DNA Research, 2017, 24(4), 419–434
doi: 10.1093/dnares/dsx014
Advance Access Publication Date: 24 April 2017
Full Paper

Full Paper

Integrated analysis of individual codon contribution to protein biosynthesis reveals a new approach to improving the basis of rational gene design

Juan C. Villada[1], Otávio José Bernardes Brustolini[2], and Wendel Batista da Silveira[1,]*

[1]Department of Microbiology, Universidade Federal de Viçosa, Viçosa 36570-900, Brazil, and [2]Department of Biochemistry and Molecular Biology, Universidade Federal de Viçosa, Viçosa 36570-900, Brazil

*To whom correspondence should be addressed. Tel: +55 31 3899 2957. Fax: +55 31 38992573.
Email: wendel.silveira@ufv.br

Edited by Prof. Kenta Nakai

Received 24 July 2016; Editorial decision 22 March 2017; Accepted 23 March 2017

Abstract

Gene codon optimization may be impaired by the misinterpretation of frequency and optimality of codons. Although recent studies have revealed the effects of codon usage bias (CUB) on protein biosynthesis, an integrated perspective of the biological role of individual codons remains unknown. Unlike other previous studies, we show, through an integrated framework that attributes of codons such as frequency, optimality and positional dependency should be combined to unveil individual codon contribution for protein biosynthesis. We designed a codon quantification method for assessing CUB as a function of position within genes with a novel constraint: the relativity of position-dependent codon usage shaped by coding sequence length. Thus, we propose a new way of identifying the enrichment, depletion and non-uniform positional distribution of codons in different regions of yeast genes. We clustered codons that shared attributes of frequency and optimality. The cluster of non-optimal codons with rare occurrence displayed two remarkable characteristics: higher codon decoding time than frequent–non-optimal cluster and enrichment at the 5′-end region, where optimal codons with the highest frequency are depleted. Interestingly, frequent codons with non-optimal adaptation to tRNAs are uniformly distributed in the *Saccharomyces cerevisiae* genes, suggesting their determinant role as a speed regulator in protein elongation.

Key words: codon usage bias, microbial biotechnology, position-dependent codon usage, rational gene design, yeast genomics

1. Introduction

Codon usage bias (CUB) has been a wide-ranging field of research in the last few decades, revealing the importance of codons in many biological processes from cell physiology to the evolution of genes and organisms.[1–9] A number of them are related to genic regulation,[10,11] folding energy of mRNA secondary structures,[12–14] mRNA stability,[15–17] alternative splicing,[18,19] miRNA–mRNA interaction,[20,21] and protein aggregation or co-translational folding.[22–27]

Consequently, synonymous mutations affect significantly both gene expression and protein level, which are determinant steps in heterologous expression.[28]

A number of traditional approaches have been widely used for the computational assessment of CUB based on measuring codon occurrence, such as Relative Synonymous Codon Usage (RSCU[29]) which is an index for evaluating the relative frequency of codons and determining when a codon is preferentially selected over its synonymous codons. The Codon Adaptation Index (CAI[2]), is another method used for determining the potential adaptability of a gene to a host genome based on frequency of codon usage. On the other hand, the tRNA Adaptation Index (tAI[5]) is based on translation efficiency as measured by the codon–tRNA interaction, which calculates the optimality of a codon to the cognate tRNAs. This method has been broadly used to describe the 'ramp theory' of translation efficiency.[30] Thus, many studies have been based on codon indexes to improve protein production through codon optimization in microorganisms.[31–34] Nevertheless, the introduction of synonymous mutations has sometimes resulted in a decrease in protein production[35] or loss of protein function due to instability, misfolding or aggregation.[24,36,37] Contrary to earlier codon optimization studies, recent works have demonstrated enhanced protein production by insertion of rare codons,[38] and, indeed, by random codon substitutions at the 5′-end region of genes.[39]

Despite significant advances in CUB research, the understanding of the role of individual codons remains a challenging problem in the production of recombinant proteins,[40] especially in yeast, where the limitations of codon optimization indexes have been reported years ago.[41] We believe that this drawback is related mainly to a lack of integrated studies that could combine different approaches into one integral explanation of the effects of individual codons on protein biosynthesis. Although considerable advances have been made in accounting for the role of codons in certain processes, an integrated study is still required to explain the implications of codons at different biological levels and integrate them into a framework that could improve in depth the understanding of CUB and its basis in rational gene design.

Another important question to be addressed in yeast CUB research is the determination of which codons are under evolutionary selective pressures at different positions in the genes. Recently, Hockenberry et al.[42] developed a method of codon quantification which was implemented in the *Escherichia coli* genome to illustrate codon deviations from uniformity as a function of their position within genes. This approach is a major step to providing support for the ramp theory and codon selection at the 5′-end of endogenous genes. However, we observed that this method is not accurate enough for comparing codons at regions distant from the start codon. To overcome this drawback, we developed a method for quantifying all the genome regions based on a binning scheme of codon quantification relative to coding sequence (CDS) length. Thus, beyond the integrated scheme to evaluate CUB as a function of the position in yeast CDSs, the present study proposes the use of a different approach to the quantification of codons in yeasts using genome-scale data.

Further, we examine various features of individual codons for five yeasts widely used in heterologous expression and value-added chemical production. We integrate the results to show the potential effects of each codon in protein biosynthesis, evidencing that position-dependent CUB is a governing rule over endogenous yeast CDSs. We show the interdependencies between the genome-scale frequency of codons and their optimality to translation efficiency. Additionally, we demonstrate that there are very specific non-optimal codons that are, necessarily, uniformly distributed in intragenic regions of genes, while other kinds of non-optimal codons—having higher decoding times—are enriched at the beginning of genes and depleted in distant regions. Finally, we exemplify the potential impacts of individual codons in protein secondary structures. Taken together, our framework and the subsequent analyses present a meaningful advance in improving the basis of rational gene design for heterologous expression and yeast synthetic biology.

2. Material and methods

2.1. Criteria for yeast selection

Because we were interested in analysing the CUB in yeasts related to biotechnological processes, principally as cell factories, we counted the yeast species names reported in the 'Yeast biotechnology' article collection (http://www.microbialcellfactories.com/series/Yeast%20Biotechnology), available in the Microbial Cell Factories journal edited by Prof. Diethard Mattanovich (last updated on 8 May 2015). The yeasts which were chosen for subsequent analysis were those with the most related publications. The yeasts selected were *Kluyveromyces lactis*, *Kluyveromyces marxianus*, *Pichia pastoris*, *Saccharomyces cerevisiae* and *Yarrowia lipolytica*.

2.2. Genome sequences

The genomic sequences of *Kluyveromyces lactis* (NRRL Y-1140: NC_006037–NC_006042), *Pichia pastoris* (GS115: NC_012963–NC_012966), *Saccharomyces cerevisiae* (S288c: NC_001133–NC_001148) and *Yarrowia* lipolytica (CLIB122: NC_006067–NC_006072) were obtained from the Fungi section of the NCBI ftp server (accessed in January 2015). The genomic sequence of *Kluyveromyces marxianus* (CCT 7735: CP009303–CP009310*)* and *Escherichia coli* (K-12: NC_000913) were downloaded from the Nucleotide database of the NCBI (accessed in February 2015).

It has been demonstrated in previous studies that signal peptides (SPs) introduce a different bias of codon usage in the 5′-end gene region.[26,43–44] Based on this, we decided to analyse CDSs containing SPs separately from those lacking SPs. Accordingly, the presence or lack of proteins with SPs were determined using SignalP-4.1[45] with the parameter $t = euk$ for yeast sequences and, $t = gram$- for *E. coli*. Default values were used for other parameters.

2.3. Characterization of codons

2.3.1. Codon adaptation

In order to analyse the adaptation of each CDS to the codon usage of its respective genome, we used the Codon Adaptation Index (CAI[2]). First, the Codon Usage Table (Supplementary Tables S3–S7) of each genome was calculated by using the *cusp* function in the local version of EMBOSS suite v.6.6.0.0.[46] Then, using the *cai* function of EMBOSS suite, the CAI was computed for each coding sequence with the correspondent codon usage table of its genome.

2.3.2. The frequency of codon usage

The relative synonymous codon usage (RSCU) is significant to the analysis of codon bias in terms of frequency. An important advantage of this index is its independence from amino acid composition bias.[47] Then, the RSCU value of each codon was calculated for all organisms. The RSCU value was determined by using Equation (1) from a script in R as follows:

$$RSCU = \frac{O_{ij}}{\left[\sum_{j}^{Ni} O_{ij}\right] \times \frac{1}{Ni}} \quad (1)$$

where O_{ij} is the frequency of the *j*th codon for the *i*th amino acid, and *Ni* the total number of synonymous codons coding for the *i*th amino acid. Hence, as designated by Nasrullah et al.,[48] a frequent codon will have an RSCU ≥ 1 and codons with RSCU < 1 are qualified as rare.

2.3.3. Translation efficiency and codon decoding time

The translation efficiency of a codon was obtained from the tRNA Adaptation Index (tAI[5]), where codon relative adaptiveness (*w*) is the adaptation of a codon to the pool of available cognate tRNAs in the genome which incorporates the different possible tRNA and the wobble pairing rules. Based on reported methods,[5] we calculated the values for *K. marxianus* and *P. pastoris* of the relative adaptiveness of a codon using CodonR including the parameters for eukaryotic microorganism analysis. The datasets of relative adaptiveness of each codon of *S. cerevisiae* were obtained from reported data.[30] In the case of *K. lactis* and *Y. lipolytica* they were retrieved from a different experiment.[27] When needed, tRNA counts were predicted locally using tRNAscan 1.4.[49] As described by Pechmann and Frydman,[27] codon relative adaptiveness (*w*) characterizes a codon as optimal if $w \geq 0.4$ or non-optimal if $w < 0.4$.

Codon decoding time (CDT) is a value related to the abstract time required to translate a codon in relation to the codon relative adaptiveness of the tAI index. Thus, as defined by Dana and Tuller,[50] codons with low relative adaptiveness to the tRNA pool are more slowly translated. The CDT for each codon was calculated using Equation (2) and is defined as:

$$CDT_i = \frac{1}{w_i} \quad (2)$$

2.4. Genome simulations

For testing spatial deviation from uniformity in codon bias, 1200 complete genomes in total, 200 for each one of the 6 organisms in this study, were simulated. We developed an R script (*SyMuGS*—see Availability section and Methods Appendix for details) under the *seqinr* package[51] to generate genomes containing coding sequences with the same codon usage, coding the same protein sequence as the original genome but scrambling the codon position in each coding sequence. This provides a robust null model to test the deviations of the original genome against the random distribution of codons in the simulated genomes.[42]

2.5. Quantification of position-dependent codon usage bias

We developed two R scripts (*QuantiCUB* and *ExVar3D*, see Availability section and Methods Appendix) under R/Bioconductor (http://www.R-project.org/) and Biostrings[52] to count the codon occurrence under a relative-to-CDSs-length binning scheme (Fig. 1). We noticed that the codon quantification method proposed in a previous work[42] is an appropriate approach to testing deviations from uniformity at initial intragenic regions but not to test subsequent positions, because of the exponential growth of the bin sizes in regions far from the start codon. When the codon position is more distant, the procedure becomes imprecise as regards retrieving codon quantification in regions in the intermediate and the 3′-end region of CDSs. Hence, it is not possible to keep codon quantification values in the same bin as they are part of different intragenic region of CDSs with dissimilar lengths. Thus, we decided to quantify the CUB by binning codons as a function of the position relative to the length of the CDSs. We integrated all the scripts developed into a toolset named '*CodG: Analyzing codon positional dependency from genome-scale data*' (See Methods Appendix in Supplementary Material and https://github.com/juanvillada/CodG).

