Contents lists available at ScienceDirect

BioSystems

journal homepage: http://www.elsevier.com/locate/biosystems

Potential role of the *X* circular code in the regulation of gene expression

Julie D. Thompson [a,**], Raymond Ripp [a], Claudine Mayer [a,b,c], Olivier Poch [a], Christian J. Michel [a,*]

[a] *Department of Computer Science, ICube, CNRS, University of Strasbourg, Strasbourg, France*
[b] *Unité de Microbiologie Structurale, Institut Pasteur, CNRS, 75724, Paris Cedex 15, France*
[c] *Université Paris Diderot, Sorbonne Paris Cité, 75724, Paris Cedex 15, France*

ARTICLE INFO

Keywords:
Gene expression
Codon optimization
Codon usage
Genetic code
Circular code

ABSTRACT

The *X* circular code is a set of 20 trinucleotides (codons) that has been identified in the protein-coding genes of most organisms (bacteria, archaea, eukaryotes, plasmids, viruses). It has been shown previously that the *X* circular code has the important mathematical property of being an error-correcting code. Thus, motifs of the *X* circular code, i.e. a series of codons belonging to *X* and called *X* motifs, allow identification and maintenance of the reading frame in genes. *X* motifs are significantly enriched in protein-coding genes, but have also been identified in many transfer RNA (tRNA) genes and in important functional regions of the ribosomal RNA (rRNA), notably in the peptidyl transferase center and the decoding center. Here, we investigate the potential role of *X* motifs as functional elements of protein-coding genes. First, we identify the codons of the *X* circular code which are frequent or rare in each domain of life (archaea, bacteria, eukaryota) and show that, for the amino acids with the highest codon bias, the preferred codon is often an *X* codon. We also observe a correlation between the 20 *X* codons and the optimal codons/dicodons that have been shown to influence translation efficiency. Then, we examined recently published experimental results concerning gene expression levels in diverse organisms. The approach used is the analysis of *X* motifs according to their density $d_s(X)$, i.e. the number of *X* motifs per kilobase in a gene sequence *s*. Surprisingly, this simple parameter identifies several unexpected relations between the *X* circular code and gene expression. For example, the *X* motifs are significantly enriched in the minimal gene set belonging to the three domains of life, and in codon-optimized genes. Furthermore, the density of *X* motifs generally correlates with experimental measures of translation efficiency and mRNA stability. Taken together, these results lead us to propose that the *X* motifs may represent a genetic signal contributing to the maintenance of the correct reading frame and the optimization and regulation of gene expression.

1. Introduction

The standard genetic code represents one of the greatest discoveries of the 20th century (Crick et al., 1961). All known life on Earth uses the (quasi-) same triplet genetic code to control the translation of genes into functional proteins. The fact that there are 64 possible nucleotide triplet combinations but only 20 amino acids to encode, means that the genetic code is redundant and most amino acids are encoded by more than one codon. For instance, the amino acid glutamine is coded by two codons {CAA,CAG} and alanine is coded by four codons {GCA,GCC,GCG,GCT}. This redundancy allows for the encoding of supplementary information in addition to the amino acid sequence (Weatheritt and Babu, 2013; Maraia and Iben, 2014), and significant efforts have been applied recently to understand the multiple layers of information or 'codes within the code' (Babbitt et al., 2018; Bergman and Tuller, 2020) that can be exploited to increase the versatility of genome decoding: for example, nucleosome positioning codes (Eslami-Mossallam et al., 2016), the histone code (Prakash and Fournier, 2018), the splicing code (Baralle and Baralle, 2018), mRNA degradation sites (Cakiroglu et al., 2016), or the protein folding code (Faure et al., 2017; Seligmann and Warthi, 2017), to name but a few.

The genetic code has a non-overlapping structure, which means that the codons in a DNA sequence must be decoded in the correct reading frame in order to produce the correct amino acid sequence. Reading the

[*] Corresponding author.
[**] Corresponding author.

E-mail addresses: thompson@unistra.fr (J.D. Thompson), raymond.ripp@unistra.fr (R. Ripp), mayer@pasteur.fr (C. Mayer), olivier.poch@unistra.fr (O. Poch), c.michel@unistra.fr (C.J. Michel).

https://doi.org/10.1016/j.biosystems.2021.104368
Received 9 December 2020; Received in revised form 18 January 2021; Accepted 20 January 2021
Available online 7 February 2021

sequence out-of-frame can have severe effects, including termination of translation if a stop codon is encountered or production of a non-functional protein sequence otherwise. The "ambush hypothesis" proposes that out-of-frame stop codons regulate translation by allowing rapid termination of frameshifted translations and it has been suggested that codons that can form stop codons when a frameshift occurs may be selected for (Seligmann and Pollock, 2004; Seligmann, 2019).

Here, we focus on an important class of genome codes, called the circular codes, first introduced by Arquès and Michel (1996) and reviewed in Michel (2008), and Fimmel and Strüngmann (2018). In coding theory, a circular code is also known as an error-correcting code or a self-synchronizing code, since no external synchronization is required for reading frame identification. In other words, circular codes have the ability to detect and maintain the correct reading frame. For example, comma-free codes are a particularly efficient subclass of circular codes, where the reading frame is detected by a single codon. The genetic code was originally proposed to be a comma-free code in order to explain how a sequence of codons could code for 20 amino acids, and at the same time how the correct reading frame could be retrieved and maintained (Crick et al., 1957). However, it was later proved that the modern genetic code could not be a comma-free code (Nirenberg and Matthaei, 1961), when it was discovered that *TTT*, a codon that cannot belong to a comma-free code, codes for phenylalanine. Furthermore, from a theoretical point of view (see Section 5. in Michel, 2020): (i) the comma-free codes cannot satisfy the coding condition of 20 amino acids (at most 13 amino acids can be coded by the 408 maximal comma-free codes); and (ii) the self-complementary comma-free codes have an incomplete circularity property (reading frame retrieval) as 12 trinucleotides among 60 must be ignored (the maximality of comma-free codes which are self-complementary, or C^3, or C^3 self-complementary, is only 16 trinucleotides). Other circular codes are less restrictive than comma-free codes, as a frameshift of 1 or 2 nucleotides in a sequence entirely consisting of codons from a circular code will not be detected immediately but after the reading of a certain number of nucleotides.

By excluding the four periodic codons {AAA,CCC,GGG,TTT} and by assigning each codon to a preferential frame (i.e. each codon is assigned to the frame in which it occurs most frequently compared to the other frames), a circular code was identified in the reading frame of protein-coding genes from eukaryotes and prokaryotes (Arquès and Michel, 1996; Michel, 2017). This so-called *X* circular code consists of 20 codons (Fig. 1):

$$X=\{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,GAG,GAT,\\ GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC\} \qquad (1)$$

and codes for the following 12 amino acids (three and one letter notation):

{Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Phe, Thr, Tyr, Val} ={A, N, D, Q, E, G, I, L, F, T, Y, V}.

Other circular codes, and notably variations of the common *X* circular code, are hypothesized to exist in different organisms (Frey and Michel, 2003, 2006; Michel, 2017).

The *X* circular code has important mathematical properties, in particular it is self-complementary (Arquès and Michel, 1996), meaning that if a codon belongs to *X* then its complementary trinucleotide also belongs to *X*. Like the comma-free codes, the *X* circular code also has the property of synchronizability. It has been shown that, in any sequence generated by the *X* circular code, at most 13 consecutive nucleotides are enough to always retrieve the reading frame (Arquès and Michel, 1996). In other words, any *X* motif containing 4 consecutive *X* codons is sufficient to determine the correct reading frame. Fig. 1 in Michel and Thompson (2020) presents the hierarchy of a set of words, a code, a circular code and a comma-free code with their circularity property (reading frame retrieval). More formal definitions of the mathematical properties (theorems) of the *X* circular code are available in a number of reviews (Michel, 2008; Fimmel and Strüngmann, 2018) and recent works (Fimmel et al., 2019, 2020).

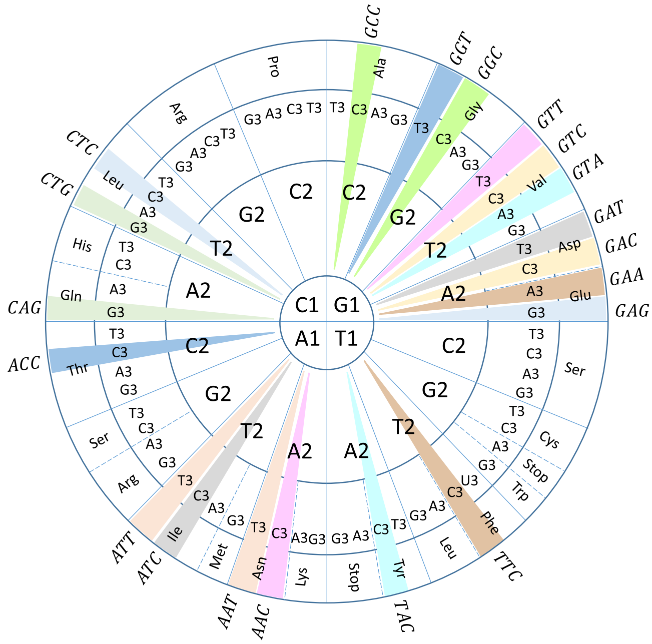


Fig. 1. Circular representation of the genetic code, adapted from Grosjean and Westhof (2016), with the 20 codons of the *X* circular code shown on the circumference. The numbers after the nucleotides indicate their position in the codon. *X* codons that are complementary to each other are highlighted in the same color. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

The hypothesis of the *X* circular code in genes is supported by evidence from several statistical analyses of modern genomes. We previously showed in a large-scale study of 138 eukaryotic genomes that *X* motifs (defined as series of at least 4 codons from the *X* circular code) are found preferentially in protein-coding genes compared to non-coding regions with a ratio of ~8 times more *X* motifs located in genes (El Soufi and Michel, 2016). More detailed studies of the complete gene sets of yeast and mammal genomes confirmed the strong enrichment of *X* motifs in genes and further demonstrated a statistically significant enrichment in the reading frame compared to frames 1 and 2 (p-value < 10^{-10}) (Michel et al., 2017; Dila et al., 2019a). In addition, it was shown that most of the mRNA sequences from these organisms (e.g. 98% of experimentally verified genes in *S. cerevisiae*) contain *X* motifs. Intriguingly, conserved *X* motifs have also been found in many tRNA genes (Michel, 2013), as well as in important functional regions of the 16S/18S ribosomal RNA (rRNA) from bacteria, archaea and eukaryotes (Michel, 2012; Dila et al., 2019b), which suggest their involvement in universal gene translation mechanisms. More recently, a circular code periodicity 0 modulo 3 was identified in the 16S rRNA, covering the region that corresponds to the primordial proto-ribosome decoding center and containing numerous sites that interact with the tRNA and messenger RNA (mRNA) during translation (Michel and Thompson, 2020). Based on these observations, it has been proposed that the *X* circular code represents an ancestor of the modern genetic code that was used to code for a smaller number of amino acids and simultaneously identify and maintain the reading frame (Dila et al., 2019b). Intriguingly, the theoretical minimal RNA rings, short RNAs designed to code for all coding signals without coding redundancy among frames, are also biased for codons from the X circular code (Demongeot and Seligmann, 2019). The question remains of whether the X motifs observed in modern genes are simply a vestige of an ancient code that might have existed in the early stages of cellular life, or whether they still play a role in the complex translation systems of extant organisms.

In this work, the coverage of the X circular code or the X motifs in genes is analyzed using a (very) simple density parameter. Unexpectedly, this coverage parameter identifies several relations between the X circular code and translation efficiency and/or kinetics. We first investigate whether a correlation exists between the X circular code and the 'optimal' codons/dicodons associated with increased gene translation efficiency and mRNA stability. Then, we examine the recent evidence resulting from high-throughput technologies such as ribosome profiling, and demonstrate that the presence of X motifs in genes can be used as a predictor of gene expression level. Taken together, these observations provide evidence supporting the idea that motifs from the X circular code represent a new genetic signal, participating in the maintenance of the correct reading frame and the optimization and regulation of gene expression.

