Evolutionary conservation and functional implications of circular code motifs in eukaryotic genomes

Gopal Dila, Christian Michel, Olivier Poch, Raymond Ripp, Julie D Thompson

Evolutionary conservation and functional implications of circular code motifs in eukaryotic genomes

Gopal Dila, Christian J. Michel, Olivier Poch, Raymond Ripp, Julie D. Thompson

CSTB, ICube
CNRS, University of Strasbourg
300 Boulevard Sébastien Brant
67400 Illkirch, France

Email: d.gopal@outlook.com, c.michel@unistra.fr, olivier.poch@unistra.fr, raymond.ripp@unistra.fr, thompson@unistra.fr

Keywords: circular code motifs; genome evolution; genetic code; gene expression.

Abstract

A set X of 20 trinucleotides has been found to have the highest average occurrence in the reading frame, compared to the two shifted frames, of genes of bacteria, archaea, eukaryotes, plasmids and viruses (Michel, 2017, 2015; Arquès and Michel, 1996). This set X has an interesting mathematical property, since X is a maximal C^3 self-complementary trinucleotide circular code (Arquès and Michel, 1996). Furthermore, any motif obtained from this circular code X has the capacity to retrieve, maintain and synchronize the reading frame in genes. In a recent study of the X motifs in the complete genome of the yeast, *Saccharomyces cerevisiae*, it was shown that they are significantly enriched in the reading frame of the genes (protein-coding regions) of the genome (Michel *et al.*, 2017). It was suggested that these X motifs may be evolutionary relics of a primitive code originally used for gene translation. The aim of this paper is to address two questions: are X motifs conserved during evolution? and do they continue to play a functional role in the processes of genome decoding and protein production? In a large scale analysis involving complete genomes from four mammals and nine different yeast species, we highlight specific evolutionary pressures on the X motifs in the genes of all the genomes, and identify important new properties of X motif conservation at the level of the encoded amino acids. We then compare the occurrence of X motifs with existing experimental data concerning protein expression and protein production, and report a significant correlation between the number of X motifs in a gene and increased protein abundance. In a general way, this work suggests that motifs from circular codes, i.e. motifs having the property of reading frame retrieval, may represent functional elements located within the coding regions of extant genomes.

1. Introduction

The same set X of trinucleotides (also known as codons) was identified in average in genes (reading frame) of bacteria, archaea, eukaryotes, plasmids and viruses (Michel 2017, 2015; Arquès and Michel, 1996). It contains the 20 following trinucleotides

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

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

$$\mathcal{X} = \{Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Phe, Thr, Tyr, Val\} = \{A, N, D, Q, E, G, I, L, F, T, Y, V\}. \quad (2)$$

This set X has several strong mathematical properties. In particular, it is self-complementary, i.e. 10 trinucleotides of X are complementary to the other 10 trinucleotides of X, e.g. $AAC \in X$ is complementary to $GTT \in X$, and it is a circular code. A circular code is defined as a set of words such that any motif obtained from this set, allows to retrieve, maintain and synchronize the original (construction) frame. Thus, the circular code X may represent a self-correcting property of the genetic code. Indeed, it has been proposed recently that the circular code X may participate in the regulation of gene transcription (El Houmami and Seligmann, 2017). Other correction properties of the genetic code

have also been proposed with roles in gene translation, including compensation of tRNA misloading (Seligmann, 2011), prevention of protein misfolding (Seligmann and Warthi, 2017), or termination of translation after ribosomal frameshifting (Seligmann and Pollock, 2004).

Motifs from the circular code X (denoted (1) above) having this frame retrieval property are called X motifs. Since 1996, the theory of circular codes in genes has mainly been developed by analysing the classes, the numbers and the mathematical properties of circular codes using probability-statistics, combinatorics and graph theory (reviews in Michel, 2008, and Fimmel and Strüngmann, 2018). More recently, the circular code theory was applied to the complete genome sequence of a living organism, namely the eukaryote *Saccharomyces cerevisiae* (Michel *et al.*, 2017). It was shown that X motifs from the circular code X (1) were significantly enriched in the genes (protein-coding regions) of the genome. The authors hypothesized that the X motifs may be evolutionary relics of a primitive code originally used for translation.