First, the start and stop codons were removed from all sequences and, in order to maintain uniformity in the length distribution of the different CDS sequences between the five yeasts, sequences with less than 40 codons were excluded from the dataset (i.e. sequences excluded for *K. lactis* = 0.02%, *K. marxianus* = 0%, *P. pastoris* = 0%, *S. cerevisiae* = 0.7% and *Y. lipolytica* = 0.1%). Thus, following the results of a recent report,[42] where it was demonstrated that no significant differences were detected when using different bin sizes, and guided by the result that shows ten as the best number of sections for visual exploration of the codon quantification, our method divides each gene into ten parts and saves the information of codon quantity per each tenth part of the CDSs thereby establishing ten bins with the same total number of codons. For each genome, the algorithm generated a matrix of 59 codons as rows (excluding the start, stop, and tryptophan codon counts) and 10 columns as bins.

Finally, to test the effectiveness of our method, we used the *E. coli* genome as a positive control (Supplementary Fig. S2), comparing it to previous reports.[42] As expected, comparable quantification values at the 5′-end were retrieved, and additionally, deviations from uniformity were detected at the 3′-end, a result which had not been observed by using a previous binning scheme.[42]

2.6. Test for uniformity in CUB

After constructing the matrix of codons per bins, we tested the uniformity of distribution of each codon by calculating the χ^2 value[42] from Equation (3) as follows:

$$\chi^2 = \sum_{i=1}^{n} Z^2 = \sum_{i=1}^{n} \frac{(O-E)^2}{\sigma^2} \quad (3)$$

where *O* is the observed count per bin in the original genome, *E* the expected value and σ the standard deviation of codon counts per bin obtained from the 200 simulated genomes, *n* the number of bins and *z* the *z*-score of each codon per bin. The squared *z*-score is the value related to deviations from uniformity, it tests the selection for or against uniformity in codon distribution as a function of intragenic position.

In order to test the enrichment or depletion of codons as a function of the position within a CDS, we used the Y-value as described mathematically in Equation 4:

$$Y = Log_2\left(\frac{O}{E}\right) \quad (4)$$

when *O* (observed value) is higher than *E* (expected value), it indicates codon enrichment at that intragenic position and the resulting value is positive. On the contrary, a depletion is given by negative values.

2.7. Codon conservation in protein structures

To understand how a similar CAI of a gene could affect translation optimization, we analysed the sites where codon categories are conserved in the structure of proteins in a similar way as that reported in

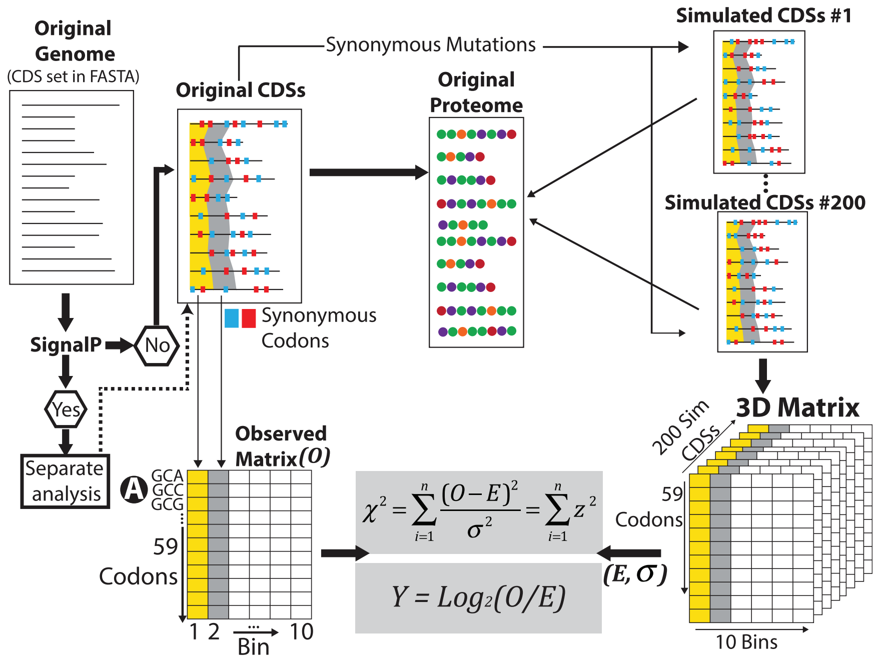

Figure 1. Quantification strategy of codon usage bias by bins relative to coding sequences' length. From each original genome, two datasets were arranged, the first including only CDSs lacking signal peptides, and the second comprising only CDSs codifying proteins with signal peptides. Then, for both datasets, all CDSs were divided into ten parts (bins) and their codons quantified and saved in bins (Observed Matrix). Synonymous mutations were introduced, conserving the original codon usage of the genome but scrambling codon positions within genes, thus generating 200 whole simulated genomes per yeast. Codons were quantified by bins as stated in the 'Material and methods' section, and a 3D matrix of 10 × 59 × 200 was generated in order to retrieve the expected value and standard deviation for each codon after codon position alteration.

the literature.[27] We examined, as a proof of concept, two proteins conserved in the five yeasts which have similar values of CAI, but different expression levels under physiological conditions, namely, the elongation factor 1-alpha (TEF1a) which is highly expressed, and the heat-shock protein SSE1 which has low expression level. We retrieved the protein homologous sequences from UniProt[53] where available, or determined them by *blast* when not. The 3D structures were obtained from the Protein Data Bank, 1F60 structure for EF1A and 2QXL structure for SSE1. The structures were visualized, and their colours were edited using PyMOL (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.). A multiple sequence alignment of amino acids and codons was formulated using ClustalW[54] and edited in MEGA 6[55] in order to avoid gaps and maintain the integrity of the *S. cerevisiae* sequence as the model. The datasets P02994 for EF1A and P32589 for SSE1 from UniProtKB were used as a reference for the secondary structure motifs and 3D structural features.

To test whether codon conservation in protein secondary structures is a generalized phenomenon, we used the data from the PDB web site database (http://www.rcsb.org/pdb). The files were retrieved by advanced search using *Saccharomyces cerevisiae* as the organism source. The proteins were ordered according to their resolution and we recovered the 300 best resolution files (Supplementary Dataset S2—PDB IDs/SGDIDs). Then we removed the redundancy of these files applying the program VAST: Vector Alignment Search Tool.[56] The set-up parameters were the *P*-value of 10e−7 and display of "PDB codes only". The outcome was 138 PDB files/proteins tagged as non-redundant PDB data set (Supplementary Dataset S2—VAST). To automatize this process we used an in-house script (Struct_*nonredund.cc*, see Availability section).

In the PDB file format documentation (http://www.wwpdb.org/documentation/file-format (January 2017, date last accessed)), there are only two secondary directly specified structures: Helix and Sheet. Thus, the analysis of codon proportion homogeneity within helix and sheet structures was performed by Chi-square test using two frequency tables, one for Helix (Supplementary Dataset S2—Helix) and another for Sheet (Supplementary Dataset S2—Sheet) created by our in-house C++ script (*Struct_frequencies.cc*, see Availability section). Finally, we tested codon proportionality through a position-dependent scheme which analysed the occurrence of codon categories at the first, second and third positions within each protein structure using an in-house script (*Struct_freq2.cc*). Next, the same analysis was applied to the $n-2$, $n-1$ and n, where n is the last codon of each structure per protein. Finally, as in PDB documentation there is not a defined coil region, we performed the coil analysis using the regions which do not belong to Helix and Sheet coordinates present in the non-redundant PDB files. The chi-square tests were also applied to the data retrieved from the coil region, i.e. loops, which connects helices and sheets. The scripts used for coil analyses can be found in Availability section (*Struct_freq2_coil.cc* and *Struct_freq3_coil.cc*).

2.8. Computation of mRNA minimum free energy

To evaluate the positional dependency of minimum free energy (MFE) in the mRNA secondary structures, we calculated the MFE in different positions of the genes by using the RNAfold program of the ViennaRNA package 2.0.[57] The MFE was computed for all the CDSs of each yeast ($n_{K.\ lactis} = 5065$, $n_{K.\ marxianus} = 4774$, $n_{P.\ pastoris} = 5019$, $n_{S.\ cerevisiae} = 5786$, $n_{Y.\ lipolytica} = 6413$). Ribosomal genes

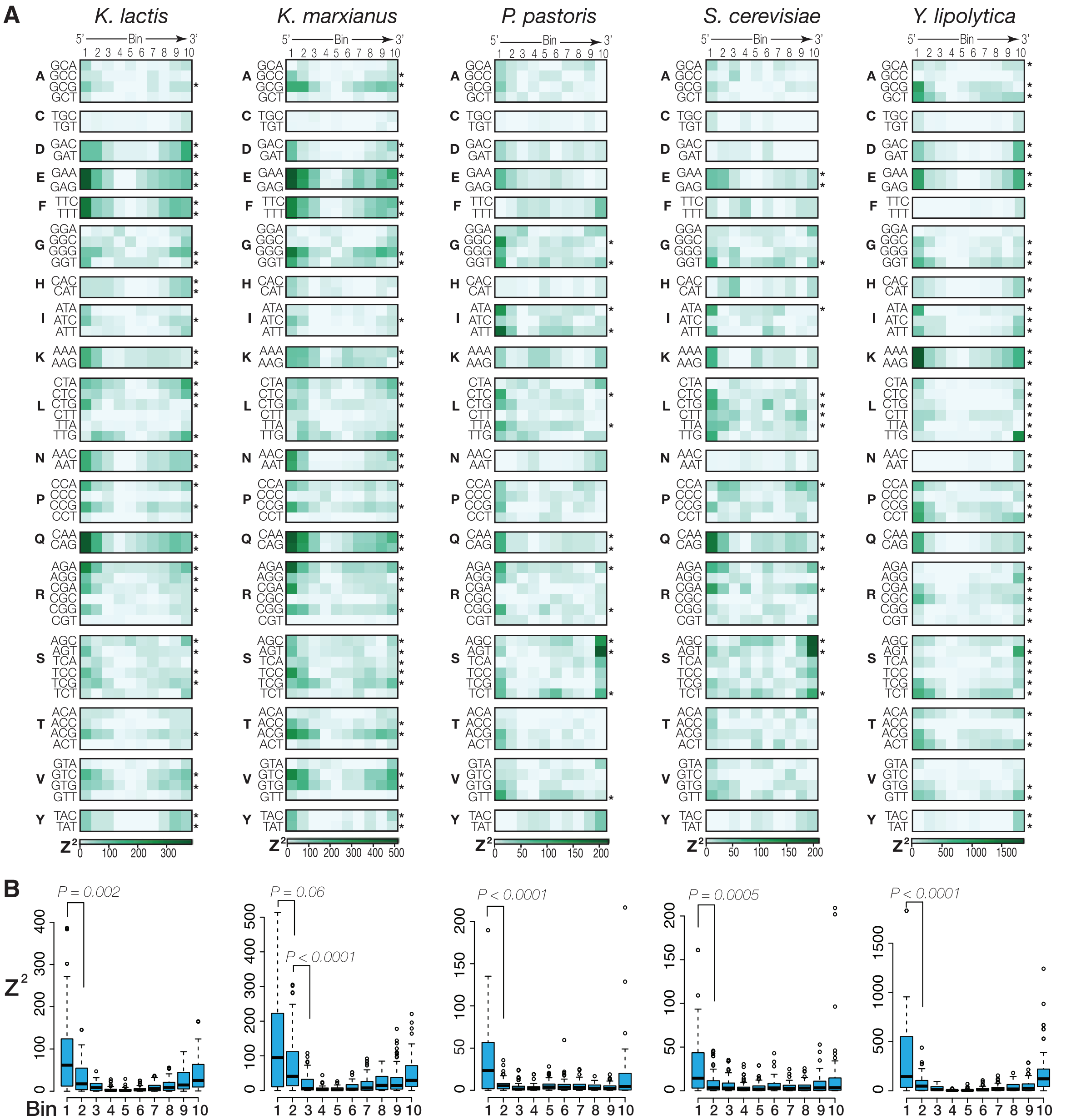

Figure 2. The positional dependency of codon usage bias relative to coding sequences length. The figure shows the analysis applied to the set of coding sequences lacking signal peptides. (A) Deviations from uniformity (squared *z*-scores) are reported graphically to illustrate the codon usage deviations from uniformity at different intragenic regions, concentrated mainly at the 5′-end close to the start codon. Squared *z*-scores are presented according to the quadratically scaled bar. (B) Distribution of deviations from uniformity by bin, illustrating the differences between first bin (5′-end region) and subsequent bins. The Mann–Whitney test was applied to determine the significance of the differences.

from each yeast genome were used as a subset of highly translated transcripts.