2. Method

The definitions and mathematical properties of circular codes are not recalled here since they are not necessary to understand the methods and results obtained in this work. We refer the reader to the reviews (Michel, 2008; Fimmel and Strüngmann, 2018) and the recent works (Fimmel et al., 2019, 2020) for this information.

2.1. Definition of the X motif density parameter

We define an X motif $m_s(X)$ as a series of at least 4 consecutive codons of the circular code X (defined in (1)) in the reading frame of a gene sequence s. For example, $m_s(X) =$ CAGGACTACGTCGAC is an X motif since CAG, GAC, TAC and GTC are codons of X. It is important to remember that any X motif with at least 4 consecutive X codons always allows the reading frame to be retrieved. Let $N(m_s(X))$ be the number of X motifs $m_s(X)$ in a gene sequence s of nucleotide length l_s. Then the density $d_s(X)$ of X motifs in a gene sequence s is defined by the number of X motifs per kilobase in s, i.e.

$$d_s(X) = \frac{1000}{l_s} N(m_s(X)). \tag{2}$$

This density $d_s(X)$ (Equation (2)) in a sequence s can easily be extended to a density $d_{\mathscr{S}}(X)$ in a set $\mathscr{S}$ of gene sequences s by dividing the total number of X motifs by the total nucleotide length $l_{\mathscr{S}} = \sum_{s \in \mathscr{S}} l_s$, i.e.

$$d_{\mathscr{S}}(X) = \frac{1000}{l_{\mathscr{S}}} \sum_{s \in \mathscr{S}} N(m_s(X)). \tag{3}$$

These densities $d_s(X)$ and $d_{\mathscr{S}}(X)$ are normalized parameters allowing to compare the coverage of X motifs in sequences of different lengths and in sequence populations of different sizes.

In order to evaluate the statistical significance of the obtained results, we also define 100 random codes R with similar properties to the X circular code, using the method described in Dila et al. (2019a). We then identified R random motifs $m_s(R)$ from these codes in the gene sequences and calculated the densities $d_s(R)$ and $d_{\mathscr{S}}(R)$ of R motifs in a gene sequence s and a set $\mathscr{S}$ of gene sequences s, respectively, as for the X motifs.

2.2. Data sources for analysis of optimal codons

As a measure of the optimality of each codon, we used the codon stabilization coefficient (CSC), defined by Bazzini et al. (2016) as the Pearson correlation coefficient between the occurrence of each codon and the half-life of each mRNA. Thus, codons found more frequently in genes with longer mRNA half-lives have higher CSC values. The 61 coding codons can then be ranked according to their CSC scores in different organisms. We obtained the CSC rankings for each codon from previous studies in four species: zebrafish (Bazzini et al., 2016), *Xenopus* (Bazzini et al., 2016), *Drosophila* (Burow et al., 2018) and *S. cerevisiae* (Hanson and Coller, 2018). We then calculated the mean CSC ranking for each codon in these four species.

Dicodons associated with reduced protein expression in *S. cerevisiae* were taken from a previous study (Gamble et al., 2016). Dicodons associated with low abundance or high abundance proteins were obtained from a previous study (Diambra, 2017).

2.3. Minimal gene set analysis

The minimal gene set of 81 genes conserved in all species was obtained from a previous study (Koonin, 2000). We used the *Mycoplasma genitalium* genes provided in the original study as a query, and searched for orthologues in the reference set of complete genomes for 317 species (144 eukaryotes, 142 bacteria and 31 archaea) in the OrthoInspector 3.0 database (Nevers et al., 2019). This resulted in a set of 15822 protein sequences (5503 eukaryotes, 9205 bacteria and 1114 archaea). For each protein sequence, we retrieved the mRNA sequences from the Uniprot database (www.uniprot.org) and identified all X motifs in the reading frame with a minimum length of 4 codons, using in-house developed software.

2.4. Data sources for codon-optimized genes

Experimental data for synthetic genes re-designed for optimized protein expression were obtained from the SGDB database (Wu et al., 2007). SGDB contains sequences and associated experimental information for synthetic (artificially engineered) genes from published peer-reviewed studies. We selected the gene entries where the synthetic sequence contained only synonymous codon changes, and where experimental protein expression levels were available for both the wild type and the synthetic gene. This resulted in a set of 45 gene pairs (wild type and synthetic gene), for which we identified all X motifs in the reading frame with a minimum length of 4 codons, using in-house developed software as before.

2.5. Estimation of translation rates based on ribosome profiling data

Computational estimations of translation rates for 5450 *S. cerevisiae* genes were obtained from a previous study (Diament et al., 2018). The authors performed a simulation of translation based on the totally asymmetric simple exclusion process (TASEP) model, using experimental measurements of the number of ribosomes on each transcript as well as RNA copy numbers to calibrate the parameters. For each of the 5450 genes, we identified the X motifs using in-house developed software.

3. Results

In this section, we first compare the 20 codons of the *X* circular code with the optimal codons and dicodons that have been shown to influence translation efficiency. Then, using previously published experimental data, we investigate whether a correlation exists between the density of *X* motifs in genes and the level of gene expression.

3.1. *X* codons and codon usage

Synonymous codons are observed with different frequencies between species, a phenomenon known as codon usage bias (CUB). This means that in different species, there is a preference for different codons with some being used more frequently than others. CUB is also observed within an organism at the gene level, since codon frequencies can vary between genes in the same genome. The ways in which CUB influences different aspects of protein production have been studied for a long time (Grantham et al., 1981) and it has become clear that codon choice has effects at many stages, including transcription (Zhou et al., 2016), translation efficiency (Qian et al., 2012), mRNA stability (Presnyak et al., 2015), protein folding (Buhr et al., 2016) and protein function (Bali and Bebok, 2015).

In Section 4 in Michel (2020), several theoretical properties are provided in detail to explain that the codon usage parameter is unable to identify a (maximal) circular code. In other words, the circularity statistical property (reading frame retrieval) of a code cannot be associated with the codon usage. In this section, we will verify this assertion from a statistical (experimental) point of view for the circular code *X*. It is important to stress that if such an assertion is observed for the circular code *X*, which has the highest occurrence in genes on average, then this assertion is of course verified for the other circular codes, in particular with the variant circular codes specific to some genes (Frey and Michel, 2003, 2006; Michel, 2017). A second objective of this section is to identify some relationships between the codons of the circular code *X* and codon usage. We recall here that the codons of the circular code *X* are found preferentially in the reading frame as compared to the other frames.

A number of factors are known to influence codon usage in individual organisms, including nucleotide frequencies (GC content), especially at the third codon position, also known as the wobble position. For example, in the codon usage tables provided in appendix Table A1, in the fungi *S. cerevisiae*, A or T is preferred at wobble positions, whereas in the animalia *Drosophila*, the fungi *Neorospora crassa* and the bacteria *Escherichia coli*, C or G is more frequent (Palidwor et al., 2010). If we consider the nucleotide frequencies for the 20 codons of the *X* circular code, no overall bias is observed (each nucleotide is used with the same frequency), however some differences can be seen at the different positions (Table 1). For example, at the wobble position of *X* codons, C is the most frequent nucleotide, followed by T. In contrast, at the first position, G is strongly preferred and is present in half of the *X* codons. Interestingly, the codons coding for the most ancient amino acids are also enriched in G in the first position (Trifonov, 2000).

Dinucleotide frequencies are also known to affect codon usage in gene-coding regions, since a combination of mutational biases against CpG in genomic DNA and selection against TpA for increased stability in mRNA gives rise to a non-random pattern for each of the 16 possible dinucleotides (Simmonds et al., 2013). Table 2 shows the dinucleotide frequencies in *X* codons. Interestingly, the *X* circular code has no codons containing the CG dinucleotide, and only two codons containing the TA dinucleotide: GTA (Val) and TAC (Tyr). RNA rings, a theoretical construct that mimics natural genomes, are also devoid of CG dinucleotides (Demongeot et al., 2020).

Table 1
Nucleotide numbers in the 3 positions of 20 codons of the *X* circular code.

	1st position	2nd position	3rd position	Total
A	5	8	2	15
C	3	2	10	15
G	10	2	3	15
T	2	8	5	15
Total	20	20	20	60

Table 2
Dinucleotide numbers in the 20 codons of the *X* circular code, either at the 1st and 2nd positions, or at the 2nd and 3rd positions.

Dinucleotide	*X* circular code	Dinucleotide	*X* circular code
AA	3	GA	4
AC	4	GC	2
AG	2	GG	2
AT	4	GT	4
CA	1	TA	2
CC	2	TC	4
CG	0	TG	1
CT	2	TT	3

At the codon level, we compared the 20 codons belonging to the *X* circular code with the CUB observed in the three domains of life (Fig. 2). Although the frequencies of each codon are different in the three domains, some universal trends can be seen. For example, many of the codons that are observed frequently in all three domains belong to *X*, including GAC, GAT, GAA, GAG, coding for Asp and Glu. This might be explained in two different ways: first that the *X* codons are simply the codons used more frequently in extant organisms, or second that these codons are found more frequently because they belong to the *X* circular code. Nevertheless, the relation between codon frequency and *X* codons is more complex, since some codons that are rare in at least one of the domains also belong to *X*, such as GTA (Val), ACC (Thr), GGT (Gly) and TAC (Tyr) (in agreement with the assertion that the circularity property of a code cannot be identified with the codon usage).

At the amino acid level, some amino acids like Cys, His, Met and Trp are generally rarer in all three domains and none of the codons coding for these amino acids belong to the *X* circular code. In fact, as stated earlier, the *X* circular code codes for only 12 amino acids. Of these, 8 correspond to the early amino acids (EAA) that were identified in prebiotic chemistry experiments as well as in meteorites, as discussed previously in Michel et al. (2017). Another interesting observation concerns the amino acids with the highest codon bias (i.e. where one codon is preferred over the other synonymous codons). Here, the preferred codon is often an *X* codon, even when different organisms have different preferred codons. For example, CTC is the most frequent codon coding for Leu in archaea, while CTG coding for Leu is preferred in eukaryotes and bacteria. Both CTC and CTG belong to the *X* code (see again the above assertion about the circularity property of a code that cannot be identified with the codon usage).

Based on these observations, we hypothesize that the *X* circular code may be another moderator of codon usage.

3.2. *X* codons correlate with the codons associated with increased expression

We compared the 20 codons that belong to the *X* circular code with the ‘codon optimality code’ resulting from various statistical and experimental studies in metazoan (Bazzini et al., 2016), as well as in *S. cerevisiae* (Hanson and Coller, 2018). In these studies, the codon stabilization coefficient (CSC) was used as a robust and conserved measure of how individual codons contribute to shape mRNA stability and translation efficiency. Fig. 3 shows the mean ranking of optimal codons, according to the CSC score, from four different experiments (in *S. cerevisiae*, zebrafish, *Xenopus* and *Drosophila*), where the highest ranking codon is the most optimal one. The *X* codons are ranked significantly higher than non-*X* codons (i.e. the 41 coding codons which do not belong to the circular code *X*), according to a Mann-Whitney