In this article, we describe a large-scale study of the X motifs in two independent sets of complete genomes. The first set is composed of four mammal genomes, representing highly evolved species and closely related genomes. The second set is built from nine yeast genomes, representing the simplest eukaryotes with more divergent genome sequences. Each set includes a well-studied and annotated 'reference' genome: the human genome for the first set and the *Saccharomyces cerevisiae* genome for the second set. We first highlight specific evolutionary pressures on the X motifs in the genes of both sets of genomes, and identify important new properties of X motif conservation at the level of the encoded amino acids. Thus, the 20 trinucleotides of the circular code X (1) are grouped according to the amino acids they encode, leading to a new hypothesis for the evolution of the genetic code where each amino acid was coded by the most constrained circular codes, namely strong comma-free and comma-free codes.

Then, we investigate the potential functional role of X motifs in the regulation of gene expression. To achieve this, we compare the occurrence of X motifs with existing experimental data concerning protein expression and production, and report a significant correlation between the number of X motifs in a gene and increased protein abundance. Taken together, the results represent compelling evidence suggesting that X motifs may indeed contribute to the complex mechanisms of protein synthesis in extant genomes.

2. Method

After recalling a few basic definitions of circular codes, we define three classes of motifs: the X motifs constructed from the circular code X (1) identified in genes, as well as non-X motifs and random motifs used to evaluate the significance of X motifs. The statistical analyses of these motifs are based on very simple statistics, namely frequencies and mean frequencies, leading to clear biological results.

2.1. Definitions of circular code

We recall a few definitions without detailed explanation (i.e. without examples and figures) that are necessary for understanding the main properties of the X motifs obtained from the trinucleotide circular code X identified in genes (Michel, 2017, 2015; Arquès and Michel, 1996).

Notation 1. Let us denote the nucleotide 4-letter alphabet $B = \{A, C, G, T\}$ where A stands for adenine, C stands for cytosine, G stands for guanine and T stands for thymine. The trinucleotide set over B is denoted by $B^3 = \{AAA, \dots, TTT\}$. The set of non-empty words (words, respectively) over B is denoted by B^+ (B^*, respectively).

Definition 1. A set $S \subseteq B^+$ is a *code* if, for each $x_1, \dots, x_n, y_1, \dots, y_m \in S, n, m \geq 1$, the condition $x_1 \cdots x_n = y_1 \cdots y_m$ implies $n = m$ and $x_i = y_i$ for $i = 1, \dots, n$.

Definition 2. Any non-empty subset of the code B^3 is a code and called *trinucleotide code*.

Definition 3. The genetic code is a trinucleotide code. It defines a surjective map $\mathcal{g}: \tilde{B}^3 \longrightarrow P$ where $\tilde{B}^3 = B^3 \backslash \{TAA, TAG, TGA\}$ and P is the set of the 20 peptide components (amino acids). We also use the following notation: a sequence s of trinucleotides, i.e. a gene, codes a sequence noted $\mathcal{g}(s)$ of amino acids, i.e. a protein.

Example 1. $\mathcal{g}(GGA) = Gly$, $\mathcal{g}^{-1}(Gly) = \{GGA, GGC, GGG, GGT\}$ and $\mathcal{g}(GACATCCTG)$=*DIL* where *D*, *I* and *L* are amino acids.

Definition 4. A trinucleotide code $X \subseteq B^3$ is *circular* if, for each $x_1, \dots, x_n, y_1, \dots, y_m \in X$, $n, m \geq 1$, $r \in B^*$, $s \in B^+$, the conditions $sx_2 \cdots x_n r = y_1 \cdots y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ (empty word) and $x_i = y_i$ for $i = 1, \dots, n$.

We briefly recall the proof used to determine whether a code is circular or not, with the most recent and powerful approach which relates an oriented (directed) graph to a trinucleotide code.

Definition 5. (Fimmel *et al.*, 2016). Let $X \subseteq B^3$ be a trinucleotide code. The directed graph $\mathcal{G}(X) = (V(X), E(X))$ associated with X has a finite set of vertices $V(X)$ and a finite set of oriented edges $E(X)$ (ordered pairs $[v, w]$ where $v, w \in X$) defined as follows:

$$\begin{cases} V(X) = \{N_1, N_3, N_1N_2, N_2N_3 \colon N_1N_2N_3 \in \mathrm{X}\} \\ E(X) = \{[N_1, N_2N_3], [N_1N_2, N_3] \colon N_1N_2N_3 \in \mathrm{X}\} \end{cases}.$$

The theorem below gives a relation between a trinucleotide code which is circular and its associated graph.