2.9. Statistical tests

The *P*-values (Figs 2B and 5A) are the probabilities related to the Wilcoxon rank sum test (Mann–Whitney test), a non-parametric test for expected values with a two-sided alternative hypothesis for independent samples. For χ^2 calculations, we report the significance of each codon by a *P*-value < 0.00017 after a Bonferroni correction for 59 tests where *P*-value $= 0.01$ contrasting them with the χ^2 distributions with $n - 1$ degrees of freedom.

3. Results and discussion

3.1. Yeast cell factories

The metabolic diversity encoded in the yeast's genome has led the scientific community to explore its capability at the industrial level as cell factories to produce value-added chemicals and enzymes. To determine the codon usage particularities in yeasts of biotechnological relevance, we used data mining to retrieve the most researched yeast species using as database all the articles in the collection of "Yeast Biotechnology" from the Microbial Cell Factories journal. It was observed that 104 papers, relating to 15 different yeast species, have been published in the last four years (Supplementary Table S1).

However, most of them were limited to five species: *Kluyveromyces lactis*, *Kluyveromyces marxianus*, *Pichia pastoris*, *Saccharomyces cerevisiae* and *Yarrowia lipolytica*. It should be noted that there was an increase in the proportion of works published on non-conventional yeasts compared with the conventional *S. cerevisiae* (Supplementary Fig. S1). This being the case, we decided to work with the aforementioned five species.

3.2. Strong signals of non-uniform distribution of codons relative to CDS length

We evaluated the positional dependency of CUB, integrating recent findings which state that gene expression levels in eukaryotes are shaped by CDS length and CUB,[58–63] that also supported our scheme in order to overcome the drawback experienced by using a previous method,[42] i.e. codon comparison between genic regions distant from the start codon. We decided to create a binning scheme based on the quantification of codons as a function of their relative position within genes and in relation to the length of their CDSs, forming ten bins by position (Fig. 1) to identify increased signals of deviations from uniformity [squared *z*-score from Equation (3)] as a function of the relative intragenic position.

Previous studies have evaluated CUB, using bins with fixed lengths to analyse genes of both prokaryotes[42] and eukaryotes.[30] These studies reported that the 5′-end region presented unusual CUB behaviour. A detailed inspection of the results presented in the Tuller *et al.* manuscript,[30] shows that the *E. coli* genome had substantial variation in standard deviations in the 3′-end of the averaged translation efficiency (as measured by the tRNA adaptation index). However, in a recent study, significant signals of deviations from uniformity (given by the *z*-score of the chi-square test) were not detected in the 3′-end of *E. coli*.[42] Even though each is a different measurement of CUB, we would expect that, under a positional dependency approach, deviations from uniformity would also be significant at the 3′-end region as codons contribute differentially to the translation efficiency.

We consider that this drawback may be due to the use of the fixed length of bins for deviations from uniformity computation, since the algorithm that uses the fixed length of bins is not capable of comparing the equivalent region of two CDSs (i.e. neither the middle nor the 3′-end) when the difference in CDS length is great. For example, if we have one CDS (CDS-A) with 200 codons and another with 100 codons (CDS-B), in a fixed configuration, we will be doing a comparison between the middle region of CDS-A and the 3′-end region of CDS-B which will give an inconsistent result. It is important to point out that our proposed binning scheme is aimed at overcoming this drawback.

Bearing this in mind, we hypothesized that, based on the correlation of both CUB and CDS length, increased selection signals against the uniformity of codon distribution at different positions on the genes would be observed, taking into account the relative position of codons by gene, instead of an absolute pre-established quantity of codons per bin. In order to test this, we implemented our method and used published results[42] as the positive control, drawing on research by authors who analysed the *Escherichia coli* genome and quantified high selection values against the uniformity of many codons at the 5′-end of the genes. As expected, by using our proposed codon quantification scheme on the same *E. coli* genome, we found comparable patterns of codon deviations from uniformity at the 5′-end (Supplementary Fig. S2). Nevertheless, it is important to point out that we also obtained significant deviations from uniformity in a number of codons at the 3′-end of genes, which had not been detected by the previously reported binning scheme. Thus, as shown in the positive control, our method allows for the detection of deviations from uniformity at both the 5′- and 3′-gene extremities.

The strong selection against uniformity detected at both gene extremities contrasts with the absence thereof in intragenic regions (distant from the extremities) of the CDSs. Considering that selection forces on CUB are determinant features of protein biosynthesis,[64] our proposed method is useful for describing the codon distribution patterns in endogenous genes and consequently further codon optimization. These results are explained principally by the accuracy of our binning scheme which is able to retrieve the same quantity of codons per bin, without binning bias at the 3′-, 5′- and all other intragenic positions. Therefore, our method exposes the differences, in terms of codon bias, between both ends and the inner regions of the CDSs, supporting recent findings described through ribosomal-protected footprint counts,[65] where it was observed that codons have different footprint count distributions in a position-dependent profile, conserved along intragenic regions but divergent at the 5′- and 3′-end. Notably, a similar profile was observed when our genome-scale codon quantification method on the *E. coli* genome (Supplementary Fig. S2) and on five yeast species (Fig. 2A and B) was applied.

3.3. The positional dependency of the genome-scale CUB in yeast

Using our codon quantification method, which had been previously tested on the *E. coli* genome, we evaluated the position-dependent CUB in the yeast cell factory genomes. For this, a chi-square test per codon [Equation (3)] was conducted after filtering out the genes encoding proteins with signal peptides (SPs) in order to avoid biases attributable to their particular CUB.[26,43,44,66] In a subsequent experiment, the same equation and methods were used to detect deviations from uniformity in the CDSs containing SPs. The occurrence of codons by bin was counted in the original (observed) genome and then compared with a null model formed by the expected value and standard deviation of 200 simulated genomes, where each genome conserved overall gene CUB but is produced with random codon positional distribution (see details in Material and Methods). This allowed us to determine the significance of the original CUB genome against a null model, calculating the squared *z*-score (deviations from uniformity) as a function of codon position and CDS length.

Out of 59 redundant codons (Fig. 2A), we found 37 in *K. lactis*, 39 in *K. marxianus*, 14 in *P. pastoris*, 16 in *S. cerevisiae* and 50 in *Y. lipolytica* with significant non-uniform distribution in terms of their position (P-value < 0.00017, chi-square test after Bonferroni correction; Supplementary Table S2). These codons are highlighted with an asterisk symbol in Fig. 2A. Analyzing graphically the squared *z*-scores [Equation (3)] of the chi-square statistics, it was possible to observe greater uniformity deviations at both gene extremities, mainly at the 5′-end, and low values, close to zero, in intermediate gene positions. In addition, a Mann–Whitney test was applied because we were interested in testing the meaningful difference of the *z*-score distribution at the 5′-end bin to evaluate the statistical significance of the differences between deviations from uniformity by position as a whole.

Remarkably, using our quantification method constrained by the relativity of CDS length, the statistical test revealed that there are strong differences in codon selection against uniformity between the first part and subsequent position within genes. Moreover, when we tested the uniformity deviations at different positions within CDSs, we found significant differences (P-value ≤ 0.002) between the first and second bin for all yeasts, except for *K. marxianus* in which higher significance was found between the second and third bin, although we found higher squared z-score values at the 5′-end of CDSs compared with *K. lactis*, *P. pastoris* and *S. cerevisiae*. Accordingly, when we divided all the CDSs into an equal amount of parts (ten bins in this case), all yeasts had the highest deviations from uniformity at the 5′-end of genes. Furthermore, it should be noted that we found high squared z-scores at the 3′-end (Fig. 2A and B), suggesting that certain codons also present strong selection against uniformity at this extremity in agreement with the results obtained in a recent study.[65] Consequently, the results reported here denote the contribution of individual codons to the generation of strong deviations from uniformity at the extremities of CDSs, showing how the ramp theory is also presented as a function of CDS length at the evolutionary level in yeast.

3.4. Genome-scale signals of frequent codons with low optimality

Because we were interested in determining the characteristics of frequency and optimality of the translation efficiency of each codon, we calculated two broadly used indexes: the Relative Synonymous Codon Usage (RSCU) using the codon usage table (CUSP) of each genome (Supplementary Tables S3–S7) and the codon optimality (w, formally defined as the codon relative adaptiveness) to the tRNA pool using published data,[30] and when required, we predicted the tRNA counts (see Materials and Methods, tRNA counts in Supplementary Table S8). We used the RSCU values [Equation (1)] to characterize a codon as frequent or rare (Supplementary Table S9) and the codon adaptiveness (w) to characterize codons as optimal when $w \geq 0.4$ or non-optimal if $w < 0.4$ (Supplementary Table S10) in agreement with a previous study, in which w had been assessed in yeast genomes.[27]

Based on principles of codon optimization for heterologous protein expression in microorganisms, we expected to observe a high frequency of optimal codons, and complementarily, a low frequency of non-optimal codons. Visual inspection of the frequency values (RSCU in Fig. 3A), reveals that they were very similar among *K. lactis*, *K. marxianus*, *P. pastoris* and *S. cerevisiae*; however, the optimality of codons (w values) showed significant variations (Fig. 3B). On the other hand, we observed variations in *Y. lipolytica* in both codon frequency and optimality when compared with the other species. Intriguingly, we found for each genome heterogeneous correspondences between frequency/rarity and optimality/non-optimality. In contrast to the expected, we observed that several rarely used codons have optimal translation efficiency whereas frequently used codons have non-optimal translation efficiency.

From an evolutionary perspective, the evidence of codons which are frequently used but have non-optimal translation efficiency seems to indicate that there are selective pressures acting positively to conserve and use non-optimal codons, probably to guarantee the effectiveness of protein biosynthesis as elongation speed regulators. Therefore, we show that non-optimal and rare codons have different characteristics, and that optimal codons are different from frequent ones. Although they are referred to as synonymous in the literature,[66] these characteristics of frequency and optimality could be positively correlated for certain codons but not in all cases. Optimality refers to the adaptation to the cognate tRNA isoacceptors and frequency/rarity refers to the occurrence of a codon in the genome. Mixed characteristics of frequency and optimality should be considered in detail because they are different approaches used in the characterization of individual codons which could impact several molecular processes. Subsequently, we decided to explore the features of individual codons in depth.