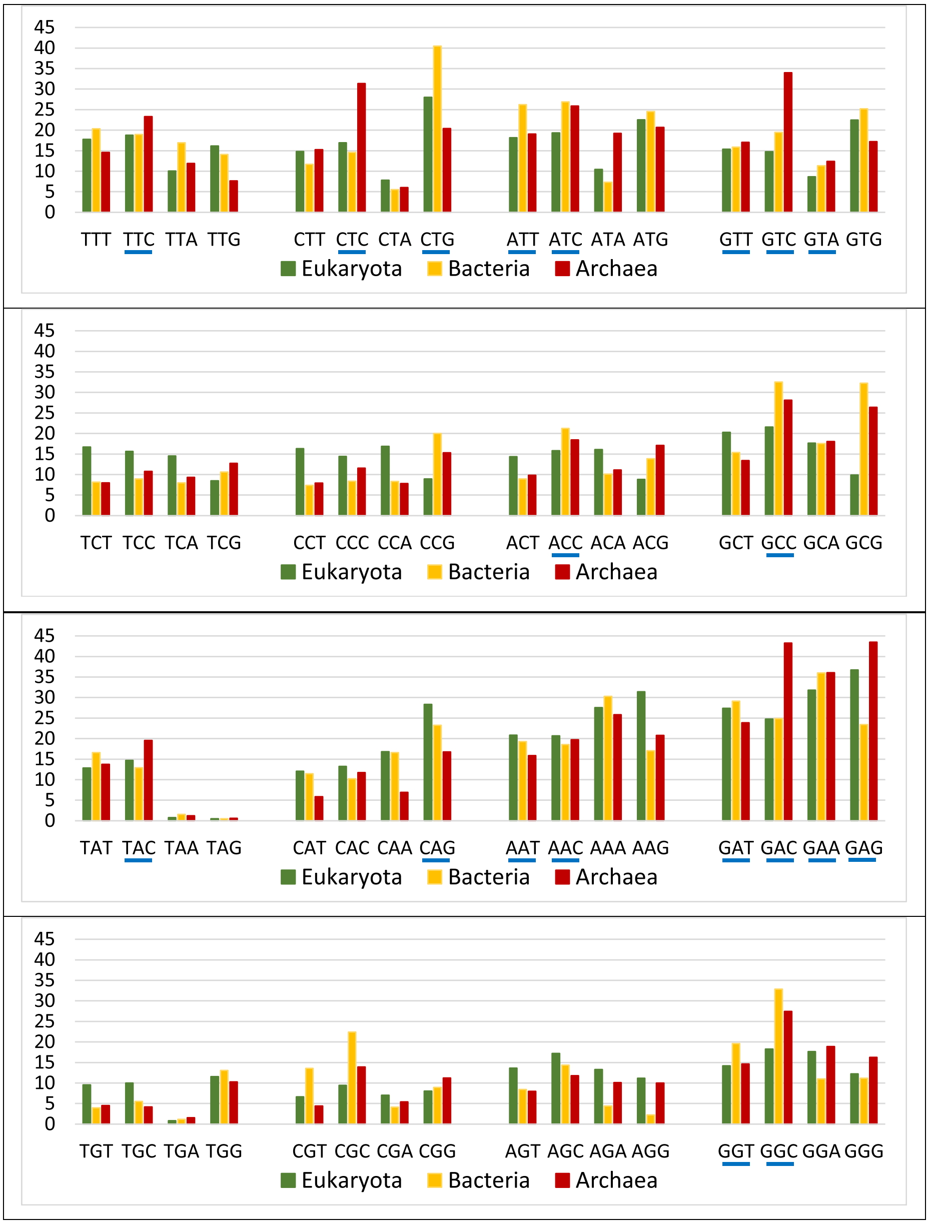


Fig. 2. Codon usage bias (CUB) in the three domains of life. The graphs show the average codon frequencies per 1000 codons observed in each domain of life (data extracted from the HIVE-CUTs database at https://hive.biochemistry.gwu.edu/cuts/). Codons belonging to the *X* circular code are underlined in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

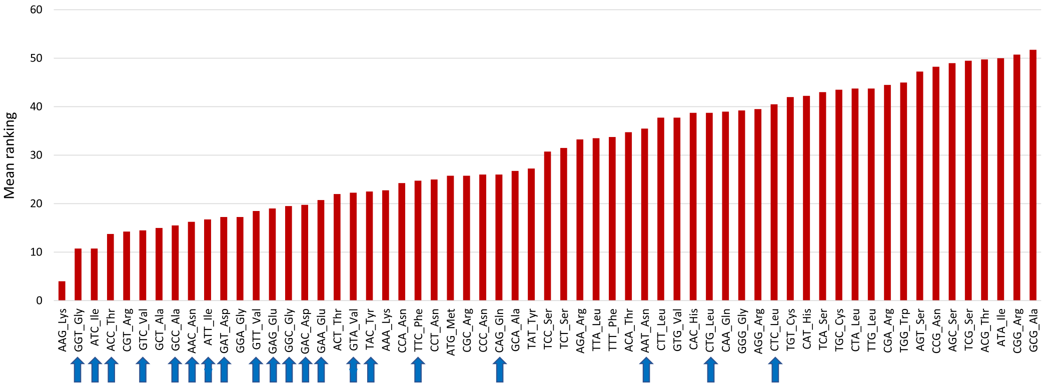


Fig. 3. Optimal codons for translation elongation rate and mRNA stability in different eukaryotic species (*S. cerevisiae*, zebrafish, *Xenopus* and *Drosophila*). Codons are ordered according to their mean ranking obtained in four different experiments. Codons belonging to the *X* circular code are identified by a blue arrow. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 3
Dicodons enriched in low or high abundance proteins, in two different studies: Gamble et al. (2016) and Diambra (2017). *X* codons are highlighted in blue.

Dicodons with low abundance		Dicodons with high abundance	
Gamble et al.	Diambra	Diambra	
AGG-CGA	AAA-ATA	**AAC-AAC**	**GAC-ACC**
AGG-CGG	**AAT**-GCA	**AAC**-AAG	**GAC-TAC**
ATA-CGA	**AAT**-TGG	**AAC-ACC**	**GAT**-GCT
ATA-CGG	AGT-AAG	AAG-TCC	**GCC-AAC**
CGA-ATA	AGT-GTG	**ACC-AAC**	**GCC**-AAG
CGA-CCG	ATA-**GGT**	**ACC**-AAG	**GCC-ACC**
CGA-CGA	**ATT**-AAA	**ACC-ACC**	**GCC-ATC**
CGA-CGG	CAA-AGT	**ACC-ATC**	**GCC-GCC**
CGA-**CTG**	**CAG**-AAA	**ACC-ATT**	**GGT-GTC**
CGA-GCG	**GAA**-AGT	**ACC-GCC**	**GTC**-AAG
CTC-CCG	**GAA**-CTA	**ACC-TTC**	**GTC-ACC**
CTG-ATA	GCA-TTT	**ATC-AAC**	**GTC-ATC**
CTG-CCG	TAT-AAA	**ATC**-AAG	**GTT-GCC**
CTG-CGA	TAT-CCG	**ATC-ACC**	**TAC-AAC**
GTA-CCG	TTT-**CAG**	**ATC-ATC**	**TAC**-AAG
GTA-CGA	TTT-TTT	**ATT-GCC**	TCC-**ACC**
GTG-CGA		CCA-CCA	**TTC-AAC**
		CGT-CGT	**TTC**-AAG
		GAC-AAC	**TTC-ACC**
		GAC-AAG	**TTC-ATC**

signed rank test (z-score = 4.3, p-value < 0.00001). In other words, optimal codons for mRNA stability and elongation rate are significantly enriched in *X* codons.

A method was previously developed for measuring the Reading Frame Retrieval (RFR) probability for the 20 *X* codons (originally called stability, Table 7 in Ahmed et al., 2010, Section 2.2 and 1st row of Table 1 in Michel and Seligmann, 2014). We do not observe a significant correlation between the RFR values and the mean ranking (Fig. 3) for the 20 *X* codons: Pearson correlation coefficient $r = 0.51$ with a p-value = 0.022 and Spearman rank correlation coefficient $\rho = 0.36$ with a p-value = 0.115.

Table 4
Number of *X* codons and non-*X* codons in low or high abundance proteins (from Table 3). There is a strong dependence between *X* codons and protein abundance with p-values < 0.0001 both with the Fischer exact and Chi-squared tests.

	Low abundance	High abundance	Total
X codons	15	64	79
Non-*X* codons	51	16	67
Total	66	80	146

Table 5
Number of *X* dicodons and non-*X* dicodons in low or high abundance proteins (from Table 3). There is a strong dependence between *X* dicodons and protein abundance with p-values < 0.0001 both with the Fischer exact and Chi-squared tests.

	Low abundance	High abundance	Total
X dicodons	0	27	27
Non-*X* dicodons	33	13	46
Total	33	40	73

3.3. *X* codons correlate with the dicodons associated with increased expression

In recent years, emerging evidence has shown that translational rates may be encoded by dicodons rather than single codons (Gamble et al., 2016; Diambra, 2017; Guo et al., 2012). For example, a large-scale screen in *S. cerevisiae* (Gamble et al., 2016) assessed the degree to which codon context modulates eukaryotic translation elongation rates beyond effects seen at the individual codon level. The authors screened yeast cell populations housing libraries containing random sets of triplet codons within an ORF encoding superfolder Green Fluorescent Protein (GFP). They found that 17 dicodons were strongly associated with reduced GFP expression, i.e. associated with a substantial reduction of the translation elongation rate. This set included the known inhibitory dicodon CGA-CGA and was enriched for codons decoded by wobble interactions. Of these 17 dicodons associated with slower translation elongation rates, none are composed of 2 *X* codons (Table 3).

A subsequent statistical analysis of coding sequences of nine organisms (Diambra, 2017) identified dicodons with significant different frequency usage for coding either lowly or highly abundant proteins. The working hypothesis was that sequences encoding abundant proteins should be optimized, in the sense of translation efficiency. 16 dicodons were identified with a preference for low abundance proteins, while 40 dicodons presented a preference for high abundance proteins. None of the 16 dicodons associated with low abundance proteins are composed

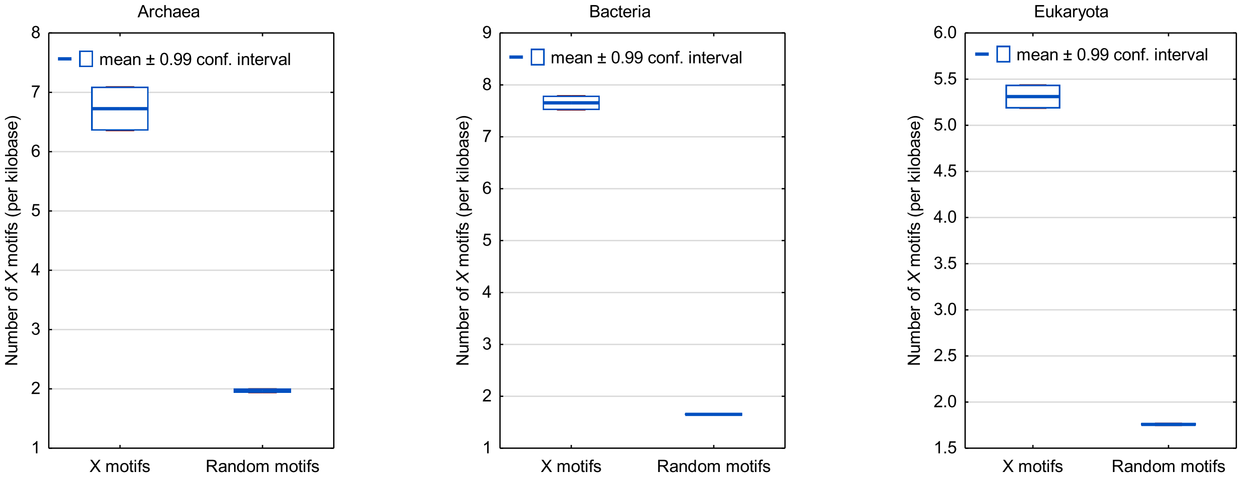


Fig. 4. Density $d_{\mathscr{X}}(X)$ of X motifs (Equation (3), i.e. number of X motifs per kilobase) in the mRNA sequences of the 'minimal gene set'. The distributions of the number of X motifs per kilobase identified in the sequences from the three domains of life are indicated by boxplots representing the mean number with a ± 0.99 confidence interval. The distributions of the number of R random motifs per kilobase (as in previous works in Michel et al., 2017; Dila et al., 2019a) identified in the same sequences are shown for statistical evaluation. There is a very strong statistical significance as confirmed by a one-sided Student's t-test with a p-value $< 10^{-100}$ for each set of sequences from archaea, bacteria and eukaryote.

of 2 X codons (Table 3). In contrast, 27 of the 40 dicodons associated with high abundance proteins correspond to 2 X codons, and only 3 dicodons do not contain any X codons (Table 3).