Theorem 1. (Fimmel *et al.*, 2016). Let $X \subseteq B^3$ be a trinucleotide code. The following statements are equivalent:

(i) The code X is circular.

(ii) The graph $\mathcal{G}(X)$ is acyclic.

We also recall the results that characterize the comma-free codes and the strong comma-free codes by the longest paths in their associated graphs.

Theorem 2. (Fimmel *et al.*, 2016). Let $X \subseteq B^3$ be a trinucleotide circular code. The following statements are equivalent:

(i) X is comma-free.

(ii) The longest path in $\mathcal{G}(X)$ is of length at most 2.

Theorem 3. (Fimmel *et al.*, 2017). Let $X \subseteq B^3$ be a trinucleotide circular code. The following statements are equivalent:

(i) X is strong comma-free.

(ii) The longest path in $\mathcal{G}(X)$ is of length at most 1.

Thus, the reading frame is retrieved after the reading of 2 nucleotides with motifs from a strong comma-free code, of 3 nucleotides (1 trinucleotide) with motifs from a comma-free code and of at most 13 nucleotides (4 trinucleotides + 1 nucleotide) with motifs from circular codes.

The trinucleotide set X coding the reading frame in genes of bacteria, archaea, eukaryotes, plasmids and viruses is a maximal (20 trinucleotides) C^3 (the 2 shifted codes by permutation of X are also circular) self-complementary (10 trinucleotides of X are complementary to the 10 other trinucleotides of X) trinucleotide circular code (Michel 2017, 2015; Arquès and Michel, 1996).

2.2. Definition of X motifs, non-X motifs and random motifs

Definition 6. As in Michel *et al.* (2017), a X motif $m(X)$ constructed from the circular code X (1), is a word with cardinality $4 \leq c \leq 20$ trinucleotides and length $l \geq c \geq 4$ trinucleotides. Here, we consider only the X motifs $m(X)$ found in reading frame of genes.

Indeed, the X motifs $m(X)$ have a cardinality $c \leq 20$ trinucleotides as the circular code X (1) has 20 trinucleotides. The minimal length $l = 4$ trinucleotides was chosen based on the requirement for 13 nucleotides in order to retrieve the reading frame. The class of motifs of X with cardinality $c < 4$ are excluded here because they are mostly associated with the "pure" trinucleotide repeats often found in non-coding regions of the genome (Michel *et al.*, 2017; El Soufi and Michel, 2017).

The fundamental property of a X motif $m(X)$ is the ability to retrieve, synchronize and maintain the reading frame. Indeed, a window of 13 nucleotides located anywhere in a sequence generated from the circular code X (1) is sufficient to retrieve the reading (correct, construction) frame of the sequence. It is important to stress again that this window for retrieving the reading frame in a sequence can be located anywhere in the sequence, i.e. no other frame signal, including start and stop trinucleotides, is required to identify the reading frame.

Example 2. For the convenience of the reader, we give an example of a X motif $m(X) = m_1$ from the circular code X (1) in a sequence $s = \cdots AAAGGTGCCGAAGCCCTGGAGGAAAAG \ldots$ In the sequence s, there is a X motif $m_1 = GGTGCCGAAGCCCTGGAGGAA$ of cardinal $c = 5$ trinucleotides $\{CTG, GAA, GAG, GCC, GGT\}$ and length $l = 7$ trinucleotides. Note that m_1 cannot be extended to the left or to the right in s due to the presence of the periodic trinucleotide AAA (left) and the trinucleotide AAG (right) which both do not belong to the circular code X. Then, the reading frame of the sequence s can easily be deduced from the X motif m_1: $\ldots, AAA, GGT, GCC, GAA, GCC, CTG, GAG, GAA, AAG, \ldots$.

Definition 7. For simplification reasons, a non-X motif $m(\overline{X})$ is any word of any cardinality and length constructed from the nucleotide 4-letter alphabet $B = \{A, C, G, T\}$ except the X motifs $m(X)$ defined in

Definition 6. As for the X motifs $m(X)$, we only consider the non-X motifs $m(\bar{X})$ found in reading frame of genes.

Note that using this simplified notation, the class of motifs of X of cardinality $c < 4$ trinucleotides and length $l < 4$ trinucleotides belong to the non-X motifs $m(\bar{X})$.