3.5. Characterization between the Frequency and Optimality of codons

In order to characterize the different correspondences between the frequency and optimality of codons, we decided to study each codon

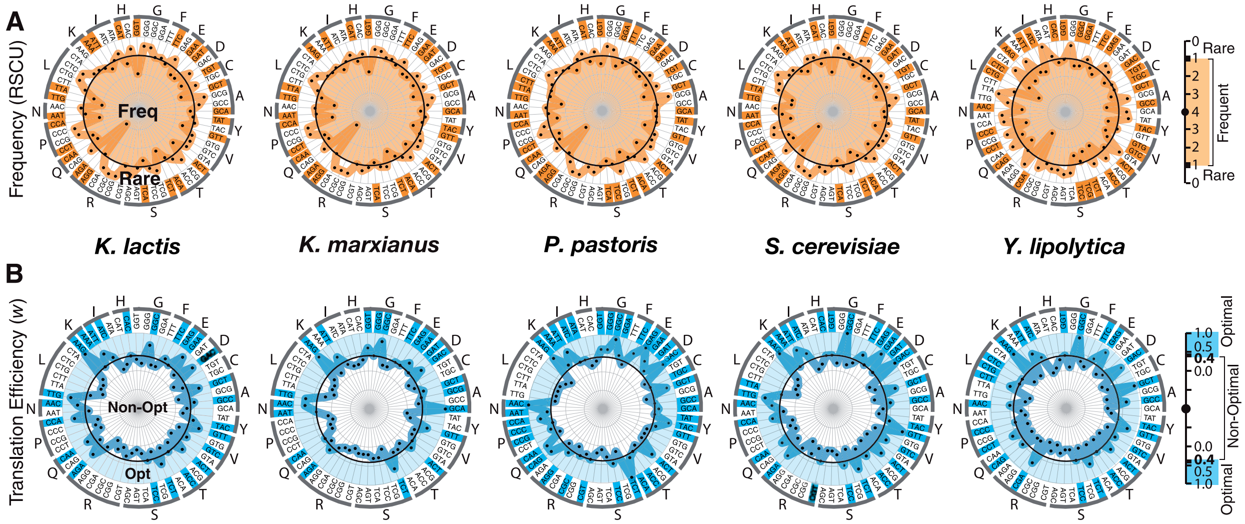

Figure 3. Heterogeneous correspondences between codon frequency and codon optimality metrics. (A) The values of the Relative Synonymous Codon Usage (RSCU) are shown here in order to characterize each codon in terms of frequency or rarity. Values near to the centre and inside the shaded region with black borderline are frequent codons, those outside the circle are rare. (B) Values of Codon Adaptiveness (w) to cognate tRNAs from the tRNA Adaptation Index (tAI) are represented to illustrate the translation efficiency of each codon. Points inside the shaded region, far apart from the centre of the plot represent "optimal" codons, opposite points near to the centre and inside the white region indicate non-optimal codons.

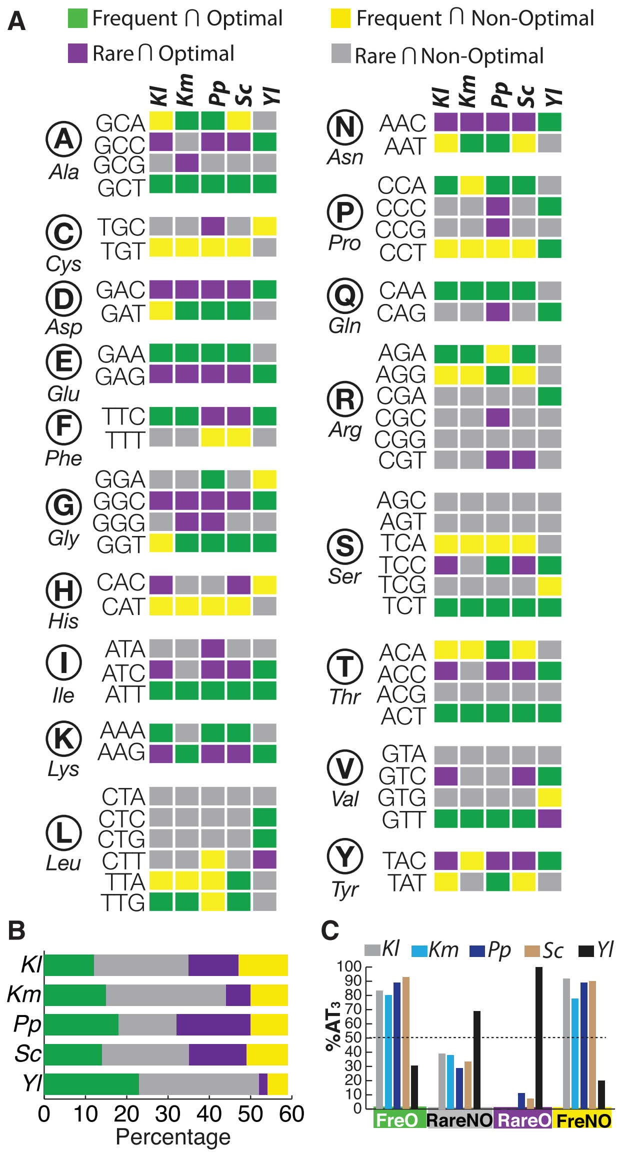

Figure 4. Integrated characterization of individual codons. (A) Each codon is characterized regarding its frequency and optimality. (B) Percentage of the four codon categories in the 59 redundant codons of each yeast genome. (C) The third-base composition of the four codon categories illustrates the differences between *Yarrowia lipolytica* and the other yeasts.

in an integrated framework (Fig. 4A) using the codon frequency and individual codon optimality for translation efficiency and then compare them to the data of the CUB positional dependency (Fig. 2A). It allowed us to define four categories for the classification of codons: (i) FreO, given by the intersection of high frequency and optimality; (ii) RareNO, given by low frequency (rare) and non-optimality; (iii) RareO, given by low frequency and optimality; and (iv) FreNO, given by high frequency and non-optimality.

Our integrated framework revealed the remarkable features of individual codons. For example, cysteine is coded by both the TGC and TGT codons, presenting uniform distribution with regard to the position within genes (Fig. 2A). In general, they have non-optimal adaptation to the tRNAs of each yeast genome (Fig. 3B), except for *P. pastoris*. In this yeast, the TGC codon is optimal for the tRNA pool (Fig. 3B); however, it is rarely used (Fig. 3A). On the other hand, codons of glutamine present a conserved non-uniform distribution as a function of the position within genes, possibly contributing to the formation of a selection ramp (Fig. 2A), with strong selection values against uniformity at the 5′-end in all yeasts. The CAA codon is frequent and optimal (Fig. 4A) for all the yeasts with the exception of *Y. lipolytica*, in which this codon is rare and non-optimal. In contrast, the CAG codon is rare and non-optimal in *K. lactis, K. marxianus* and *S. cerevisiae*, but frequent and optimal in *Y. lipolytica* (Fig. 4A). Another relevant feature is that, although CAG is rarely used and is non-optimal in the genomes of *K. lactis, K. marxianus* and *S. cerevisiae*, it is optimal in *P. pastoris* (Fig. 4A). Thus, this is a characteristic that is particular to *P. pastoris*. Correspondingly, *P. pastoris* has a higher quantity of codons that are optimal but rarely used (Fig. 4B), which seems to be an advantage for the production of recombinant proteins taking into account that these optimal codons, rarely used in the expression of endogenous genes, might be used more frequently in the expression of heterologous genes.

On the other hand, it was possible to observe the different properties in the codon selection features found in the genome of *Y. lipolytica* (Fig. 4B) when compared with the other yeasts. Although it is broadly known that *Y. lipolytica* has a very different codon usage, this microorganism seems to be under strong positive pressure from optimal codons to use them frequently and, at the same time, a strong negative selection over non-optimal codons to use them rarely.

Finally, we found important differences when comparing the third base composition of the four codon categories. *Y. lipolytica* presents very low content of AT_3 in its frequent codons and high content in the rare ones. On the other hand, all the other yeasts showed high AT_3 content of frequent codons and very low, or zero content, in rare codons (Fig. 4C). It should be pointed out that codon optimality seems to be an inaccurate characteristic to yeasts categorization considering the third base composition (Fig. 4C). It is important to clarify that data integration through this approach has been achieved by using genome sequences of microorganisms, which exhibit codon usage bias positively associated with translation efficiency. Therefore, the application of this method should consider these features before applying it to future research in non-translationally efficient organisms. Nevertheless, with the properties highlighted here it is possible to define the contribution of each codon in terms of evolutionary selection, translation efficiency and whole genome occurrences for each yeast cell factory. This integrated framework stands as a valuable tool that contributes to the improvement of knowledge on the different roles of individual codons in endogenous genes.

3.6. Rare–non-optimal codons have higher decoding time and are enriched at the beginning of CDSs

The characterization of codons led us to address the contribution of each codon category to the protein elongation process. Based on this, and the following three main statements: first, codons are translated at different rates;[26,50,65,67] second, we observed there are strong evolutionary selective pressures acting differentially on redundant codons to use them rarely or frequently; and third, codons are also differentially adapted to the tRNA pool in terms of efficiency, we hypothesized that it would be possible to find significant differences in the contribution of the four different categories of codons (FreNO, RareNO, etc.) to the kinetics of the elongation process.

In order to test this, for each codon we computed the codon decoding time [CDT, Equation (2)] as described previously[65] and

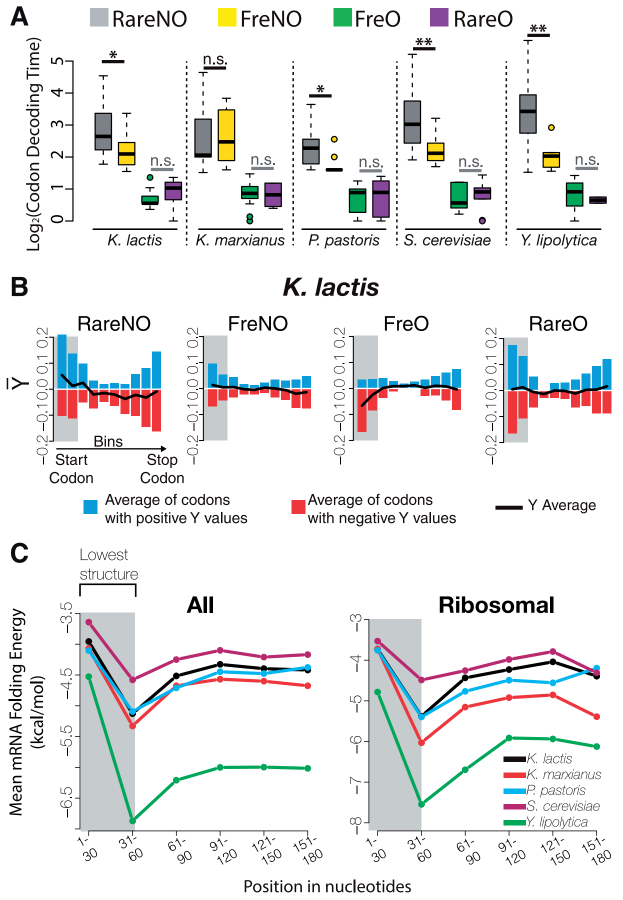

Figure 5. Codon decoding time (CDT), the average of Y-values and the mRNA free folding energy are presented to indicate the features of each codon category. (A) shows the significant differences found when CDT values of Rare–Non-optimal (RareNO) are compared with Frequent–Optimal (FreNO) codons. The translation rate is the slowest when decoding RareNO codons. No significant differences between the optimal ones (FreO and RareO) were found. (B) RareNO codons seem to be enriched at regions near to start codons in *K. lactis* and other yeasts (Supplementary Fig. S8), while Frequent–Optimal (FreO) codons contrast that profile. FreNO codons are not significantly enriched or depleted as a function of position within genes, they are uniformly distributed inside genes, probably to guarantee the accuracy in protein biosynthesis as translation speed regulators (see Supplementary Figs S3–S7 for individual values of codons per bin). (C) mRNA structure is the lowest in the 5′-end region, where RareNO codons are enriched. The MFE was computed for all the CDSs of each yeast ($n_{K.\ lactis}$ = 5065, $n_{K.\ marxianus}$ = 4774, $n_{P.\ pastoris}$ = 5019, $n_{S.\ cerevisiae}$ = 5786, $n_{Y.\ lipolytica}$ = 6413).

subsequently we formed four clusters of codons according to the categories revealed by the integrated framework. Finally, we analysed the distribution of CDT values, a measurement aimed at addressing the time required to decode each codon.