In order to evaluate the statistical significance of X codons with an increased protein expression, we computed from Table 3, the number of X codons and non-X codons in low or high abundance proteins (Table 4) and the number of X dicodons and non-X dicodons in low or high abundance proteins (Table 5). Fischer exact and Chi-squared tests show a strong dependence between X codons or X dicodons, and protein abundance with all the p-values < 0.0001.

These recent studies support the idea that codons in coding sequences are likely arranged in an organized way, and that the local sequence context contributes to the effects of codon usage bias on gene regulation. Strikingly, our observations support the hypothesis that codon context may be linked in some way to the X circular code. In the next section, we describe more detailed analyses that test this hypothesis further.

3.4. *X* motifs are enriched in the minimal gene set

Based on the increasing evidence of the importance of codon context (Guo et al., 2012; Clarke and Clark, 2008; Brar, 2016; Chevance and Hughes, 2017; Sharma et al., 2019), we hypothesized that if the X circular code plays a role in gene regulation, then we might expect to see a non-random use, or 'clusters', of X codons along the length of the gene. In previous works (Michel et al., 2017; Dila et al., 2019a), we defined an X motif as a series of consecutive X codons (of length at least 4 codons as 13 nucleotides always retrieve the reading frame, see Introduction) in a gene sequence and searched for such X motifs in the reading frames of different genes. This approach allowed us to demonstrate that the reading frames of genes in yeasts and in mammals are significantly enriched in such X motifs. To test the hypothesis that the X motifs represent a more universal signature, we analyzed a set of 81 genes that were previously defined as a 'minimal gene set' (Koonin, 2000). At that

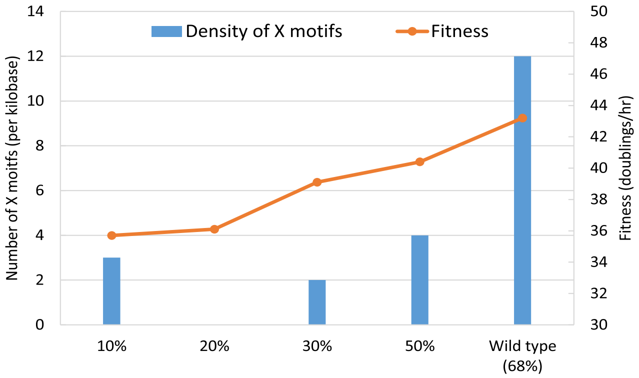


Fig. 6. Histogram of the density $d_{\mathscr{X}}(X)$ of X motifs (Equation (3), i.e. number of X motifs per kilobase) in the recoded version of the gene 10A from *Escherichia coli* K-12, compared to the wild type sequence. The orange plot indicates the viral fitness values corresponding to each construct. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

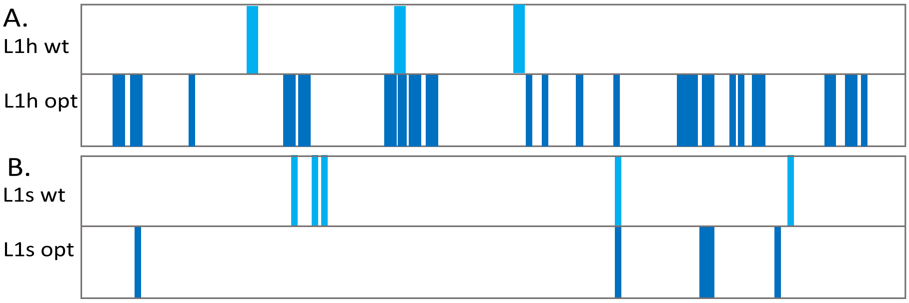


Fig. 5. Schematic view of the X motifs found in codon-optimized genes: A. L1h gene from human papillomavirus (appendix Figure A1.A; Leder et al., 2001). B. L1s gene from human papillomavirus (appendix Figure A1.B; Warzecha et al., 2003). The X motifs in the wild type sequence are shown in light blue, and the X motifs in the codon optimized sequences in dark blue. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

time, the 'minimal gene set' genes were found to be conserved in all species. We used the *Mycoplasma genitalium* genes provided in the original study, as well as 15,822 orthologous sequences (5503 eukaryotes, 9205 bacteria and 1114 archaea), and identified all *X* motifs in the reading frame with a minimum length of 4 codons. Fig. 4 shows the density $d_{\mathscr{S}}(X)$ of *X* motifs (Equation (3)) in the mRNA sequences. To evaluate the significance of the enrichment, as in previous works (Michel et al., 2017; Dila et al., 2019a), we used a randomization model in which we generated 100 random codes that preserved most of the properties to the *X* code, except the circularity. We then identified all random motifs from the 100 random codes and calculated mean values for the 100 codes.

The density of *X* motifs found in the minimal gene set sequences belonging to the three domains of life, is significantly higher than the density of random motifs according to a one-sided Student's *t*-test (*p*-value $< 10^{-100}$) for each set of sequences from archaea, bacteria and eukaryota. This study demonstrates that *X* motifs are significantly enriched in the minimal gene set, and seem to be a universal feature of gene sequences in all three domains of life.

3.5. *X motifs are enriched in codon-optimized genes*

If *X* motifs modify the codon usage in favor of optimal codons for translational efficiency, then we would expect that increasing the number of *X* motifs in a gene would increase the expression level. In an indirect way, we have shown that this is indeed the case. We previously showed that synthetic genes, which were re-designed for optimized protein expression, generally have more *X* motifs (Dila et al., 2019a). Fig. 5A and appendix Figure A1.A show an example of the protein L1h from human papillomavirus (HPV-16), optimized for expression in mammalian cell lines and leading to significantly increased expression (Leder et al., 2001). Here, the wild type gene contains only 3 *X* motifs, while the optimized gene construct has a total of 21 *X* motifs. It is important to note that classical codon optimization strategies do not always increase protein expression levels. Fig. 5B and appendix Figure A1.B show another example involving the L1s protein from human papillomavirus (HPV-11) optimized for expression in the potato *Solanum tuberosum* (Warzecha et al., 2003). In this case, only a low level of L1 expression was observed for the codon-optimized gene. In this example, we did not observe a significant difference between the number of *X* motifs in the wild type and optimized sequences (5 *X* motifs in the wild type gene compared to 4 *X* motifs in the optimized construct).

Codon replacement strategies have also been applied to the design of attenuated viruses, although in this case frequent codons are replaced with rare ones. Using quantitative proteomics and RNA sequencing, the molecular basis of attenuation in a strain of bacteriophage T7 (*Escherichia coli* K-12) was investigated (Jack et al., 2017). The authors engineered the *E. coli* major capsid protein gene (gene 10A) to carry different proportions of suboptimal, rare codons. Transcriptional effects of the recoding were not observed, but proteomic observations revealed that translation was halved for the completely recoded major capsid gene, with subsequent effects on virus fitness (measured as doublings/hour). We obtained the sequences with 10%, 20%, 30% and 50% recoding from Bull et al. (2012) and identified the density $d_{\mathscr{S}}(X)$ of *X* motifs (Equation (3)) in each construct. Fig. 6 clearly shows the correlation between the fitness obtained for each recoded sequence and the density of *X* motifs observed. The authors suggested that recoding of gene 10A reduced capsid protein abundance probably by ribosome stalling rather than ribosome fall-off.

Table 6
Comparison of density $d_{\mathscr{S}}(X)$ of *X* motifs (Equation (3), i.e. number of *X* motifs per kilobase) in the wild type gene *frq* and different optimized constructs for the *Neurospora crassa* FRQ protein. In the 'mid opt' constructs, only the non-preferred codons were changed; for 'full opt' constructs, every codon was optimized.

Region	*frq* construct	Density of *X* motifs
N-terminal (1–164)	wild type	4.1
	mid opt	6.1
	full opt	10.2
Middle (185–530)	wild type	3.9
	full opt	5.8

In general, codon optimization is a successful strategy for improving protein expression in heterologous systems. However, simply replacing all rare codons by frequent codons can have negative effects *in vivo* (Gingold and Pilpel, 2011). Rare codons have the potential to slow down the translation elongation rate, due to the relatively long dwell time of the ribosome while searching for rare tRNAs. Several studies have suggested that gene-wide codon bias in favor of slowly translated codons serves as a regulatory means to obtain low expression levels of protein when desired, for example, in the case of regulatory genes, or where excess of the protein may be detrimental or lethal to the cell. An example, in *Neurospora crassa*, demonstrated that codon optimization of the central clock protein FRQ actually abolished circadian rhythms (Zhou et al., 2013). Different optimized constructs of the wild type gene *frq* were used in the study, where either the N-terminal end (codons 1–164) or the middle region (codons 185–530) was optimized. All optimized constructs gave higher levels of FRQ protein, although this led to different structural conformations and a loss of circadian rhythms. The density $d_{\mathscr{S}}(X)$ of *X* motifs identified in the different wild type and optimized constructs is shown in Table 6. As in the previous examples, the optimized constructs contain significantly more *X* motifs (for instance, density of 10.2 in the N-terminal end of the fully optimized construct compared to 4.1 in the wild type). This example shows how non-optimal codon usage, and the associated reduction in the number of

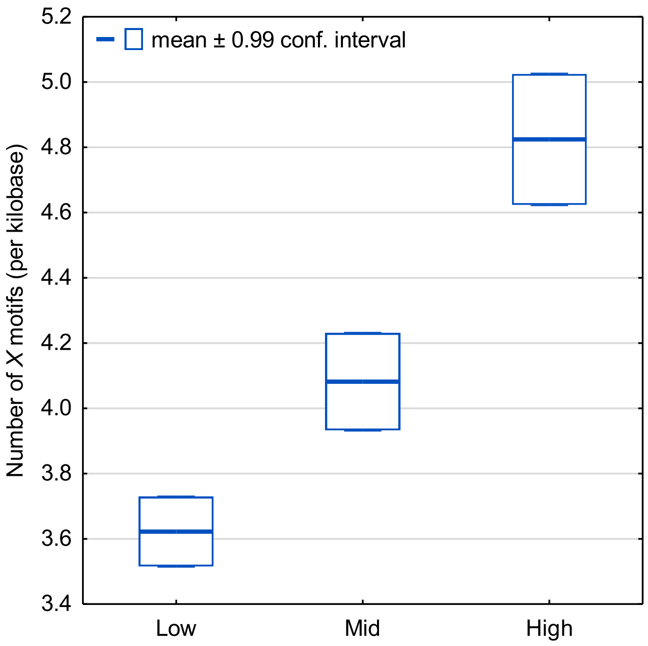


Fig. 7. Density $d_{\mathscr{S}}(X)$ of *X* motifs (Equation (3), i.e. number of *X* motifs per kilobase) for *S. cerevisiae* genes: 1323 genes with low translation rates (estimated translation rate < 0.03), 1378 genes with medium translation rates (estimated translation rate 0.05–0.09) and 1324 genes with high translation rates (estimated translation rate > 1.1). The distributions of the number of *X* motifs identified in the genes are indicated by boxplots representing the mean number with a ±0.99 confidence interval. The statistical significance is confirmed by two one-sided Student's *t*-tests with *p*-value $< 10^{-10}$ between the sequences with medium translation rates and those with low translation rates, and *p*-value $< 10^{-14}$ between the sequences with high translation rates and those with medium translation rates.

X motifs, can be used *in vivo* to regulate protein expression and to achieve optimal protein structure and function.

In this section, we have cited diverse examples of artificial codon optimization experiments and shown that the optimizations are biased to introduce *X* codons. Since it is highly unlikely that the experiments took into account the circular code theory, we conclude that this is an independent confirmation of the theory of the *X* circular code.

In nature, the translation efficiency of a gene may vary at different conditions, cell types and tissues (Charneski and Hurst, 2013; Gardin et al., 2014; Weinberg et al., 2016; Wu et al., 2019). Thus, it has been proposed that the codon optimization should take into account other factors in addition to replacing rare codons by frequent ones, a process termed 'codon harmonization' (Brule and Grayhack, 2017; Villada and Brustolini, 2017; Mignon et al., 2018). Taken together, the examples described above suggest that it may be important for such harmonization strategies to consider the effect of codon replacement on the insertion or deletion of *X* motifs.