In order to evaluate the statistical significance of X motifs in genes, we define, as in Michel *et al.* (2017), random motifs.

Definition 8. A R random motif $m(R)$ constructed from a random code R, is a word with cardinality $4 \leq c \leq 20$ trinucleotides and length $l \geq c \geq 4$ trinucleotides. A random code R is generated according to the properties of X, except its circularity property:

(i) R has a cardinality equal to 20 trinucleotides;

(ii) The total number of each nucleotide A, C, G and T in R is equal to 15 (note that $20 \times 3 = 15 \times 4$);

(iii) R has no stop trinucleotides $\{TAA, TAG, TGA\}$ and no periodic trinucleotides $\{AAA, CCC, GGG, TTT\}$;

(iv) R is not a circular code. Its associated graph $\mathcal{G}(R)$ is cyclic ($\mathcal{G}(R)$ being not shown).

As for the X motifs $m(X)$ and the non-X motifs $m(\bar{X})$, we only consider the R random motifs $m(R)$ found in reading frame of genes.

In order to obtain a statistically significant distribution of random codes, a set of 100 (different) random codes R are generated according to Definition 8. Examples of such random codes are given in Appendix in Michel *et al.* (2017).

Definition 9. We say that a letter $N_i \in B$ belongs to a motif m of length $l(m)$ if $1 \leq i \leq l(m)$.

Notation 2. $\mathcal{S} = \{X, \bar{X}, R\}$ denotes the three trinucleotide codes associated with the studied motifs $m(X)$, $m(\bar{X})$ and $m(R)$.

2.3. Multiple alignment of genes

In the following sections, we briefly recall the multiple alignment of genes and the notations used. A reference gene sequence $s_1 = \mathbb{R}$ (by convention here the reference sequence is the first sequence in the alignment) is aligned with its orthologous corresponding $n-1$ genes $s_2, \ldots, s_n$ where $s_2, \ldots, s_n \in B^+$. The genes $s_1, s_2, \ldots, s_n$ have respective lengths $|s_1|, |s_2|, \ldots, |s_n|$. Note that orthologous genes originate from a common DNA ancestral sequence and diverged after a speciation event.

A gene multiple alignment $s_1, s_2, \ldots, s_n$, $n \geq 2$, is a mapping z on the alphabet $(B \cup \{\varepsilon\})^n \setminus (\{\varepsilon\})^n$ whose projection on the 1st component is s_1, up to the projection on the nth component is s_n. Thus, a gene multiple alignment z of letter length l is noted

$$z = \begin{pmatrix} \bar{N}_{11} & \cdots & \bar{N}_{l1} \\ \bar{N}_{12} & \cdots & \bar{N}_{l2} \\ \vdots & \vdots & \vdots \\ \bar{N}_{1n} & \cdots & \bar{N}_{ln} \end{pmatrix}$$

with the reference sequence $\mathbb{R} = s_1 = \bar{N}_{11}, \cdots, \bar{N}_{l1}$, up to the sequence $s_n = \bar{N}_{1n}, \cdots, \bar{N}_{ln}$ such that the nucleotide $\bar{N}_{ji} \in B \cup \{\varepsilon\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, l$ and, where ε being classically associated with the gap symbol "-" or ".". An aligned tuple $\left(\bar{N}_{j1}, \ldots, \bar{N}_{ji}, \ldots, \bar{N}_{jn}\right)$ at the jth position such that $\bar{N}_{j1}, \bar{N}_{ji} \in B$ with

$\bar{N}_{j1} \neq \bar{N}_{ji}$ and $i \geq 2$ denotes the substitution of the jth nucleotide $\bar{N}_{j1}$ of $\mathbb{R}$ by the jth nucleotide $\bar{N}_{ji}$ of s_i. An aligned tuple $(\bar{N}_{j1}, \ldots, \bar{N}_{ji}, \ldots, \bar{N}_{jn})$ such that $\bar{N}_{j1} \in B$ and $\bar{N}_{ji} \in \{\varepsilon\}$ with $i \geq 2$ denotes the deletion of the jth nucleotide $\bar{N}_{j1}$ of $\mathbb{R}$. An aligned tuple $(\bar{N}_{j1}, \ldots, \bar{N}_{ji}, \ldots, \bar{N}_{jn})$ such that $\bar{N}_{j1} \in \{\varepsilon\}$ and $\bar{N}_{ji} \in B$ with $i \geq 2$ denotes the insertion of the jth nucleotide $\bar{N}_{ji}$ of s_i.