Confirming our hypothesis (Fig. 5A), rarely used non-optimal codons (RareNO) have significantly higher CDTs ($P < 0.05$, Wilcoxon rank sum test) than frequent non-optimal (FreNO) codons, except in the case of *K. marxianus*. In this yeast, the statistical test was not significant. It should be noted (Fig. 5A) that certain RareNO codons of *K. marxianus* have CDT values higher than the highest CDT values of FreNO codons. However, there was no significant difference between the RareNO and FreNO groups in this yeast. This particularity could be due to difficulties in the *w* value computation used for calculating CDTs. Thus, it is expected that the future availability of *K. marxianus* protein expression data will allow for constraining the computation of their respective *w* values, which in turn will contribute to improving their CDT calculation.

Additionally, the detection of higher CDTs conserved in RareNO codons seems to explain their low frequency in intermediate gene regions where the protein biosynthesis process should be highly efficient. This reveals that the decoding time of certain codons is a special feature that constrains codon selection, which explains why certain different non-optimal codons are used more frequently than others in yeast. Thus, we conclude that the insertion of RareNO codons in genes can facilitate slower protein elongation. In fact, RareNO codons introduce lower translation rates than FreNO codons, an important characteristic of note in the synthetic gene design.

The results described above also led us to hypothesize that, whether the ramp theory applies to translation efficiency or not,[10,30,67–70] we should consider these signals in terms of richness or depletion by codon category as a function of position within CDSs. Hence, a higher occurrence of RareNO codons should be expected at the beginning of the genes which will modulate the initiation of elongation. At the same time, we would expect its absence in intragenic regions where the elongation process must be more efficient. On the other hand, it is reasonable to consider that FreNO codons, which are equally non-optimal as RareNO codons, but have lower decoding times, should be found at different positions on the genes. It is probable that they act as elongation speed regulators for protein co-translational folding,[22] and that elongation will not be slowed as much as when decoded by RareNO codons. To test this, we determined if the different categories of codons are enriched or depleted as a function of position in the ten bins. We calculated the Y-value [Equation (4)] as the ratio of observed (*O*) values to expected values (*E*) on a logarithmic scale, to determine the fold-change of observed codons against expected codon occurrences in simulated genomes with scrambled position of codons in their CDSs. The computation was carried out for each yeast genome (Supplementary Figs S3–S7).

Interestingly, we found that RareNO codons are predominantly enriched at regions close to the start codon (Fig. 5B) and depleted in distant regions. In general, FreO codons are predominantly depleted at the beginning of genes. In agreement with recent data,[30] we add important evidence to the theory of the translation ramp, with two remarkable additions: first, we suggests the existence of the ramp at the evolutionary level in yeast, and second, we found a specific group of codons (RareNO codons) enriched at the beginning of genes, which are responsible for the lowest translation rates. Moreover, we observed a conserving tendency in FreNO codons to avoid either enrichment or depletion. They were found uniformly distributed through the genomes of *K. lactis*, *K. marxianus*, *P. pastoris* and *S. cerevisiae* (Supplementary Fig. S8), but strongly enriched at the beginning of genes in *Y. lipolytica*, pointing out that this yeast also has the lowest percentage of FreNO codons in comparison to the other yeasts (Fig. 4B). Another interesting point worthy of note is that our finding related to the enrichment of RareNO codons in the 5′-end region of endogenous genes in yeast could be associated with the maintaining of a decreased local mRNA structure (Fig. 5C and Supplementary Dataset S1). A similar result has been described recently in *E. coli*,[42] suggesting that rare codons assist to preserve reduced mRNA structure.

In our work, the local mRNA minimum folding energy (MFE) at the beginning of genes showed the lowest local mRNA structure when compared with intragenic positions distant from the 5′-end region. Our calculation indicates that it is a conserved characteristic when applied to the whole gene dataset of yeasts and to the highly expressed genes such as ribosomal proteins. Accordingly, this observation among mRNA, MFE and codon positional dependency in yeast suggests a similar phenomenon, as determined experimentally in bacteria,[71] where rare codons reduced the secondary structure of mRNAs. Therefore, our data reinforce, with novel genome-scale characteristics, the ramp theory of codon usage bias.[10,30,70] Additionally, the aforementioned findings validate the existence of the ramp at the evolutionary level and allow us to specify which codons are really contributing to slower translation speeds, illustrating the orchestration of selective constraints, translation efficiency, codon occurrence, local mRNA folding energies and the determinant contributions of individual codons, especially the imperative leadership of non-optimal codons, to the protein biosynthesis.

3.7. Genes with signal peptides use preferentially RareNO codons at the 5′-end

Signal peptide (SP) design and its synthetic construction is an effective method used in industrial biotechnology to improve heterologous protein production and biocatalysis.[72,73] Thus, investigation of the differential codon composition of genes with SPs has been an objective of various studies despite the different approaches employed.[26,43–44] We used our integrated framework, to analyse the CDSs containing SPs for each yeast genome, by computing the Y-value (Fig. 6) and the deviations from uniformity (Supplementary Fig. S9).

Because SPs are composed mainly of a hydrophobic core of amino acids[74] such as glycine (Gly), alanine (Ala), valine (Val), leucine (Leu), isoleucine (Ile), proline (Pro) and phenylalanine (Phe), we hypothesized that if translation efficiency controls the SPs, the aforementioned amino acids would be codified preferentially by the less efficient (non-optimal) codons, and then we would expect their enrichment in the 5′-end region of CDSs.

In agreement with our hypothesis, it is possible to observe that in the first bin (Fig. 6) Leu is codified preferentially by CTG in *K. lactis* (*Kl*), CTG in *K. marxianus* (*Km*), CTC in *P. pastoris* (*Pp*), CTC in *S. cerevisiae* (*Sc*) and TTA in *Y. lipolytica* (*Yl*). Although they are different codons, all of them are RareNO codons. On the other hand, for the same amino acid, the most depleted codon is TTG in *Kl*, TTG in *Km*, TTA in *Pp*, TTG in *Sc* and CTT in *Yl*, which are optimal codons except for *Pp* that does not present optimal codons for this amino acid. In other cases, Val is preferentially codified by GTA in *Kl*, GTA in *Km*, GTG in *Pp*, GTG in *Sc* and GTA in *Yl*, being all RareNO codons. Phe is preferentially codified by TTT in *Kl*, *Km* and *Pp*, which is a non-optimal codon, and at the same time avoids TTC which is optimal in all cases; *Sc* and *Yl* do not present codon bias for this amino acid. Gly is preferentially codified by GGG in *Kl*, GGA in *Km*, GGG in *Sc* and GGG in *Yl*, which are all RareNO codons. *Pp* does not present codon bias for this amino acid, while *Km*, *Sc* and *Yl* avoid the use of the GGT codon which is FreO in all cases.

Interestingly, a number of researchers reported significant improvement in yeast protein secretion when Arg was inserted exactly in the N-terminal of synthetics SPs[74,75] and, as shown in Fig. 6, the amino acid with the highest bias (conserved characteristic among the yeasts) is indeed Arg. We checked the features of its codons and

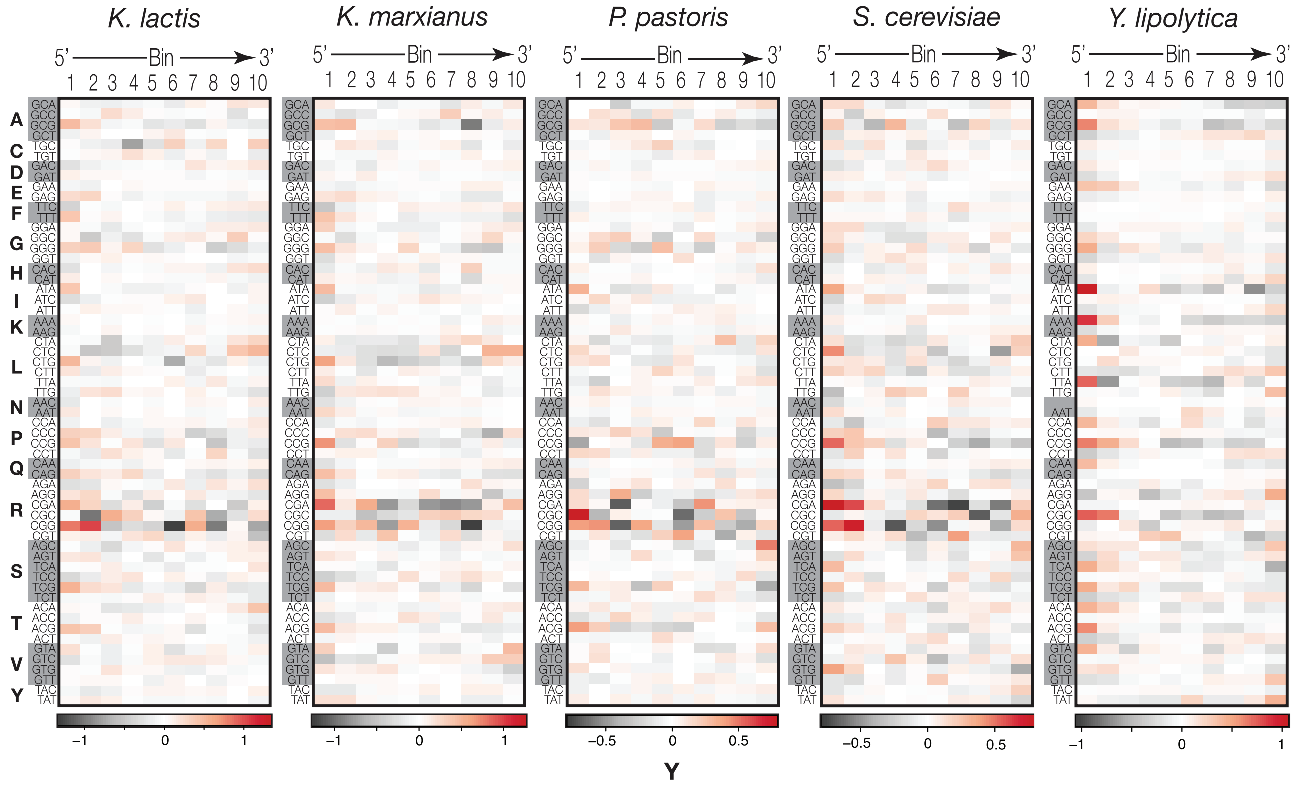

Figure 6. The codon positional dependency in genes coding proteins with signal peptide. The figure shows the analysis applied to the set of CDSs which have signal peptides. The Y-value (see Materials and methods) is reported graphically to illustrate the codon enrichment or depletion at different intragenic positions.

found that Arg is preferentially codified in the 5′-end region by CGG in *Kl*, CGA in *Km*, CGA in *Sc* and CGC in *Yl*, which are RareNO codons, while *Pp* uses preferentially CGC (RareO) and CGG (RareNO). Therefore, this is an important feature to be considered in experimental designs of synthetic SPs.

In general, we observed in this study a conserved feature: amino acids that are part of the SPs are codified preferentially by non-optimal codons, most of them being RareNO codons. We also verified that this genic region is commonly depleted in optimal codons, demonstrating through this integrated approach, that positional dependency governs the codon arrangement in SPs, which seems to provide slow translation rates at the 5′-end and maintains the ramp of translation efficiency of genes,[30] to guarantee co-translational interaction between SPs and the signal recognition particle.[26]

3.8. Codon category contribution to protein structure and the necessity of new optimization methods

Protein co-translational folding is one of the most important features of individual codons in protein biosynthesis.[22,67,76,77] We analysed the relationships between different categories of codons and the protein tertiary structure, as well as important issues related to the use of the classical method of codon optimization such as the Codon Adaptation Index (CAI).