3.6. *X* motifs correlate with translation efficiency and mRNA stability

Previously, we showed that the reading frames of genes in *S. cerevisiae* are significantly enriched in *X* motifs (Michel et al., 2017). Since then, ribosomal profiling has enabled a detailed study of translation efficiency for a large set of 5450 genes from this organism (Diament et al., 2018). A central assumption of ribosome profiling is that indirect measurement of the kinetics of translation *via* ribosome footprint occupancy on transcripts is directly reflective of true protein synthesis. The authors thus estimated the average translation rate of each gene, using experimental measurements of ribosome occupancy. Again, we identified the *X* motifs in the complete set of 5450 genes and calculated the density $d_{\mathcal{F}}(X)$ of *X* motifs (Equation (3)) in three subsets of the genes having different estimated translation rates (Fig. 7). We observed that genes with higher translation rates had significantly more *X* motifs than those with lower translation rates. The density of *X* motifs is higher for the sequences with medium translation rates than for those with low translation rates (one-sided Student's *t*-test *p*-value $< 10^{-10}$) and for the sequences with high translation rates than for those with medium translation rates (one-sided Student's *t*-test *p*-value $< 10^{-14}$). This result demonstrates the link between the total time needed for ribosome transition on a mRNA and density of *X* motifs along the length of the sequence.

To investigate whether *X* motifs might play a role in modulating ribosome speed in specific regions in mRNA, we considered single protein studies, where local translation elongation rate has been studied in detail. The first example concerns the study of a gene in *S. cerevisiae*, to investigate the link between translational elongation and mRNA decay (Boël et al., 2016). In this study, various HIS3 protein constructs (length of 699 nucleotides) were designed with increasing codon optimality (measured by the CSC index) from 0% to 100%. We identified *X* motifs in the different constructs as before and compared them to the experimentally measured mRNA half-life. As the authors point out, the mRNA half-life is largely determined by the codon-dependent rate of translational elongation, since mRNAs whose translation elongation rate is slowed by inclusion of non-optimal codons are specifically degraded. The density of *X* motifs ranges from 0 in the 0% optimized construct to more than 7 in the 100% optimized sequence (Fig. 8). The results suggest that the introduction of individual *X* motifs in specific regions can be used to increase the mRNA half-life.

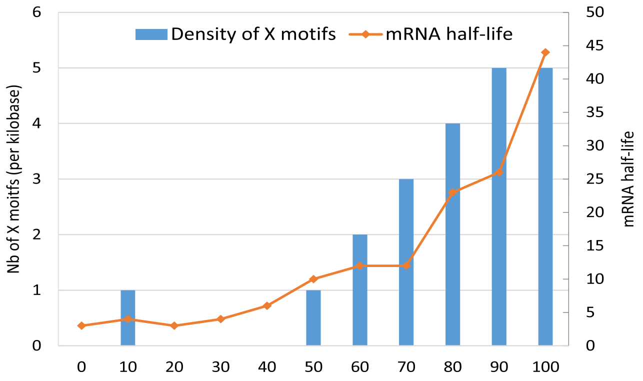


Fig. 8. Histogram of the density $d_{\mathcal{F}}(X)$ of *X* motifs (Equation (3), i.e. number of *X* motifs per kilobase) for different constructs corresponding to the *S. cerevisiae* HIS3 gene with 0–100% optimized codons. The orange plot indicates the mRNA half-life values corresponding to each construct. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

The second example concerns a *Drosophila* cell-free translation system that was used to directly compare the rate of mRNA translation elongation for different luciferase constructs with synonymous substitutions (Błażej et al., 2018). The OPT construct was designed with the most preferred codons in all positions except for the first 10 codons, while the dOPT construct had the least preferred codons in all positions. The N-OPT, M-OPT and C-OPT constructs were created by replacing the N-terminal part (codons 11–223), middle part (codons 224–423) and C-terminal part (codons 424–550) of the dOPT sequence with the corresponding optimized sequence, respectively. For each construct, the authors measured the time when the luminescence signal was first detected after start of translation. The time of first appearance (TFA) should thus reflect the speed of translation process. Higher TFA values were observed for each construct in the order dOPT $<$ C-OPT $<$ M-OPT $<$ N-OPT $<$ OPT, correlating well with an increasing density of *X* motifs (Fig. 9). These results suggest that the introduction of *X* motifs in different regions of the gene significantly increased the rate of translation elongation, probably by speeding up ribosome movement on the mRNA.

We have highlighted the potential effects of *X* motifs on translation elongation speed, protein folding and function. The examples selected include studies in very different organisms, including viruses, fungi and insects with different codon usage bias (codon usage tables for these organisms are provided in appendix Table A1), but in all the examples a strong correlation is observed between 'optimal' codons and *X* codons. Taken together, the results support the idea that the use of *X* motifs is a conserved mechanism from viruses to animals that may participate in the modulation or regulation of the translation elongation rate along the mRNA.

4. Discussion

In this work, we have combined two very distinct research domains: gene translation through the genetic code and the theory of circular codes which allows two processes simultaneously: reading frame retrieval and amino acid coding. Our hypothesis is that at least two codes operate in genes: the standard genetic code, experimentally proved to be functional, and the *X* circular code that has been shown to be statistically enriched in genes. Recent studies have suggested that the *X* circular code is involved in the regulation of ribosomal frameshifting

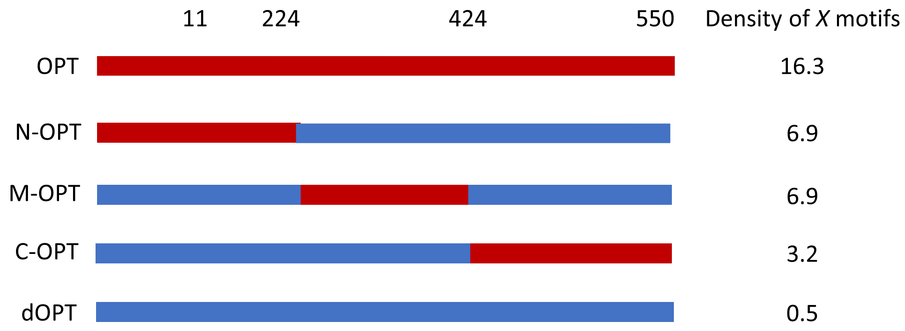


Fig. 9. Density $d_{\mathscr{S}}(X)$ of X motifs (Equation (3), i.e. number of X motifs per kilobase) in the different constructs corresponding to the *Drosophila* luciferase gene. Sequence regions shown in blue are codon optimized, and in red are the wild type sequence. The numbers above the sequences indicate the codon positions of the optimized regions. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

mechanisms (Warthi and Seligmann, 2019; Dila et al., 2020). For the first time here, we shed light on a number of experimental results related to the regulation of gene expression, by using the definition of a very simple parameter analyzing the density of X motifs in genes, i.e. motifs from the circular code X.

We would first like to make some comments about the mathematical structure of these two codes. The standard genetic code consists of 60 codons coding for 19 amino acids, the start codon ATG that codes for the amino acid Met and establishes the reading frame, and three non-coding stop codons {TAA,TAG,TGA}. The genetic code has a weak mathematical structure: a surjective coding map for the 60 codons and an incomplete self-complementary property for the 60 codons (e.g. the complementary codon of TTA coding the amino acid Leu is the stop codon TAA). The set of start and stop codons is not self-complementary. The circular code X consists of 20 codons coding for 12 amino acids and has a strong mathematical structure: circularity for retrieving the reading frame, a surjective coding map, a complete self-complementary property for the 20 codons, a C^3 property, etc. (reviewed in Michel, 2008; Fimmel and Strüngmann, 2018).

We propose that the theory of circular codes can be used to shed light on many of the observed phenomena related to optimal codons/dicodons and the effects of codon optimization on different factors of gene expression, from transcriptional regulation to translation initiation, retrieval of the open reading frame, translation elongation velocities, and protein folding. We showed that optimal codons at the species and gene levels correlate well with the 20 codons that define the X circular code. Importantly, the optimal codons identified in diverse species (Bazzini et al., 2016) that increase translation elongation rates and mRNA stability are significantly enriched in X codons. We then studied a number of published experiments that used recent technologies to perform more detailed investigations of codon usage along the length of a gene, which suggest that codon context and local clusters of optimal or non-optimal codons may represent important regulatory signals for translation bursts and pauses (Yu et al., 2015; Rodnina, 2016). In all these experiments, increased translation efficiency correlates with the number of X motifs present in the gene sequences. These observations raise the question: do X motifs somehow orchestrate elongation rate? Since it is known that translational elongation rate is intimately connected to mRNA stability, it is also tempting to suggest that X motifs are linked to the universal code of codon-mediated mRNA decay proposed by Chen and Coller (2016).

The theory of the X circular code will have practical implications for improving the prediction of gene expression levels based on the gene sequence. Most of the current codon usage measures are dependent on the studied organism and the chosen expression system. In contrast, the presence of X motifs represents a universal signature that is significantly correlated with increased expression. Our previous work has already shown that X motifs can predict functional versus dubious genes in yeast (Michel et al., 2017), in coronaviruses (Michel et al., 2020), and can be used for rational gene design (Dila et al., 2019a).

Translation of mRNA by the ribosome is a universal mechanism, and the most parsimonious explanation for the observed correlation between the presence of X motifs and increased translation elongation rates is that X motifs are somehow recognized by the ribosome. It is known that codon usage has effects on the major steps of translation elongation, including codon-anticodon decoding and peptide bond formation (Rodnina, 2016), as well as translocation which can be slowed down by mRNA secondary structure elements, such as pseudoknots and stem-loops (Wu et al., 2018). Our hypothesis that X motifs in mRNA are recognized by the ribosome is further supported by recent ribosome profiling experiments in *Neurospora crassa*, which suggest that codon optimization increases the rate of ribosome movement on mRNA (Zhou et al., 2016), and by the observation that translation elongation and mRNA stability are coupled through the ribosomal A-site (Hanson et al., 2018). Interestingly, our previous work has identified X motifs in the anticodon region of multiple tRNA genes, as well as in important functional regions of the ribosomal rRNA including the decoding center (Michel, 2012; Dila et al., 2019b).

How could motifs from the X circular code work? If the decoding unit at the ribosome is the anticodon then the comma-free codes would immediately return to the reading phase while the general circular codes would have a delay associated with reading at most four codons (exactly 13 nucleotides). If the decoding unit at the ribosome is the anticodon with adjacent nucleotides then the general circular codes could also immediately return to the reading phase. Does the self-complementary property of the X circular code contribute to coordination between X motifs in mRNA and X motifs in tRNA and/or rRNA?

So far we have mainly discussed the effects of codon choices on the throughput of translation, but changes in the translation elongation process can clearly affect translation fidelity and accuracy, reviewed in Liu et al. (2017). For example, clustering of rare codons could deplete cognate tRNAs, increasing the probability of a near- or non-cognate tRNA occupying the decoding site, and this probability could be reflected in the frequency of miss-incorporation. In this case, it has been shown that the standard genetic code minimizes the impact of the mutations on the translated protein (Błażej et al., 2018). Clustering of identical rare codons also increases the probability of a frameshift during translation. Ribosome stalling at Lys codons triggers ribosome sliding on successive AAA codons. When ribosomes resume translation, they may shift in an incorrect reading frame. The ribosomes translating in the −1 or +1 frame usually quickly encounter out-of-frame stop codons that result in termination. Again, it has been suggested that the genetic code might be in some way optimized for frameshift mutations (Geyer and Madany Mamlouk, 2018; Bartonek et al., 2020), although we have shown recently that in the case of +1 frameshifts, the *X* circular code is in fact more optimized than the standard genetic code (Dila et al., 2020). Given the inherent error correcting properties of circular codes, it is thus possible that the *X* circular code may play a role in the synchronization of the correct reading frame.

In the future, we hope to show that the simple parameter defined in this work to estimate the coverage of *X* motifs in genes is a useful factor that should be taken into account in codon optimization strategies or other experimental approaches involving gene expression. We also plan to investigate more complex parameters linked to *X* motifs, such as localized density patterns within specific regions of the genes.