The X motifs $m(X)$, the non-X motifs $m(\bar{X})$ and the R random motifs $m(R)$ located in the gene multiple alignment belong to the alphabet $B \cup \{\varepsilon\}$, i.e. they may contain gaps. Those which are located in the gene sequence s_i, $i = 1, \ldots, n$, are noted $m(X, s_i)$, $m(\bar{X}, s_i)$ and $m(R, s_i)$, respectively, in particular, $m(X, \mathbb{R})$, $m(\bar{X}, \mathbb{R})$ and $m(R, \mathbb{R})$, respectively, in the reference gene $\mathbb{R}$. Figure 1 shows an example of a part of a gene multiple alignment.

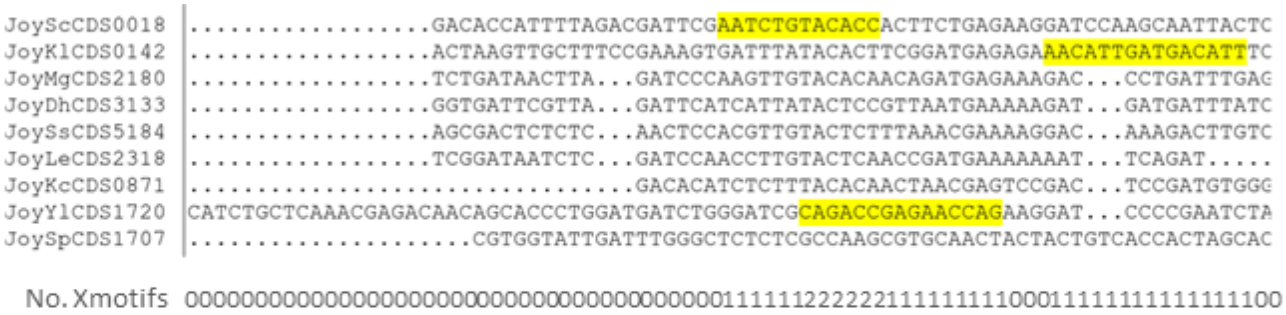

Figure 1. Screenshot of a yeast gene multiple alignment. The X motifs $m(X)$ in the reading frame of genes (Definition 6: cardinality $4 \leq c \leq 20$ trinucleotides and length $l \geq c \geq 4$ trinucleotides) are coloured in yellow. The reference (first) gene $s_1 = \mathbb{C}$ (*Saccharomyces Cerevisiae Sc*) contains one X motif $m(X, \mathbb{C})$ in reading frame and two non-X motifs $m(\bar{X}, \mathbb{C})$ in reading frame (Definition 7) without colour. Two X motifs are identified in the reading frame of two other yeast genes: $m(X, \mathbb{L})$ (*Kluyveromyces lactis Kl*, $s_2 = \mathbb{L}$) and $m(X, s_8)$ (*Yarrowia lipolyticai Yl*). The number of X motifs (last row) is used in the calculation of the positional conservation parameter (Section 2.4).

2.4. Positional conservation parameter of X motifs and random motifs

Here, we consider whether the position of X motifs is preserved within the genes from different organisms. To do this, for each column of a gene alignment, the number of organisms with a X motif at this position was calculated. For example, in Figure 1, the number of organisms with a X motif is equal to 0, 1, or 2. This number was normalized by the number of organisms having a nucleotide at that position in the alignment and not a gap.

Formally, we define a simple statistical parameter for analysing the positional conservation of motifs in the reference genes in the multiple alignments.