We assessed in each yeast genome a conserved CDS with a similar CAI value, but with different protein expression levels under similar physiological conditions. In this way, we analysed the TEF1a (Translation elongation factor 1 alpha) and SSE1 (coding a heat-shock protein) genes which have similar CAI values in each host yeast genome (Fig. 7A). Subsequently, we performed codon multiple alignments as described in a previous study[27] and then we considered the corresponding codon category and determined the levels of conservation among the 5 yeasts (Fig. 7B). We represented the codon-category conserved sites in the 3D structure of both proteins, which allowed us to illustrate conserved evolutionary types of translation efficiency and co-translational folding over two differently expressed genes, despite having close CAI values.

First, it is important to highlight that by using our codon categorization, a potential impact on essential structural features was uncovered. Our novel categorization scheme allowed us to identify patterns of conserved optimal codons in agreement with previous findings,[27] but going beyond the categorizing of individual codons by optimality properties, the proposed method herein takes into account the fact that these optimal codons are also frequent codons in

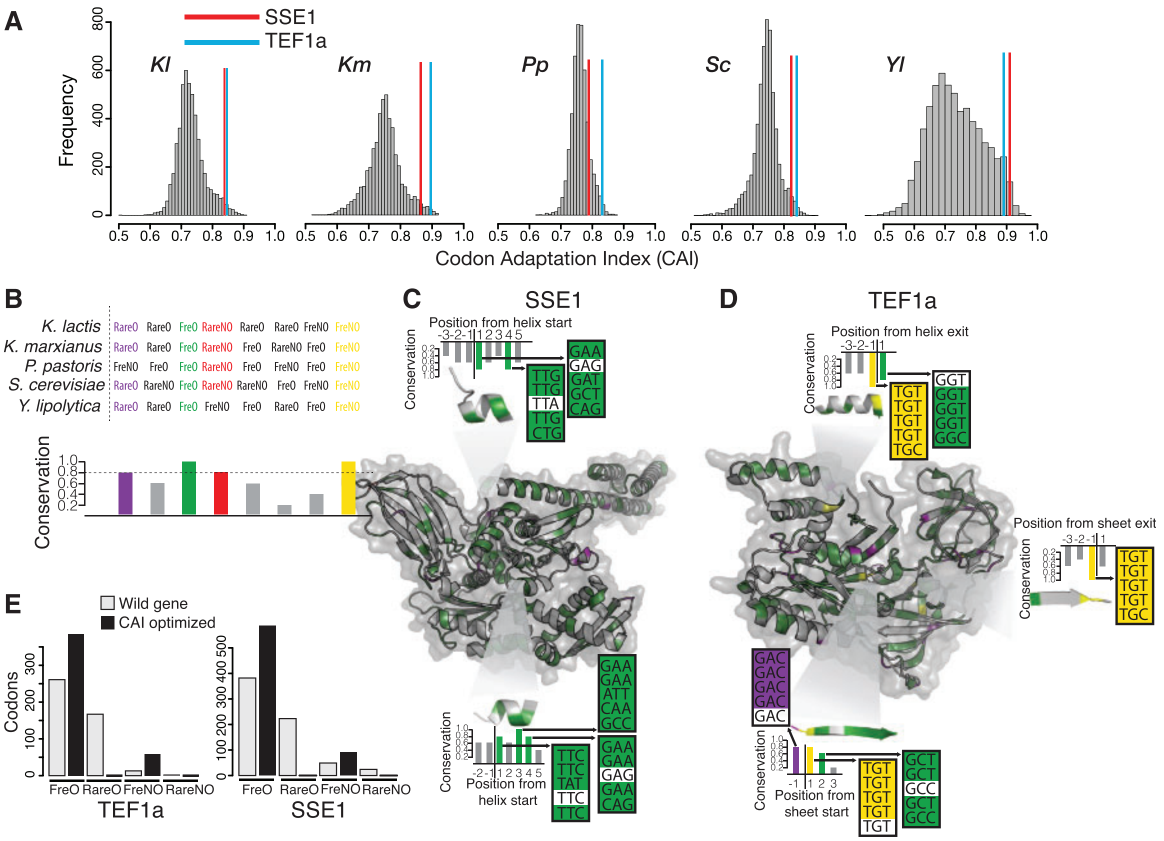

Figure 7. Proof of concept on codon category contribution to structural features of two different proteins. (A) Similar values of Codon Adaptation Index (CAI) were found for the SSE1 (expressed as a response of heat-shock) and the TEF1a proteins (highly and constitutively expressed). As both present similar CAI values to their host genomes, and have very different expression rates in physiological conditions, they are an interesting example to test our concept. CAI, a commonly used index, seems to be a non-well fitted method for predicting expression levels of proteins and thus is not recommended for the secure optimization of yeast genes. (B) Method for multiple sequence alignment based on codons to determine the conservation of codon categories in coding sequences. If a codon category is conserved at least in four out of five yeasts then the codon is defined as conserved and its category is illustrated with its correspondent colour in the protein structure. (C) Conserved codon categories are highlighted in SSE1 protein structure. Motifs of conserved Frequent–Optimal codons (FreO) at determinant positions in alpha-helices. (D) Conserved codon categories are highlighted in TEF1a protein structure. Positions with conserved Frequent–Non-optimal codons (FreNO), illustrating conserved non-optimal sites within genes at different positions to possibly regulate the efficiency and accuracy of the protein elongation process, being a potential requirement to the co-translational folding of proteins. (E) Changes in codon category composition when both proteins were optimized by CAI algorithm. CAI optimization introduces silent mutations that could impact translation rates by insertion of FreO and FreNO codons.

a genome-scale analysis. Thus, we observed FreO codons conserved in positions 1 and 4 from the helix start, and additionally, the enrichment of conserved FreO codons in the alpha-helices (Fig. 7C and D). One important point to be noted is the absence of FreNO conserved sites in the structure of the SSE1 protein (Fig. 7C), whose translation is low, but presents a conserved RareNO site at the beginning of the gene in the second codon/amino acid position. However, it was not possible to illustrate it in the protein structure due to lack of information of this amino acid in the original 3D protein model. In addition, it is important to observe (Fig. 7C and D) that even different codons, or indeed, when different amino acids are being coded, the codon category is conserved among the yeasts, suggesting the importance of the properties of individual codons to the conformation of structural features in proteins.

In the case of TEF1a, a highly expressed protein, we found optimal (FreO and RareO) and FreNO conserved sites in important structural locations (Fig. 7D), such as the last position of beta-sheets and alpha-helices. It is reasonable to claim that this FreNO conservation provides the control needed over the protein elongation rate. It seems that the ribosome suffers a decrease in translation speed to retain inside its tunnel the previously translated amino acids in order to fold the structural motif. The conserved FreNO sites are found in general at ∼20–30 codons downstream of complete alpha-helices or beta-sheets, fitting perfectly to the length of the ribosomal tunnel (∼30 amino acids), the region responsible for protein co-translational folding.[26,78,79] After they have been analysed, we focused our investigation on a more generalized perspective about the role of the four categories of codons in secondary protein structures.

First, we hypothesized that codon features affect the translation efficiency of the ribosome, and assist protein secondary structure conformation. Therefore, we expected to observe that the secondary structures would be translated at different rates and their codon composition would not be equally distributed between the cluster of frequently used codons (FreO and FreNO) and the cluster of rarely used codons (RareO and RareNO). To test this, we retrieved 300 structures from the RSCB Protein Data Bank and filtered out the redundant ones, resulting in 138 non-redundant protein structures. Next, we evaluated the occurrence of the four codon categories inside all their alpha-helices, beta-sheets and coil regions (coil was assumed to be the regions which do not belong to helix and sheet coordinates in PDB files).

We found that the codon composition of alpha-helices and beta-sheets is similar (∼41% FreO, ∼16% FreNO, ∼18% RareNO and ∼23% RareO; Supplementary Fig. S10A), also ∼65% of the codons were optimal and ∼35% non-optimal, suggesting that a substantial proportion of the secondary structures is translated slowly. Contrary, coil regions presented a notable difference in codon composition (Supplementary Fig. S10A). We observed that the composition of RareNO codons in coils was the highest compared with the other two structures, approximately 26% of its total composition. This result indicates that, at least for the proteins analysed, coils have numerous points with the slowest translation efficiency as determined by codon decoding times in Fig. 5A. Moreover, codon optimality distribution in coil regions is particularly characterized by an increased presence of non-optimal codons (∼44% non-optimal and ∼56% optimal; Supplementary Dataset S2).

We believe that these observations are attributed to the dynamics of the protein elongation process. As an example, we consider a simplified part of a protein composed by a helix followed downstream by a coil. When the region corresponding to the helix has been already (optimally) translated, it would be fitted inside the ribosome tunnel where would occur the co-translational folding.[26,78,79] Probably, the time required to take place this folding is provided by the decrease of translation rate in coil region, which would be associated to the presence of RareNO codons (enriched in the coil). Another possible explanation to this observation could be attributed to the evolutionary pressures that act over the sheet and helix structures, as they determine the correct function of proteins one will expect that these secondary structures tend to be adapted for efficient translation while coils could accumulate different silent mutations. This reasoning is still a theoretic approximation to the observations and further approaches should be applied to unveil the mechanisms behind these intriguing results. We suggest that synthetic optimization of codons in those structures should consider that changes in local translation rates could impact the protein co-translational folding and ultimately its function.

Finally, we were interested in examining the codon composition at the extremities of the secondary structures which seems to be specifically under higher selective pressures considering the previous TEF1a/SSE1 inspection. To analyse the positional dependency of the structure-codon correspondence we studied the differences in the codon composition of the four codon categories at different positions. Because the secondary structures of proteins present a high degree of variation in length, we decided to constrain our analysis using the first three and the last three codon positions. Notably, the codon composition of FreNOs presented the highest increment in the first codon of helices when compared with the overall (Supplementary Fig. S10B). Although FreO codons are preferentially found at this first position (∼39%; Supplementary Dataset S2), FreNO codons presented their highest occurrence (∼21%) at this position when compared with all other positions. At subsequent positions (Codons 2 and 3) the FreNO composition is decreased and similar to the overall, in the third codon position FreO codons were found to be relatively increased. At the last position on the helices (Codon n), FreO codons decreased their occurrence and the non-optimal codons presented a relative increment. Also, we observed an increased occurrence of RareO codons (Supplementary Fig. S10A). In general, our analysis indicates that helices are relatively depleted in optimal codons at the first and last position of the structure (Supplementary Fig. S11).

These results denote that the extremities of the helices, could serve as a transition point in the translation rate to start or terminate structure translation. This transition point in beta-sheets is characterized, in contrast to helices, by a decrease in FreO codons occurrence at the end of beta-sheets (Supplementary Fig. S10B), where FreNO codons are relatively enriched, and by a relative depletion of non-optimal codons at the beginning of the structure (Supplementary Fig. S11). In the case of the coil regions, we observed that the first three positions presented a remarkable enrichment of RareNO codons (Supplementary Fig. S10B), specifically, the position corresponding to the second codon presented the highest increment of RareNO codons in all the structures, suggesting that the translation at the beginning of coil regions is characterized by a very slow rate. Even though the occurrence of non-optimal codons in coil regions is mainly concentrated at the beginning of coils, we also found them increased at the end of the structure (Supplementary Fig. S11). Although codon overall composition of protein secondary structures seems not very dissimilar (Supplementary Fig. S10A), it is possible to perceive that the three protein secondary structures presented contrasting profiles in the extremities (Supplementary Fig. S11). In general, this intriguing outcome is a theoretical evidence that protein secondary structures could be ruled by the codon positional dependency like the observed properties of coding sequences at genome-scale, having potential implications in the understanding of protein evolution as well as biotechnological applications. However, this is an unexpected result and further research should be aimed at collecting more evidence on the transition points presence and the position-dependent translation efficiency coupled to the co-translational folding in inner regions of protein secondary structures.