Declaration of competing interest

The authors report no conflict of interest.

Acknowledgements

This work was supported by Institute funds from the French Centre National de la Recherche Scientifique and the University of Strasbourg. The authors would like to thank the BigEst and BICS Bioinformatics Platforms for their assistance. This work was supported by French Infrastructure Institut Français de Bioinformatique (IFB) ANR-11-INBS-0013, and the ANR Grants Elixir-Excelerate: GA-676559 and RAinRARE: ANR-18-RAR3-0006-02.

Appendix A

Table A.1
Codon usage tables for the species used in the different studies described in the main text. **A.** *Saccharomyces cerevisiae*. **B.** *Drosophila melanogaster*. **C.** *Neorospora crassa*. **D.** *Escherichia coli*. Data are from the HIVE-CUTS database at https://hive.biochemistry.gwu.edu/cuts/. *X* codons are highlighted in blue.

A. *Saccharomyces cerevisiae*

TTT	26.27	TCT	23.34	TAT	19.08	TGT	7.86
TTC	17.91	TCC	14.06	TAC	14.60	TGC	4.79
TTA	26.32	TCA	19.05	TAA	0.97	TGA	0.61
TTG	26.47	TCG	8.72	TAG	0.47	TGG	10.36
CTT	12.33	CCT	13.57	CAT	13.90	CGT	6.28
CTC	5.55	CCC	6.92	CAC	7.76	CGC	2.65
CTA	13.51	CCA	17.79	CAA	27.06	CGA	3.12
CTG	10.66	CCG	5.43	CAG	12.42	CGG	1.84
ATT	30.10	ACT	20.22	AAT	36.56	AGT	14.59
ATC	16.98	ACC	12.47	AAC	24.78	AGC	9.97
ATA	18.32	ACA	18.17	AAA	42.81	AGA	21.03
ATG	20.70	ACG	8.16	AAG	30.47	AGG	9.46
GTT	21.45	GCT	20.26	GAT	38.01	GGT	22.55
GTC	11.22	GCC	12.13	GAC	20.36	GGC	9.79
GTA	12.07	GCA	16.27	GAA	45.72	GGA	11.19
GTG	10.73	GCG	6.18	GAG	19.53	GGG	6.07

C. *Neorospora crassa*

TTT	12.25	TCT	12.25	TAT	8.61	TGT	3.54
TTC	21.43	TCC	20.02	TAC	16.95	TGC	7.50
TTA	2.83	TCA	9.57	TAA	0.56	TGA	0.78
TTG	15.40	TCG	14.94	TAG	0.57	TGG	13.31
CTT	14.33	CCT	15.59	CAT	9.82	CGT	8.58
CTC	26.24	CCC	22.34	CAC	14.75	CGC	17.35
CTA	6.04	CCA	12.81	CAA	17.29	CGA	7.18
CTG	18.29	CCG	15.09	CAG	25.47	CGG	8.52
ATT	13.90	ACT	11.42	AAT	10.68	AGT	8.90
ATC	26.01	ACC	24.74	AAC	26.35	AGC	17.79
ATA	4.23	ACA	11.13	AAA	11.95	AGA	8.01
ATG	21.51	ACG	13.90	AAG	38.73	AGG	11.82
GTT	14.03	GCT	20.93	GAT	24.27	GGT	17.51
GTC	24.14	GCC	35.26	GAC	32.18	GGC	28.48
GTA	5.54	GCA	13.00	GAA	23.13	GGA	13.81
GTG	15.78	GCG	17.46	GAG	41.77	GGG	11.44

B. *Drosophila melanogaster*

TTT	12.55	TCT	7.33	TAT	10.70	TGT	5.83
TTC	20.48	TCC	19.65	TAC	17.40	TGC	13.14
TTA	4.56	TCA	8.51	TAA	0.63	TGA	0.44
TTG	15.78	TCG	17.48	TAG	0.54	TGG	9.24
CTT	8.67	CCT	7.46	CAT	10.81	CGT	8.95
CTC	13.30	CCC	18.40	CAC	15.76	CGC	17.76
CTA	8.07	CCA	14.78	CAA	16.74	CGA	8.54
CTG	36.78	CCG	16.62	CAG	37.45	CGG	8.04
ATT	16.39	ACT	10.07	AAT	21.64	AGT	12.15
ATC	22.00	ACC	21.54	AAC	25.90	AGC	21.05
ATA	9.44	ACA	11.83	AAA	16.43	AGA	5.12
ATG	22.48	ACG	14.89	AAG	38.31	AGG	6.09
GTT	11.32	GCT	14.47	GAT	27.85	GGT	13.62
GTC	13.52	GCC	32.95	GAC	23.87	GGC	26.47
GTA	6.53	GCA	13.23	GAA	22.05	GGA	18.09
GTG	27.02	GCG	14.13	GAG	42.53	GGG	4.61

D. *Escherichia coli*

TTT	22.53	TCT	8.47	TAT	16.16	TGT	5.20
TTC	16.56	TCC	8.66	TAC	12.17	TGC	6.47
TTA	13.96	TCA	7.18	TAA	2.10	TGA	0.97
TTG	13.79	TCG	8.91	TAG	0.23	TGG	15.19
CTT	11.06	CCT	7.04	CAT	12.96	CGT	20.92
CTC	11.07	CCC	5.47	CAC	9.67	CGC	21.99
CTA	3.94	CCA	8.44	CAA	15.42	CGA	3.55
CTG	52.78	CCG	23.25	CAG	28.83	CGG	5.29
ATT	30.58	ACT	8.92	AAT	17.71	AGT	8.73
ATC	25.18	ACC	23.36	AAC	21.54	AGC	16.02
ATA	4.35	ACA	7.05	AAA	33.69	AGA	2.07
ATG	27.80	ACG	14.38	AAG	10.27	AGG	1.17
GTT	18.28	GCT	15.20	GAT	32.07	GGT	24.72
GTC	15.29	GCC	25.50	GAC	19.04	GGC	29.57
GTA	10.93	GCA	20.16	GAA	39.46	GGA	7.88
GTG	26.29	GCG	33.75	GAG	17.77	GGG	11.04

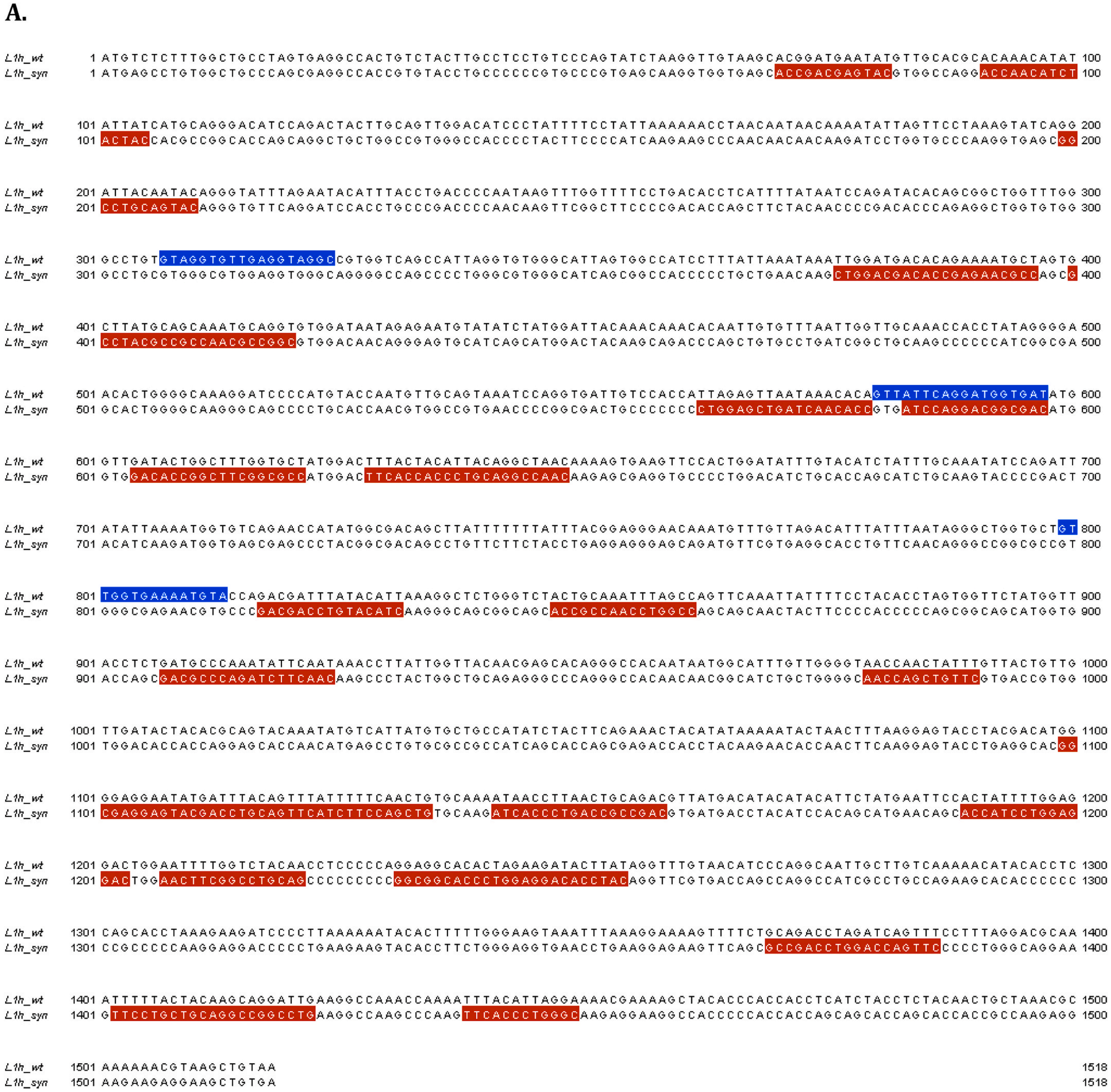
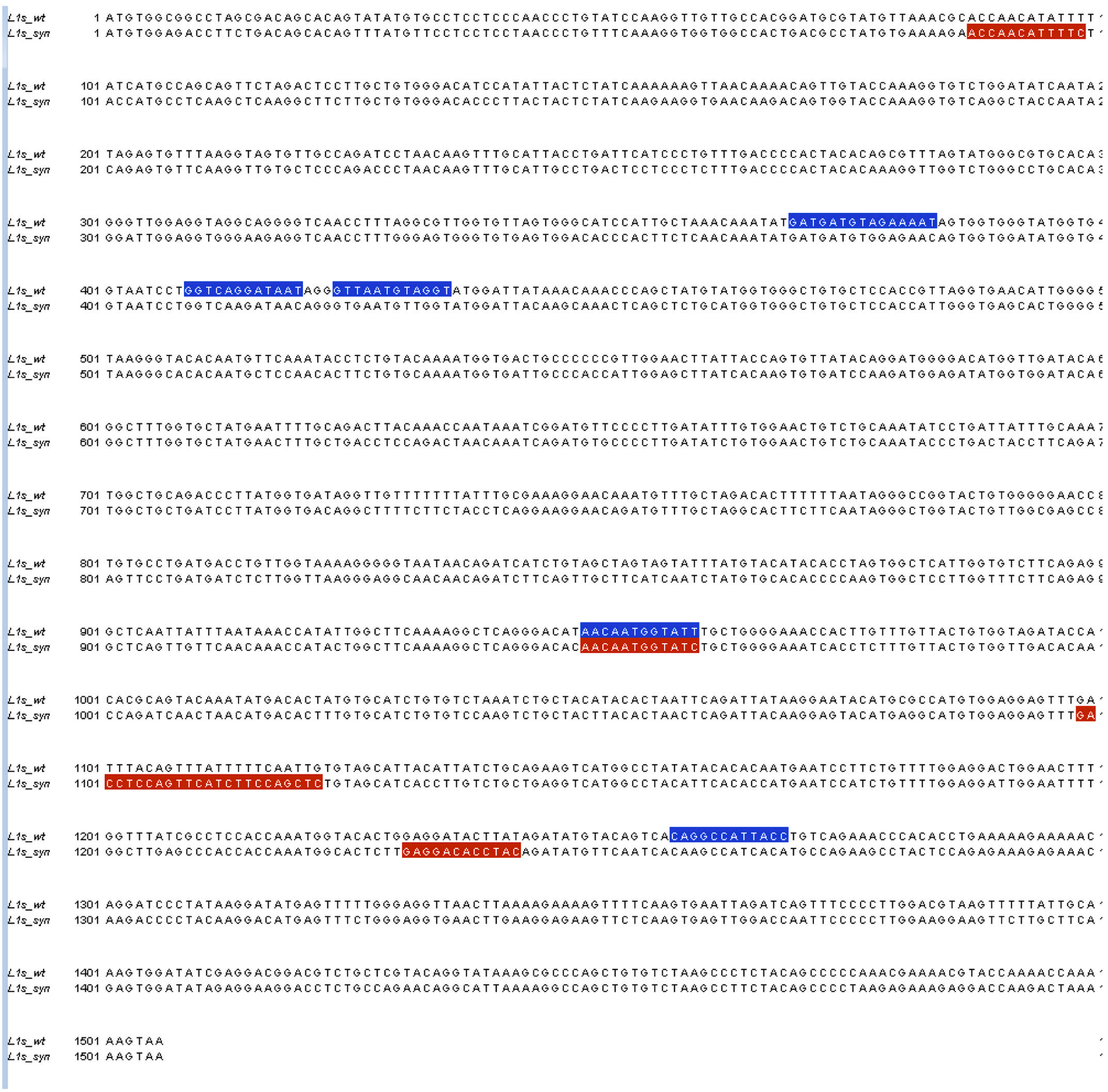