Definition 10. The positional conservation score $Ppc(m)$ of all motifs $m = (m(\mathcal{S}, \mathbb{R}))$, $\mathcal{S} \in \{X, R\}$ for studying X motifs and R random motifs, of letter lengths $l(m)$, m on the alphabet $B \cup \{\varepsilon\}$ (with gaps), in the reference genes $s_1 = \mathbb{R}$ in all the gene multiple alignments $s_1, s_2, \ldots, s_n$ is equal to

$$Ppc(m) = Ppc\big(m(\mathcal{S}, \mathbb{R})\big) = \frac{1}{\sum_{m \in \mathbb{R}} l(m)} \sum_{m \in \mathbb{R}} \sum_{j=1}^{l(m)} \frac{1}{Nb_j} \sum_{i=1}^{n} \delta_{i,j}$$

where

$$\delta_{i,j} = \begin{cases} 1 \text{ if } N_{ji} \in B \text{ and } N_{ji} \in m(\mathcal{S}, s_i) \\ 0 \text{ otherwise} \end{cases},$$

Nb_j is the number of nucleotides without gaps at position j in the multiple alignment of n genes, $2 \leq Nb_j \leq n$ for $j = 1, \ldots, l(m)$. The condition $N_{ji} \in B$ and $N_{ji} \in m(\mathcal{S}, s_i)$ signifies that the letter N_{ji} at the jth position in the gene s_i is a nucleotide and not a gap, and belongs to a motif m.

Remark 1. The positional score $Ppc(m) \in]0,1]$. The positional conservation Ppc of the motif m is the lowest in the alignment when $Ppc(m) \approx 0$, i.e. when the motif m in the reference genome is aligned with zero motifs in the other genomes. The positional conservation Ppc of the motif m is the highest in the alignment when $Ppc(m) = 1$ corresponding to the case where all genes without gaps have X motifs in the same position as the reference genes.

2.5. Pairwise alignment parameters of X motifs and non-X motifs

A pairwise alignment is a multiple alignment z with $n = 2$ sequences of letter length l such that their nucleotides $N \in B \cup \{\varepsilon\}$ (with gaps). Several classical pairwise alignment parameters are used to estimate the conservation of a pairwise alignment, including (i) the percentage of alignment positions that contain identical nucleotides, and (ii) the ratio of synonymous to non-synonymous substitutions. These parameters are briefly recalled in the following definitions and are illustrated with examples from Figure 2, and Table 1 and Table 2.

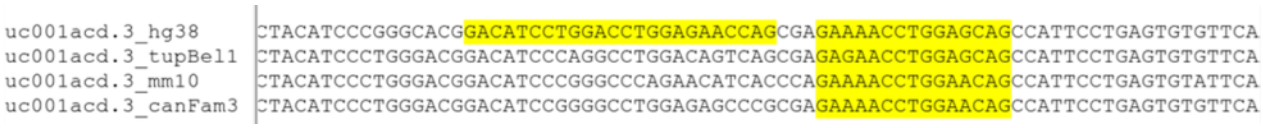

Figure 2. Screenshot of a mammal gene multiple alignment. The X motifs $m(X)$ in the reading frame of genes (Definition 6: cardinality $4 \leq c \leq 20$ trinucleotides and length $l \geq c \geq 4$ trinucleotides) are coloured in yellow. The reference (first) gene $s_1 = \mathbb{H}$ (*Homo sapiens hg38*) contains two X motifs $m(X, \mathbb{H})$ in reading frame and three non-X motifs $m(\bar{X}, \mathbb{H})$ in reading frame (Definition 7) without colour. The 2nd non-X motif $m(\bar{X}, \mathbb{H})$ is composed of one trinucleotide $CGA \notin X$ (1). Three X motifs are identified in the reading frame of three other mammal genes: $m(X, s_2)$ (*Tupaia Belangeri tupBel1*), $m(X, \mathbb{M})$ (*Mus musculus mm10*, $s_3 = \mathbb{M}$) and $m(X, s_4)$ (*Canis Lupus Familiaris canFam3*).

Reference gene $\mathbb{H}$	1st X motif $m(X, \mathbb{H})$								2nd X motif $m(X, \mathbb{H})$				
Protein $g(s)$ of $\mathbb{H}$	*D*	*I*	*L*	*D*	*L*	*E*	*N*	*Q*	*E*	*N*	*L*	*E*	*Q*
Gene s of $\mathbb{H}$	*GAC*	*ATC*	*CTG*	*GAC*	*CTG*	*GAG*	*AAC*	*CAG*	*GAA*	*AAC*	*CTG*	*GAG*	*CAG*
Gene s' of $\mathbb{M}$	*GAC*	*ATC*	*CCG*	*GGC*	*CCA*	*GAA*	*CAT*	*CAC*	*GAA*	*AAC*	*CTG*	*GAA*	*CAG*
Protein $g(s')$ of $\mathbb{M}$	*D*	*I*	*P*	*G*	*P*	*E*	*H*	*H*	*E*	*N*	*L*	*E*	*Q*

Table 1. From Figure 2, the alignment of the two X motifs $m(X, \mathbb{H})$ in reading frame of total length 39 nucleotides of the reference gene $s_1 = \mathbb{H}$ (*Homo sapiens hg38*) and the gene $\mathbb{M}$ (*Mus musculus mm10*). There are 30 identical nucleotide pairs and 9 different nucleotide pairs (underlined). The protein alignment associated with the gene alignment is given by applying the universal genetic code map g (Definition 3) to each trinucleotide.