Although we have reported these correspondences between codon categories and protein secondary structures of 138 non-redundant proteins, we suggest that codon-structure correspondence analysis, aimed at heterologous expression experiments, should be performed on a gene-by-gene basis, to identify the codon category in relevant protein structures such as the N-terminus (which in helices seems to be generally enriched in FreNO and depleted in optimal codons) and the structural transition points. Accordingly, both optimal and non-optimal codons have molecular tasks important to cell development. Furthermore, the high complexity of individual codon contribution to protein biosynthesis in yeast cell factories should be considered in codon optimization.

This reasoning led us to show the potential effects of traditional codon optimization methods such as the Codon Adaptation Index (CAI) approach for optimization of genes. In our experiment (Fig. 7E), we obtained genes with a significant number of frequent codons which should not be considered as optimal; thus, it may be considered normal to obtain a sequence enriched in FreNO codons after gene optimization using the CAI. Then, these FreNO codons could affect protein translation elongation by decreasing the local translation rate and possibly generating unexpected protein qualities.

Consequently, individual codons have different features that contribute in diverse ways to protein biosynthesis. It depends on several characteristics such as the selective forces that define which codons are enriched and depleted as a function of position, the frequency of codon usage, the adaptation of each codon to the cognate tRNAs available in the cell, the GC content, and the mRNA folding energy. Additionally, each codon has a different role as a speed regulator in protein translation, property that orchestrates the accuracy, efficiency and co-translational folding in protein biosynthesis. Hence, simplistic manipulation of codons for heterologous expression does not take into account the complex biological implications of individual codons.

We propose that the systematic overview presented in this work must be considered when designing synthetic genes. We define this proposed approach as Codon Transliteration, in which a gene to be used in heterologous expression needs to be studied in depth.

In this proposed approach, each codon that encodes the same amino acid in the original organism must have its properties characterized. Subsequently, each codon and its features in the new host must be identified. As an example, we consider a gene from *K. marxianus* to be expressed in *Y. lipolytica*. The codon TGT of *K. marxianus* is categorized as FreNO (Fig. 7D), that is, it has features such as non-optimal codon and frequent occurrence in different positions within genes, a slow translation rate, but a codon decoding time lower than RareNO codons. Nevertheless, the same codon (TGT) is RareNO in *Y. lipolytica* and for this reason, it should not be chosen for expression in this yeast because despite encoding the same amino acid (cysteine), it contributes differently to the protein elongation process. Therefore, the introduction of this codon will lead to a decrease in the elongation rate, and generate unexpected results related to protein co-translational folding. On the other hand, the TGC codon also encodes cysteine in *Y. lipolytica* and it is categorized as FreNO codon, and conserves the same features found in *K. marxianus*. For these reasons, the TGC codon should be chosen in *Y. lipolytica*.

On the other hand, we expect that genes which are not highly expressed could be efficiently expressed in heterologous hosts by transliterating their codon features (FreNO, RareNO, etc.) to the new host. Thus, in order to optimize individual codons to the new host in terms of their role in the protein secondary structure, we suggest that the category of codons encoding amino acids inside alpha-helices and structural transition points should not be modified. Nevertheless, those that contribute to other structures could be optimized with a view to improving the translation rate of ribosomes.

In conclusion, by using traditional approaches, synonymous mutations can affect the protein synthesis in terms of quantity, folding and function (Fig. 8). As described through this integrated analysis, each codon seems to play a determinant role in gene translation efficiency and co-translational protein folding, and different impacts could be generated when a gene sequence is manipulated without the details deserved.

By insertion of synonymous (silent) mutations (Fig. 8A), a hypothetical protein-coding gene, shaped by different codon category emerges. Its translation efficiency is maintained by the absence of slower translated codons (RareNO, see Fig. 5A for details), and the presence of both FreO codons, that enhance the translation speed, and FreNO codons, which assist protein co-translational folding which slow the translation rate in order to support the polypeptide

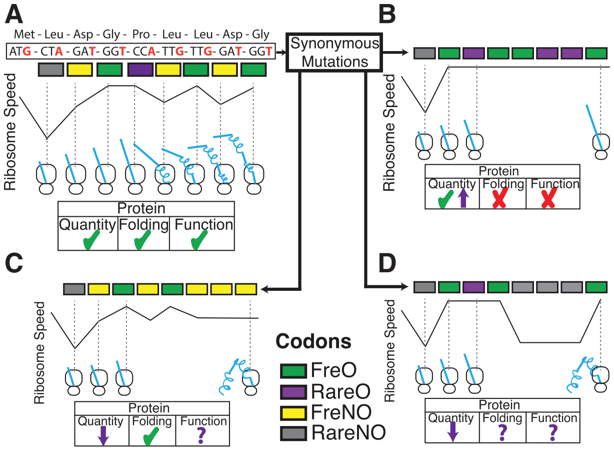

Figure 8. Theoretical scheme of codon contribution to translation efficiency and co-translational protein folding. The figure shows the potential impacts in quantity, folding and function of proteins by insertion of synonymous (silent) mutations. The four categories of codons determined in the present work have been used to illustrate each codon contribution for protein biosynthesis. (A) A hypothetical protein-coding gene shaped by different codons category. (B) The gene has been modified in order to improve its optimality for available tRNA isoacceptors. Translation speed is enhanced, positively affecting protein quantity. On the other hand, this modification can affect negatively the co-translational folding and with that the protein function. (C) An optimized version in terms of frequency (e.g. by CAI) of the wild-type gene. The insertion of FreNO codons affects negatively the quantity of protein produced, but on the other hand supports the maintenance of a slower speed in the translation steps and thereby assists in co-translational protein folding. (D) Silent mutation inserting RareNO codons (slowest translated codons). It affects the translation rate, decreases protein quantity and leads to unknown features in co-translational folding and function. Abbreviations: FreO: Frequent and Optimal codons; RareO: Rare and Optimal; FreNO: Frequent and Non-optimal; RareNO: Rare and Non-optimal.

folding inside the ribosome tunnel. This arrangement of a different category of codons allows for normal protein production in terms of functionality, quantity and folding. However, when the gene has been modified (Fig. 8B) in order to improve its optimality to available tRNA isoacceptors (e.g. by tAI), the insertion of synonymous mutations of FreO and RareO codons leads to the enhancement of translation speed which affects the protein quantity positively, but the co-translational folding and, consequently, its function negatively.

In another case, an 'optimized' version, in terms of codon frequency (e.g. by CAI), of the wild-type gene is presented (Fig. 8C). The insertion of FreNO codons affects the quantity of protein synthesized negatively, but, on the other hand, maintains a slow speed in the translational periods, and by doing this, assists co-translational protein folding. However, excessive insertion of FreNO codons implies more sites at which the ribosome decreases its translational speed and generates potential misfolded structures in the protein with a different function to the wild-type. Finally, when silent mutations are created in a gene, containing more RareNO codons (Fig. 8D) than the wild-type gene, they could affect its translation rate, decrease protein quantity and lead to unknown features in co-translational folding and function.

In summary, it seems that codons, through a positional dependency function, are arranged so as to exert control over different molecular events such as the initiation of translation by means of mRNA folding energy, the translation per se by the codon adaptiveness to the tRNA isoaceptors, the yield protein production by the codon translation efficiency, and the protein function in terms of their different roles in co-translational folding. The approach developed in this study, based on the integrated analysis of individual codons, contributes to an in-depth understanding of the biological roles of codons, and features an improvement in the basis of rational gene design in heterologous expression and synthetic biology.

Availability

Algorithms description, codes and tutorial: https://github.com/juanvillada/CodG

Acknowledgements

We thank Maria Fernanda Duran Meneses and Professor Flavia Maria Lopes Passos from Laboratorio de Fisiologia de Microrganismos at the UFV for research design support; Professor Denise Mara Soares Bazzoli from Laboratorio de Genetica Molecular de Bactérias and Professor Hilario Cuqueto Mantovani from Laboratorio de Microbiologia do Rumen e de Anaeróbios at UFV for their critical review of the manuscript. This study was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES), the National Science and Technology Development Council (CNPq) and the Foundation for Research Support of the State of Minas Gerais (FAPEMIG). Support for JCV was obtained through the Organization of American States (OAS) and the Coimbra Group of Brazilian Universities (GCUB) Scholarship Award.

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at *DNARES* Online.

Author contribution

JCV and WBDS designed the research; JCVA and OJBB performed the experiments; JCV, OJBB, WBDS analysed the data, and JCVA and WBDS wrote the manuscript.

References

1. Ikemura, T. 1981, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes, *J. Mol. Biol.*, **146**, 1–21.
2. Sharp, P. 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**, 1281–95.
3. Dong, H., Nilsson, L. and Kurland, C.G. 1996, Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates, *J. Mol. Biol.*, **260**, 649–63.
4. Hahn, M.W., Mezey, J.G., Begun, D.J., et al. 2005, Evolutionary genomics: codon bias and selection on single genomes, *Nature*, **433**, E5–6.
5. 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**, 5036–44.
6. Gingold, H. and Pilpel, Y. 2011, Determinants of translation efficiency and accuracy, *Mol. Syst. Biol.*, **7**, 481.
7. Kanaya, S., Yamada, Y., Kudo, Y. and Ikemura, T. (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis, *Gene*, **238**, 143–55.
8. Boël, G., Letso, R., Neely, H., et al. 2016, Codon influence on protein expression in *E. coli* correlates with mRNA levels, *Nature*, **529**, 358–63.
9. Najafabadi, H.S., Goodarzi, H. and Salavati, R. 2009, Universal function-specificity of codon usage, *Nucleic Acids Res.*, **37**, 7014–23.
10. Tuller, T. and Zur, H. 2015, Multiple roles of the coding sequence 5′ end in gene expression regulation, *Nucleic Acids Res.*, **43**, 13–28.
11. Novoa, E.M. and de Pouplana, L.R. 2012, Speeding with control: codon usage, tRNAs, and ribosomes, *Trends Genet.*, **28**, 574–81.
12. Ding, Y., Shah, P. and Plotkin, J.B. 2012, Weak 5′-mRNA secondary structures in short eukaryotic genes, *Genome Biol. Evol.*, **4**, 1046–53.
13. Tuller, T., Waldman, Y.Y., Kupiec, M. and Ruppin, E. 2010, Translation efficiency is determined by both codon bias and folding energy, *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 3645–50.
14. Zur, H. and Tuller, T. 2012, Strong association between mRNA folding strength and protein abundance in *S. cerevisiae*, *EMBO Rep.*, **13**, 272–7.
15. Presnyak, V., Alhusaini, N., Chen, Y.H., et al. 2015, Codon optimality is a major determinant of mRNA stability, *Cell*, **160**, 1111–24.
16. Gu, W., Zhou, T. and Wilke, C.O. 2010, A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes, *PLoS Comput. Biol.*, **6**, e1000664.
17. Mao, Y., Wang, W., Cheng, N., Li, Q. and Tao, S. 2013, Universally increased mRNA stability downstream of the translation initiation site in eukaryotes and prokaryotes, *Gene*, **517**, 230–5.
18. Chaney, J.L. and Clark, P.L. 2014, Roles for synonymous codon usage in protein biogenesis, *Annu. Rev. Biophys.*, **44**, 143–66.
19. Willie, E. and Majewski, J. 2004, Evidence for codon bias selection at the pre-mRNA level in eukaryotes, *Trends Genet.*, **20**, 534–8.
20. Gu, W., Wang, X., Zhai, C., Zhou, T. and Xie, X. 2013, Biological basis of miRNA action when their targets are located in human protein coding region, *PLoS One*, **8**, e63403.
21. Gu, W., Wang, X., Zhai, C., Xie, X. and Zhou, T. 2012, Selection on synonymous sites for increased accessibility around miRNA binding sites in plants, *Mol. Biol. Evol.*, **29**, 3037–44.
22. Yu, C.H., Dang, Y., Zhou, Z., et al. 2015, Codon usage influences the local rate of translation elongation to regulate co-translational protein folding, *Mol. Cell*, **59**, 744–54.