Fig. A.1. *X* motifs in the wild type and codon-optimized sequences. **A.** L1h gene from human papillomavirus (Leder et al., 2001). **B.** L1s gene from human papillomavirus (Warzecha et al., 2003). The *X* motifs in the wild type sequence are shown in blue, and in the codon optimized sequences in red.

B.



```
L1s_wt     1 ATGTGGCGGCCTAGCGACAGCACAGTATATGTGCCTCCTCCCAACCCTGTATCCAAGGTTGTTGCCACGGATGCGTATGTTAAACGCACCAACATATTTT
L1s_syn    1 ATGTGGAGACCTTCTGACAGCACAGTTTATGTTCCTCCTCCTAACCCTGTTTCAAAGGTGGTGGCCACTGACGCCTATGTGAAAAGAACCAACATTTTCT

L1s_wt   101 ATCATGCCAGCAGTTCTAGACTCCTTGCTGTGGGACATCCATATTACTCTATCAAAAAAGTTAACAAAACAGTTGTACCAAAGGTGTCTGGATATCAATA
L1s_syn  101 ACCATGCCTCAAGCTCAAGGCTTCTTGCTGTGGGACACCCTTACTACTCTATCAAGAAGGTGAACAAGACAGTGGTACCAAAGGTGTCAGGCTACCAATA

L1s_wt   201 TAGAGTGTTTAAGGTAGTGTTGCCAGATCCTAACAAGTTTGCATTACCTGATTCATCCCTGTTTGACCCCACTACACAGCGTTTAGTATGGGCGTGCACA
L1s_syn  201 CAGAGTGTTCAAGGTTGTGCTCCCAGACCCTAACAAGTTTGCATTGCCTGACTCCTCCCTCTTTGACCCCACTACACAAAGGTTGGTCTGGGCCTGCACA

L1s_wt   301 GGGTTGGAGGTAGGCAGGGGTCAACCTTTAGGCGTTGGTGTTAGTGGGCATCCATTGCTAAACAAATATGATGATGTAGAAAATAGTGGTGGGTATGGTG
L1s_syn  301 GGATTGGAGGTGGGAAGAGGTCAACCTTTGGGAGTGGGTGTGAGTGGACACCCACTTCTCAACAAATATGATGATGTGGAGAACAGTGGTGGATATGGTG

L1s_wt   401 GTAATCCTGGTCAGGATAATAGGGTTAATGTAGGTATGGATTATAAACAAACCCAGCTATGTATGGTGGGCTGTGCTCCACCGTTAGGTGAACATTGGGG
L1s_syn  401 GTAATCCTGGTCAAGATAACAGGGTGAATGTTGGTATGGATTACAAGCAAACTCAGCTCTGCATGGTGGGCTGTGCTCCACCATTGGGTGAGCACTGGGG

L1s_wt   501 TAAGGGTACACAATGTTCAAATACCTCTGTACAAAATGGTGACTGCCCCCCGTTGGAACTTATTACCAGTGTTATACAGGATGGGGACATGGTTGATACA
L1s_syn  501 TAAGGGCACACAATGCTCCAACACTTCTGTGCAAAATGGTGATTGCCCACCATTGGAGCTTATCACAAGTGTGATCCAAGATGGAGATATGGTGGATACA

L1s_wt   601 GGCTTTGGTGCTATGAATTTTGCAGACTTACAAACCAATAAATCGGATGTTCCCCTTGATATTTGTGGAACTGTCTGCAAATATCCTGATTATTTGCAAA
L1s_syn  601 GGCTTTGGTGCTATGAACTTTGCTGACCTCCAGACTAACAAATCAGATGTGCCCCTTGATATCTGTGGAACTGTCTGCAAATACCCTGACTACCTTCAGA

L1s_wt   701 TGGCTGCAGACCCTTATGGTGATAGGTTGTTTTTTTATTTGCGAAAGGAACAAATGTTTGCTAGACACTTTTTTAATAGGGCCGGTACTGTGGGGGAACC
L1s_syn  701 TGGCTGCTGATCCTTATGGTGACAGGCTTTTCTTCTACCTCAGGAAGGAACAGATGTTTGCTAGGCACTTCTTCAATAGGGCTGGTACTGTTGGCGAGCC

L1s_wt   801 TGTGCCTGATGACCTGTTGGTAAAAGGGGGTAATAACAGATCATCTGTAGCTAGTAGTATTTATGTACATACACCTAGTGGCTCATTGGTGTCTTCAGAG
L1s_syn  801 AGTTCCTGATGATCTCTTGGTTAAGGGAGGCAACAACAGATCTTCAGTTGCTTCATCAATCTATGTGCACACCCCAAGTGGCTCCTTGGTTTCTTCAGAG

L1s_wt   901 GCTCAATTATTTAATAAACCATATTGGCTTCAAAAGGCTCAGGGACATAACAATGGTATTTGCTGGGGAAACCACTTGTTTGTTACTGTGGTAGATACCA
L1s_syn  901 GCTCAGTTGTTCAACAAACCATACTGGCTTCAAAAGGCTCAGGGACACAACAATGGTATCTGCTGGGGAAATCACCTCTTTGTTACTGTGGTTGACACAA

L1s_wt  1001 CACGCAGTACAAATATGACACTATGTGCATCTGTGTCTAAATCTGCTACATACACTAATTCAGATTATAAGGAATACATGCGCCATGTGGAGGAGTTTGA
L1s_syn 1001 CCAGATCAACTAACATGACACTTTGTGCATCTGTGTCCAAGTCTGCTACTTACACTAACTCAGATTACAAGGAGTACATGAGGCATGTGGAGGAGTTTGA

L1s_wt  1101 TTTACAGTTTATTTTTCAATTGTGTAGCATTACATTATCTGCAGAAGTCATGGCCTATATACACACAATGAATCCTTCTGTTTTGGAGGACTGGAACTTT
L1s_syn 1101 CCTCCAGTTCATCTTCCAGCTCTGTAGCATCACCTTGTCTGCTGAGGTCATGGCCTACATTCACACCATGAATCCATCTGTTTTGGAGGATTGGAATTTT

L1s_wt  1201 GGTTTATCGCCTCCACCAAATGGTACACTGGAGGATACTTATAGATATGTACAGTCACAGGCCATTACCTGTCAGAAACCCACACCTGAAAAAGAAAAAC
L1s_syn 1201 GGCTTGAGCCCACCACCAAATGGCACTCTTGAGGACACCTACAGATATGTTCAATCACAAGCCATCACATGCCAGAAGCCTACTCCAGAGAAAGAGAAAC

L1s_wt  1301 AGGATCCCTATAAGGATATGAGTTTTTGGGAGGTTAACTTAAAAGAAAAGTTTTCAAGTGAATTAGATCAGTTTCCCCTTGGACGTAAGTTTTTATTGCA
L1s_syn 1301 AAGACCCCTACAAGGACATGAGTTTCTGGGAGGTGAACTTGAAGGAGAAGTTCTCAAGTGAGTTGGACCAATTCCCCCTTGGAAGGAAGTTCTTGCTTCA

L1s_wt  1401 AAGTGGATATCGAGGACGGACGTCTGCTCGTACAGGTATAAAGCGCCCAGCTGTGTCTAAGCCCTCTACAGCCCCCAAACGAAAACGTACCAAAACCAAA
L1s_syn 1401 GAGTGGATATAGAGGAAGGACCTCTGCCAGAACAGGCATTAAAAGGCCAGCTGTGTCTAAGCCTTCTACAGCCCCTAAGAGAAAGAGGACCAAGACTAAA

L1s_wt  1501 AAGTAA
L1s_syn 1501 AAGTAA
```

Fig. A.1. (*continued*).

References

Ahmed, A., Frey, G., Michel, C.J., 2010. Essential molecular functions associated with the circular code evolution. J. Theor. Biol. 264, 613–622.

Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein-coding genes. J. Theor. Biol. 182, 45–58.

Babbitt, G.A., Coppola, E.E., Mortensen, J.S., Ekeren, P.X., Viola, C., Goldblatt, D., Hudson, A.O., 2018. Triplet-based codon organization optimizes the impact of synonymous mutation on nucleic acid molecular dynamics. J. Mol. Evol. 86, 91–102.

Bali, V., Bebok, Z., 2015. Decoding mechanisms by which silent codon changes influence protein biogenesis and function. Int. J. Biochem. Cell Biol. 64, 58–74.

Baralle, M., Baralle, F.E., 2018. The splicing code. Biosystems 164, 39–48.

Bartonek, L., Braun, D., Zagrovic, B., 2020. Frameshifting preserves key physicochemical properties of proteins. Proceedings of the National Academy of Sciences USA 117, 5907–5912.

Bazzini, A.A., Del Viso, F., Moreno-Mateos, M.A., Johnstone, T.G., Vejnar, C.E., Qin, Y., Yao, J., Khokha, M.K., Giraldez, A.J., 2016. Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. EMBO J. 35, 2087–2103.

Bergman, S., Tuller, T., 2020. Widespread non-modular overlapping codes in the coding regions. Phys. Biol. 1088, 1478–3975/ab7083.

Błażej, P., Wnętrzak, M., Mackiewicz, D., Mackiewicz, P., 2018. Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm. PloS One 13, e0201715.

Boël, G., Letso, R., Neely, H., Price, W.N., Wong, K.H., Su, M., Luff, J., Valecha, M., Everett, J.K., Acton, T.B., Xiao, R., Montelione, G.T., Aalberts, D.P., Hunt, J.F., 2016. Codon influence on protein expression in E. coli correlates with mRNA levels. Nature 529, 358–363.

Brar, G.A., 2016. Beyond the triplet code: context cues transform translation. Cell 167, 1681–1692.

Brule, C.E., Grayhack, E.J., 2017. Synonymous codons: choose wisely for expression. Trends Genet. 33, 283–297.

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M., Komar, A.A., 2016. Synonymous codons direct cotranslational folding toward different protein conformations. Mol. Cell 61, 341–351.

Bull, J.J., Molineux, I.J., Wilke, C.O., 2012. Slow fitness recovery in a codon-modified viral genome. Mol. Biol. Evol. 29, 2997–3004.

Burow, D.A., Martin, S., Quail, J.F., Alhusaini, N., Coller, J., Cleary, M.D., 2018. Attenuated codon optimality contributes to neural-specific mRNA decay in drosophila. Cell Rep. 24, 1704–1712.

