressInHost: A codon tuning tool for the expression of recombinant proteins in host microorganisms. *J. Open Res. Softw.* 11, 385. doi:10.5334/jors.385

Russell, S., and Norvig, P. (2010). *Artificial intelligence: A modern approach*. Third. Upper Saddle River, NJ: Prentice Hall.

Sabi, R., and Tuller, T. (2014). Modelling the efficiency of codon-tRNA interactions based on codon usage bias. *DNA Res.* 21, 511–525. doi:10.1093/dnares/dsu017

Satya, R. V., Mukherjee, A., and Ranga, U. (2003). A pattern matching algorithm for codon optimization and CpG motif-engineering in DNA expression vectors. *Comput. Syst. Bioinforma.* 2, 294–305. doi:10.1109/CSB.2003.1227330

Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., and Sockett, R. E. (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* 33, 1141–1153. doi:10.1093/nar/gki242

Sharp, P. M., and Li, W. H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281–1295. doi:10.1093/nar/15.3.1281

Sharp, P. M., and Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* 24, 28–38. doi:10.1007/BF02099948

Sun, X., Yang, Q., and Xia, X. (2013). An improved implementation of effective number of codons (Nc). *Mol. Biol. Evol.* 30, 191–196. doi:10.1093/molbev/mss201

Thengade, A., and Dondal, R. (2012). "Genetic algorithm – survey paper," in MPGI National Multi Conference International Journal of Computer Applications, 2012, 975–8887. Available at: https://www.ijcaonline.org/proceedings/ncrtc/number5/6549-1039.

Wan, X. F., Xu, D., Kleinhofs, A., and Zhou, J. (2004). Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol. Biol.* 4, 19. doi:10.1186/1471-2148-4-19

Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 15, 3163–3168. doi:10.1002/pmic.201400441

Watts, A., Sankaranarayanan, S., Watts, A., and Raipuria, R. K. (2021). Optimizing protein expression in heterologous system: Strategies and tools. *Meta Gene* 29, 100899. doi:10.1016/j.mgene.2021.100899

Wright, F. (1990). The "effective number of codons" used in a gene. *Gene* 87, 23–29. doi:10.1016/0378-1119(90)90491-9

Yang, X. S. (2014). "Nature-inspired optimization algorithms," in *Nature-inspired optimization algorithms* (Oxford: Elsevier). doi:10.1016/B978-0-12-416743-8.00017-8