23. Deane, C.M. and Saunders, R. 2011, The imprint of codons on protein structure, *Biotechnol. J.*, **6**, 641–9.
24. Tsai, C.J., Sauna, Z.E., Kimchi-Sarfaty, C., Ambudkar, S.V., Gottesman, M.M. and Nussinov, R. 2008, Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima, *J. Mol. Biol.*, **383**, 281–91.
25. Nissley, D.A. and O'Brien, E.P. 2014, Timing is everything: unifying codon translation rates and nascent proteome behavior, *J. Am. Chem. Soc.*, **136**, 17892–8.
26. Pechmann, S., Chartron, J.W. and Frydman, J. 2014, Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo, *Nat. Struct. Mol. Biol.*, **21**, 1100–5.
27. Pechmann, S. and Frydman, J. 2013, Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding, *Nat. Struct. Mol. Biol.*, **20**, 237–43.
28. Gustafsson, C., Govindarajan, S. and Minshull, J. 2004, Codon bias and heterologous protein expression, *Trends Biotechnol.*, **22**, 346–53.
29. Sharp, P.M. and Li, W.H. 1986, An evolutionary perspective on synonymous codon usage in unicellular organisms, *J. Mol. Evol.*, **24**, 28–38.
30. Tuller, T., Carmi, A., Vestsigian, K., et al. 2010, An evolutionarily conserved mechanism for controlling the efficiency of protein translation, *Cell*, **141**, 344–54.
31. Ragionieri, L., Vitorino, R., Frommlet, J., et al. 2015, Improving the accuracy of recombinant protein production through integration of bioinformatics, statistical and mass spectrometry methodologies, *FEBS J.*, **282**, 769–87.
32. Zhou, W.J., Yang, J.K., Mao, L. and Miao, L.H. 2015, Codon optimization, promoter and expression system selection that achieved high-level production of *Yarrowia lipolytica* lipase in *Pichia pastoris*, *Enzyme Microb. Technol.*, **71**, 66–72.
33. Wang, J.R., Li, Y.Y., Liu, D.N., et al. 2015, Codon optimization significantly improves the expression level of α-amylase gene from *Bacillus licheniformis* in *Pichia pastoris*, *Biomed Res. Int.*, **2015**, 248680.
34. Yu, P., Yan, Y., Gu, Q. and Wang, X. 2013, Codon optimisation improves the expression of *Trichoderma viride* sp. endochitinase in *Pichia pastoris*, *Sci. Rep.*, **3**, 3043.
35. Agashe, D., Martinez-Gomez, N.C., Drummond, D.A. and Marx, C.J. 2012, Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme, *Mol. Biol. Evol.*, **30**, 549–60.
36. Rosano, G.L. and Ceccarelli, E.A. 2009, Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted *Escherichia coli* strain, *Microb. Cell Fact.*, **8**, 41.
37. Spencer, P.S., Siller, E., Anderson, J.F. and Barral, J.M. 2012, Silent substitutions predictably alter translation elongation rates and protein folding efficiencies, *J. Mol. Biol.*, **422**, 328–35.
38. Finger, C., Gamer, M., Klunkelfuß, S., Bunk, B. and Biedendieck, R. 2015, Impact of rare codons and the functional coproduction of rate-limiting tRNAs on recombinant protein production in *Bacillus megaterium*, *Appl. Microbiol. Biotechnol.*, **99**, 8999–9010.
39. Cheong, D.E., Ko, K.C., Han, Y., et al. 2015, Enhancing functional expression of heterologous proteins through random substitution of genetic codes in the 5′ coding region, *Biotechnol. Bioeng.*, **112**, 822–6.
40. Puxbaum, V., Mattanovich, D. and Gasser, B. 2015, Quo vadis? The challenges of recombinant protein folding and secretion in *Pichia pastoris*, *Appl. Microbiol. Biotechnol.*, **99**, 2925–38.
41. Friberg, M., von Rohr, P. and Gonnet, G. 2004, Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in *Saccharomyces cerevisiae*, *Yeast*, **21**, 1083–93.
42. Hockenberry, A.J., Sirer, M.I., Amaral, L.A.N. and Jewett, M.C. 2014, Quantifying position-dependent codon usage bias, *Mol. Biol. Evol.*, **31**, 1880–93.
43. Zalucki, Y., Beacham, I. and Jennings, M. 2009, Biased codon usage in signal peptides: a role in protein export, *Trends Microbiol.*, **17**, 146–50.
44. Power, P., Jones, R., Beacham, I., Bucholtz, C. and Jennings, M. 2004, Whole genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of *Escherichia coli*, *Biochem. Biophys. Res. Commun.*, **322**, 1038–44.
45. Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. 2007, Locating proteins in the cell using TargetP, SignalP and related tools, *Nat. Protoc.*, **2**, 953–71.
46. Rice, P., Longden, I. and Bleasby, A. 2000, EMBOSS: the European Molecular Biology Open Software Suite, *Trends Genet.*, **16**, 276–7.
47. Stenico, M., Lloyd, A. and Sharp, P. 1994, Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases, *Nucleic Acids Res.*, **22**, 2437–46.
48. Nasrullah, I., Butt, A.M., Tahir, S., Idrees, M. and Tong, Y. 2015, Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution, *BMC Evol. Biol.*, **15**, 174.
49. Schattner, P., Brooks, A. and Lowe T. 2005, The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs, *Nucleic Acids Res.*, **33(suppl 2)**, W686–9.
50. Dana, A. and Tuller, T. 2014, The effect of tRNA levels on decoding times of mRNA codons, *Nucleic Acids Res.*, **42**, 9171–81.
51. Charif, D. and Lobry, J. 2007, SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis, In: Bastolla, U., Porto, M., Roman, H.E. and Vendruscolo, M., editors, *Structural approaches to sequence evolution*, Springer: Berlin, Heidelberg, pp. 207–32.
52. Pages, H., Aboyoun, P., Gentleman, R. and DebRoy, S. 2009, String objects representing biological sequences, and matching algorithms, *R package version*. **2**, 2.
53. UniProt Consortium. 2015, UniProt: a hub for protein information, *Nucleic Acids Res.*, **43**, D204–12.
54. Larkin, M.A., Blackshields, G., Brown, N.P., et al. 2007, Clustal W and Clustal X version 2.0, *Bioinformatics*, **23**, 2947–8.
55. Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. 2013, MEGA6: Molecular evolutionary genetics analysis version 6.0, *Mol. Biol. Evol.*, **30**, 2725–9.
56. Madej, T., Lanczycki, C.J., Zhang, D., et al. 2013, MMDB and VAST+: tracking structural similarities between macromolecular complexes, *Nucleic Acids Res.*, **42**, D297–303.
57. Lorenz, R., Bernhart, S.H., Zu Siederdissen, C., et al. 2011, ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 1.
58. Moriyama, E. and Powell, J. 1998, Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*, *Nucleic Acids Res.*, **26**, 3188–93.
59. Duret, L. and Mouchiroud, D. 1999, Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis, *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 4482–7.
60. Xu, C., Dong, J., Tong, C., Gong, X., Wen, Q. and Zhuge, Q. 2013, Analysis of synonymous codon usage patterns in seven different citrus species, *Evol. Bioinf. Online*, **9**, 215–28.
61. Wei, L., He, J., Jia, X., et al. 2014, Analysis of codon usage bias of mitochondrial genome in *Bombyx mori* and its relation to evolution, *BMC Evol. Biol.*, **14**, 262.
62. Jia, X., Liu, S., Zheng, H., et al. 2015, Non-uniqueness of factors constraint on the codon usage in *Bombyx mori*, *BMC Genomics*, **16**, 356.
63. Williford, A. and Demuth, J. 2012, Gene expression levels are correlated with synonymous codon usage, amino acid composition, and gene architecture in the red flour beetle, *Tribolium castaneum*. *Mol. Biol. Evol.*, **29**, 3755–66.
64. Clarke, T. and Clark, P. 2010, Increased incidence of rare codon clusters at 5′ and 3′ gene termini: implications for function, *BMC Genomics*, **11**, 118.
65. Dana, A. and Tuller, T. 2014, Properties and determinants of codon decoding time distributions. *BMC Genomics*, **15**, S13.
66. Li, Y.D., Li, Y.Q., Chen, J.S., Dong, H.J., Guan, W.J. and Zhou, H. 2006, Whole genome analysis of non-optimal codon usage in secretory signal sequences of *Streptomyces coelicolor*, *BioSystems*, **85**, 225–30.
67. Quax, T., Claassens, N., Söll, D. and van der Oost, J. 2015, Codon bias as a means to fine-tune gene expression, *Mol. Cell*, **59**, 149–61.
68. Plotkin, J. and Kudla, G. 2010, Synonymous but not the same: the causes and consequences of codon bias, *Nat. Rev. Genet.*, **12**, 32–42.

69. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. and Plotkin, J.B. 2013, Rate-limiting steps in yeast protein translation, *Cell*, **153**, 1589–601.
70. Navon, S. and Pilpel, Y. 2011, The role of codon selection in regulation of translation efficiency deduced from synthetic libraries, *Genome Biol.*, **12**, R12.
71. Goodman, D.B., Church, G.M. and Kosuri, S. 2013, Causes and effects of N-terminal codon bias in bacterial genes. *Science*, **342**, 475–9.
72. Massahi, A. and Çalık, P. 2015, In-silico determination of *Pichia pastoris* signal peptides for extracellular recombinant protein production, *J. Theor. Biol.*, **364**, 179–88.
73. Tang, H., Bao, X., Shen, Y., et al. 2015, Engineering protein folding and translocation improves heterologous protein secretion in *Saccharomyces cerevisiae*, *Biotechnol. Bioeng.*, **112**, 1872–82.
74. Yarimizu, T., Nakamura, M., Hoshida, H. and Akada, R. 2015, Synthetic signal sequences that enable efficient secretory protein production in the yeast *Kluyveromyces marxianus*, *Microb. Cell Fact.*, **14**, 1.
75. Tsuchiya, Y., Morioka, K., Shirai, J., Yokomizo, Y. and Yoshida, K. 2003, Gene design of signal sequence for effective secretion of protein, In: *Nucleic Acids Symposium Series*, Vol. **3**. Oxford University Press, pp. 261–2.
76. Chartier, M., Gaudreault, F. and Najmanovich, R. 2012, Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events, *Bioinformatics*, **28**, 1438–45.
77. Shabalina, S., Spiridonov, N. and Kashina, A. 2013, Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity, *Nucleic Acids Res.*, **41**, 2073–94.
78. Nilsson, O.B., Hedman, R., Marino, J., et al. 2015, Cotranslational protein folding inside the ribosome exit tunnel, *Cell Rep.*, **12**, 1533–40.
79. Wilson, D. and Beckmann, R. 2011, The ribosomal tunnel as a functional environment for nascent polypeptide folding and translational stalling, *Curr. Opin. Struct. Biol.*, **21**, 274–82.