Cakiroglu, S.A., Zaugg, J.B., Luscombe, N.M., 2016. Backmasking in the yeast genome: encoding overlapping information for protein-coding and RNA degradation. Nucleic Acids Res. 44, 8065–8072.

Charneski, C.A., Hurst, L.D., 2013. Positively charged residues are the major determinants of ribosomal velocity. PLoS Biol. 11, e1001508.

Chen, Y.H., Coller, J., 2016. A universal code for mRNA stability? Trends Genet. 32, 687–688.

Chevance, F.F.V., Hughes, K.T., 2017. Case for the genetic code as a triplet of triplets. Proceedings of the National Academy of Sciences USA 114, 4745–4750.

Clarke, T.F., Clark, P.L., 2008. Rare codons cluster. PloS One 3, e3412.

Crick, F.H., Barnett, L., Brenner, S., Watts-Tobin, R.J., 1961. General nature of the genetic code for proteins. Nature 192, 1227–1232.

Crick, F.H., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proceedings of the National Academy of Sciences USA 43, 416–421.

Demongeot, J., Seligmann, H., 2019. Spontaneous evolution of circular codes in theoretical minimal RNA rings. Gene 705, 95–102.

Demongeot, J., Moreira, A., Seligmann, H., 2020. Negative CG dinucleotide bias: an explanation based on feedback loops between Arginine codon assignments and theoretical minimal RNA rings. Bioessays, e2000071.

Diambra, L.A., 2017. Differential bicodon usage in lowly and highly abundant proteins. PeerJ 5, e3081.

Diament, A., Feldman, A., Schochet, E., Kupiec, M., Arava, Y., Tuller, T., 2018. The extent of ribosome queuing in budding yeast. PLoS Comput. Biol. 14, e1005951.

Dila, G., Michel, C.J., Poch, O., Ripp, R., Thompson, J.D., 2019a. Evolutionary conservation and functional implications of circular code motifs in eukaryotic genomes. Biosystems 175, 57–74.

Dila, G., Michel, C.J., Thompson, J.D., 2020. Optimality of circular codes versus the genetic code after frameshift errors. Biosystems 195, 104134, 1–11.

Dila, G., Ripp, R., Mayer, C., Poch, O., Michel, C.J., Thompson, J.D., 2019b. Circular code motifs in the ribosome: a missing link in the evolution of translation? RNA 25, 1714–1730.

El Soufi, K., Michel, C.J., 2016. Circular code motifs in genomes of eukaryotes. J. Theor. Biol. 408, 198–212.

Eslami-Mossallam, B., Schram, R.D., Tompitak, M., van Noort, J., Schiessel, H., 2016. Multiplexing genetic and nucleosome positioning codes: a computational approach. PloS One 11, e0156905.

Faure, G., Ogurtsov, A.Y., Shabalina, S.A., Koonin, E.V., 2017. Adaptation of mRNA structure to control protein folding. RNA Biol. 14, 1649–1654.

Fimmel, E., Strüngmann, L., 2018. Mathematical fundamentals for the noise immunity of the genetic code. Biosystems 164, 86–198.

Fimmel, E., Michel, C.J., Pirot, F., Sereni, J.-S., Strüngmann, L., 2019. Mixed circular codes. Math. Biosci. 317, 108231.

Fimmel, E., Michel, C.J., Pirot, F., Sereni, J.-S., Starman, M., Strüngmann, L., 2020. The relation between *k*-circularity and circularity of codes. Bull. Math. Biol. 82 (105), 1–34.

Frey, G., Michel, C.J., 2003. Circular codes in archaeal genomes. J. Theor. Biol. 223, 413–431.

Frey, G., Michel, C.J., 2006. Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. Comput. Biol. Chem. 30, 87–101.

Gamble, C.E., Brule, C.E., Dean, K.M., Fields, S., Grayhack, E.J., 2016. Adjacent codons act in concert to modulate translation efficiency in yeast. Cell 166, 679–690.

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., Futcher, B., 2014. Measurement of average decoding rates of the 61 sense codons in vivo. Elife 3.

Geyer, R., Madany Mamlouk, A., 2018. On the efficiency of the genetic code after frameshift mutations. PeerJ 6, e4825.

Gingold, H., Pilpel, Y., 2011. Determinants of translation efficiency and accuracy. Mol. Syst. Biol. 7, 481.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., Mercier, R., 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9, r43–74.

Grosjean, H., Westhof, E., 2016. An integrated, structure- and energy-based view of the genetic code. Nucleic Acids Res. 44, 8020–8040.

Guo, F.B., Ye, Y.N., Zhao, H.L., Lin, D., Wei, W., 2012. Universal pattern and diverse strengths of successive synonymous codon bias in three domains of life, particularly among prokaryotic genomes. DNA Res. 19, 477–485.

Hanson, G., Alhusaini, N., Morris, N., Sweet, T., Coller, J., 2018. Translation elongation and mRNA stability are coupled through the ribosomal A-site. RNA 24, 1377–1389.

Hanson, G., Coller, J., 2018. Codon optimality, bias and usage in translation and mRNA decay. Nat. Rev. Mol. Cell Biol. 19, 20–30.

Jack, B.R., Boutz, D.R., Paff, M.L., Smith, B.L., Bull, J.J., Wilke, C.O., 2017. Reduced protein expression in a virus attenuated by codon deoptimization. G3 (Bethesda) 7, 2957–2968.

Koonin, E.V., 2000. How many genes can make a cell: the minimal-gene-set concept. Annu. Rev. Genom. Hum. Genet. 1, 99–116.

Leder, C., Kleinschmidt, J.A., Wiethe, C., Müller, M., 2001. Enhancement of capsid gene expression: preparing the human papillomavirus type 16 major structural gene L1 for DNA vaccination purposes. J. Virol. 75, 9201–9209.

Liu, Y., Sharp, J.S., Do, D.H., Kahn, R.A., Schwalbe, H., Buhr, F., Prestegard, J.H., 2017. Mistakes in translation: reflections on mechanism. PloS One 12, e0180566.

Maraia, R.J., Iben, J.R., 2014. Different types of secondary information in the genetic code. RNA 20, 977–984.

Michel, C.J., 2008. A 2006 review of circular codes in genes. Comput. Math. Appl. 55, 984–988.

Michel, C.J., 2012. Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. Comput. Biol. Chem. 37, 24–37.

Michel, C.J., 2013. Circular code motifs in transfer RNAs. Comput. Biol. Chem. 45, 17–29.

Michel, C.J., 2017. The maximal C^3 self-complementary trinucleotide circular code *X* in genes of bacteria, archaea, eukaryotes, plasmids and viruses. Life 7, 20.

Michel, C.J., 2020. The maximality of circular codes in genes statistically verified. Biosystems 197, 104201, 1–7.

Michel, C.J., Seligmann, H., 2014. Bijective transformation circular codes and nucleotide exchanging RNA transcription. Biosystems 118, 39–50.

Michel, C.J., Thompson, J.D., 2020. Identification of a circular code periodicity in the bacterial ribosome: origin of codon periodicity in genes? RNA Biol. 17, 571–583.

Michel, C.J., Mayer, C., Poch, O., Thompson, J.D., 2020. Characterization of accessory genes in coronavirus genomes. Virol. J. 17 (131), 1–13.

Michel, C.J., Ngoune, V.N., Poch, O., Ripp, R., Thompson, J.D., 2017. Enrichment of circular code motifs in the genes of the yeast Saccharomyces cerevisiae. Life 7 (52), 1–20.

Mignon, C., Mariano, N., Stadthagen, G., Lugari, A., Lagoutte, P., Donnat, S., Chenavas, S., Perot, C., Sodoyer, R., Werle, B., 2018. Codon harmonization - going beyond the speed limit for protein expression. FEBS (Fed. Eur. Biochem. Soc.) Lett. 592, 1554–1564.

Nevers, Y., Kress, A., Defosset, A., Ripp, R., Linard, B., Thompson, J.D., Poch, O., Lecompte, O., 2019. OrthoInspector 3.0: open portal for comparative genomics. Nucleic Acids Res. 47, D411–D418.

Nirenberg, M.W., Matthaei, J.H., 1961. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proceedings of the National Academy of Sciences USA 47, 1588–1602.

Palidwor, G.A., Perkins, T.J., Xia, X., 2010. A general model of codon bias due to GC mutational bias. PloS One 5, e13431.

Prakash, K., Fournier, D., 2018. Evidence for the implication of the histone code in building the genome structure. Biosystems 164, 49–59.

Presnyak, V., Alhusaini, N., Chen, Y.H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R., Coller, J., 2015. Codon optimality is a major determinant of mRNA stability. Cell 160, 1111–1124.

Qian, W., Yang, J.R., Pearson, N.M., Maclean, C., Zhang, J., 2012. Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet. 8, e1002603.

Rodnina, M.V., 2016. The ribosome in action: tuning of translational efficiency and protein folding. Protein Sci. 25, 1390–1406.

Seligmann, H., Pollock, D.D., 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. DNA Cell Biol. 23, 701–705.

Seligmann, H., Warthi, G., 2017. Genetic code optimization for cotranslational protein folding: codon directional asymmetry correlates with antiparallel betasheets, tRNA synthetase classes. Comput. Struct. Biotechnol. J. 15, 412–424.

Seligmann, H., 2019. Localized context-dependent effects of the "ambush" hypothesis: more off-frame stop codons downstream of shifty codons. DNA Cell Biol. 38, 786–795.

Sharma, A.K., Sormanni, P., Ahmed, N., Ciryam, P., Friedrich, U.A., Kramer, G., O'Brien, E.P., 2019. A chemical kinetic basis for measuring translation initiation and elongation rates from ribosome profiling data. PLoS Comput. Biol. 15, e1007070.

Simmonds, P., Xia, W., Baillie, J.K., McKinnon, K., 2013. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla –selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genom. 4, 610.

Trifonov, E.N., 2000. Consensus temporal order of amino acids and evolution of the triplet code. Gene 261, 139–151.

Villada, J.C., Brustolini, O.J.B., Batista da Silveira, W., 2017. Integrated analysis of individual codon contribution to protein biosynthesis reveals a new approach to improving the basis of rational gene design. DNA Res. 24, 419–434.

Warthi, G., Seligmann, H., 2019. Transcripts with systematic nucleotide deletion of 1-12 nucleotide in human mitochondrion suggest potential non-canonical transcription. PloS One 14, e0217356.

Warzecha, H., Mason, H.S., Lane, C., Tryggvesson, A., Rybicki, E., Williamson, A.L., Clements, J.D., Rose, R.C., 2003. Oral immunogenicity of human papillomavirus-like particles expressed in potato. J. Virol. 77, 8702–8711.

Weatheritt, R.J., Babu, M.M., 2013. Evolution. The hidden codes that shape protein evolution. Science 342, 1325–1326.

Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B., Bartel, D.P., 2016. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. Cell Rep. 14, 1787–1799.

Wu, B., Zhang, H., Sun, R., Peng, S., Cooperman, B.S., Goldman, Y.E., Chen, C., 2018. Translocation kinetics and structural dynamics of ribosomes are modulated by the conformational plasticity of downstream pseudoknots. Nucleic Acids Res. 46, 9736–9748.

Wu, C.C., Zinshteyn, B., Wehner, K.A., Green, R., 2019. High-resolution ribosome profiling defines discrete ribosome elongation states and translational regulation during cellular stress. Mol. Cell 73, 959–970 e5.

Wu, G., Zheng, Y., Qureshi, I., Zin, H.T., Beck, T., Bulka, B., Freeland, S.J., 2007. SGDB: a database of synthetic genes re-designed for optimizing protein over-expression. Nucleic Acids Res. 35, D76–D79.

Yu, C.H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M.S., Liu, Y., 2015. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Mol. Cell 59, 744–754.

Zhou, M., Guo, J., Cha, J., Chae, M., Chen, S., Barral, J.M., Sachs, M.S., Liu, Y., 2013. Non-optimal codon usage affects expression, structure and function of FRQ clock protein. Nature 495, 111–115.

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C.H., Fu, J., Chen, S., Liu, Y., 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proceedings of the National Academy of Sciences USA 113, E6117–E6125.