Reference gene $\mathbb{H}$	1st non-X motif $m(\bar{X},\mathbb{H})$					2nd $m(\bar{X},\mathbb{H})$	3rd non-X motif $m(\bar{X},\mathbb{H})$					
Protein $g(s)$ of $\mathbb{H}$	Y	I	P	G	T	R	P	F	L	S	V	F
Gene s of $\mathbb{H}$	TAC	ATC	CCG	GGC	ACG	CGA	CCA	TTC	CTG	AGT	GTG	TTC
Gene s' of $\mathbb{M}$	TAC	ATC	$CC\underline{T}$	$GG\underline{G}$	ACG	$C\underline{C}A$	CCA	TTC	CTG	AGT	$GT\underline{A}$	TTC
Protein $g(s')$ of $\mathbb{M}$	Y	I	P	G	T	P	P	F	L	S	V	F

Table 2. From Figure 2, the alignment of the three non-X motifs $m(\bar{X},\mathbb{H})$ in reading frame of total length 36 nucleotides of the reference gene $s_1 = \mathbb{H}$ (*Homo sapiens hg38*) and the gene $\mathbb{M}$ (*Mus musculus mm10*). There are 32 identical nucleotide pairs and 4 different nucleotide pairs (underlined). The protein alignment associated with the gene alignment is given by applying the universal genetic code map g (Definition 3) to each trinucleotide.

Definition 11. The percentage $Pid(m)$ of identical nucleotides of all motifs $m = \big(m(\mathcal{S},\mathbb{R})\big), \mathcal{S} \in \{X,\bar{X}\}$ for studying X motifs and non-X motifs, of letter lengths $l(m), m$ on the alphabet $B \cup \{\varepsilon\}$ (with gaps), in the reference genes $s_1 = \mathbb{R}$ in all the gene pairwise alignments s_1 and s_2 is equal to

$$Pid(m) = Pid\big(m(\mathcal{S},\mathbb{R})\big) = \frac{1}{\sum_{m\in\mathbb{R}} l(m)} \sum_{m\in\mathbb{R}} \sum_{i=1}^{l(m)} \delta_i$$

where the operator δ_i, $1 \le i \le l(m)$, associated with a pair of letters N is defined by

$$\delta_i = \begin{cases} 1 & \text{if } N_{i1} \in B \text{ and } N_{i1} = N_{i2} \\ 0 & \text{otherwise} \end{cases}.$$

Example 3. From Table 1, $Pid\big(m(X,\mathbb{H})\big) = 30/39 = 76.92\%$.

Definition 12. Let $f_i(c)$, $g_i(c)$ respectively, be the fraction of synonymous, non-synonymous respectively, potential substitutions at the ith site, $i = 1,2,3$, of a given codon $c = N_1N_2N_3$. Then, the numbers $Ns(c)$, $Nns(c)$ respectively, of synonymous, non-synonymous respectively, sites for a given codon c, are defined according to Nei and Gojobori (1986) by $Ns(c) = \sum_{i=1}^{3} f_i(c)$ and $Nns(c) = \sum_{i=1}^{3} g_i(c) = \sum_{i=1}^{3}\big(1 - f_i(c)\big) = 3 - Ns(c)$.

Example 4. In the case of the codon $c = CTG$ coding the amino acid $g(c) = Leu$, $f_1(Leu) = \frac{1}{3}$ as only the 1st site substitution $CTG \longrightarrow TTG$ is synonymous out of ATG, GTG and TTG, $f_2(Leu) = 0$ as there is no 2nd site synonymous substitution out of CAG, CCG and CGG, and $f_3(Leu) = \frac{3}{3} = 1$ as all the 3rd site substitutions are synonymous out of CTA, CTC and CTT. Then, $Ns(Leu) = \frac{1}{3} + 0 + 1 = \frac{4}{3}$ and $Nns(Leu) = 3 - \frac{4}{3} = \frac{5}{3}$.

The definitions of $Ns(c)$ and $Nns(c)$ for a given codon are naturally extended to a motif m.

Definition 13. The potential numbers $Ns(m)$, $Nns(m)$ respectively, of synonymous, non-synonymous respectively, sites for a motif $m = \big(m(\mathcal{S},\mathbb{R})\big), \mathcal{S} \in \{X,\bar{X}\}$ for studying X motifs and non-X motifs, of letter length l, m on the alphabet B (without gaps), are equal to $Ns(m) = \sum_{c\in m} Ns(c)$ where $Ns(c)$ is defined in Definition 12, and $Nns(m) = l - Ns(m)$. Then, $Ns(m)$ and $Nns(m)$ are computed for all motifs m in the reference sequence s_1 of the gene pairwise alignments.

Example 5. From Table 1, the potential numbers $Ns(m)$ and $Nns(m)$ of synonymous and non-synonymous sites for the two X motifs are $Ns\big(m(X,\mathbb{H})\big) = \frac{1}{3}+\frac{2}{3}+\frac{4}{3}+\frac{1}{3}+\frac{4}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{4}{3}+\frac{1}{3}+\frac{1}{3} = \frac{23}{3} \approx 7.67$ and $Nns\big(m(X,\mathbb{H})\big) = 39 - \frac{23}{3} = \frac{94}{3} \approx 31.33$.

Definition 14. Let $Os(m)$, $Ons(m)$ respectively, be the observed numbers of synonymous, non-synonymous respectively, substitutions of a reference motif $m = \big(m(\mathcal{S},\mathbb{R})\big), \mathcal{S} \in \{X, \bar{X}\}$ in the gene $s_1 = \mathbb{R}$ for studying X motifs and non-X motifs, in the motif m' in gene s_2 of letter lengths $l(m) = l(m')$, m, m' on the alphabet B (without gaps), in all gene pairwise alignments s_1 and s_2.

Remark 2. $Os(m) + Ons(m) = l(m) - \sum_{i=1}^{l(m)} \delta_i$ where δ_i is defined in Definition 11.

Example 6. From Table 1, $Os\big(m(X,\mathbb{H})\big) = 4$ (four synonymous substitutions: the 3rd site of CTG (L), GAG $(E) \longrightarrow GAA$ (E), the 3rd site of AAC (N) and GAG $(E) \longrightarrow GAA$ (E)) and $Ons\big(m(X,\mathbb{H})\big) = 5$ (five non-synonymous substitutions: CTG $(L) \longrightarrow CCG$ (P), GAC $(D) \longrightarrow GGC$ (G), CTG $(L) \longrightarrow CCA$ (P), AAC $(N) \longrightarrow CAT$ (H) and CAG $(Q) \longrightarrow CAC$ (H)).

Definition 15. The percentages $Ps(m)$, $Pns(m)$ respectively, of synonymous, non-synonymous respectively, substitutions of a reference motif $m = \big(m(\mathcal{S},\mathbb{R})\big)$, $\mathcal{S} \in \{X, \bar{X}\}$ in the gene $s_1 = \mathbb{R}$ for studying X motifs and non-X motifs, in the motif m' in the gene s_2 of letter lengths $l(m) = l(m')$, m, m' on the alphabet B (without gaps), in all gene pairwise alignments s_1 and s_2, are equal to $Ps(m) = Os(m)/Ns(m)$ and $Pns(m) = Ons(m)/Nns(m)$ where $Os(m)$ and $Ons(m)$ are defined in Definition 14, and $Ns(m)$ and $Nns(m)$ in Definition 13.

Example 7. From Table 1, $Ps\big(m(X,\mathbb{H})\big) = \frac{4}{\frac{23}{3}} = \frac{12}{23} \approx 0.52$ and $Pns\big(m(X,\mathbb{H})\big) = \frac{5}{\frac{94}{3}} = \frac{15}{94} \approx 0.16$.

Example 8 summarizes the parameters for the three non-X motifs of Table 2.

Example 8. From Table 2, $Pid\big(m(\bar{X},\mathbb{H})\big) = 32/36 = 88.89\%$, $Ns\big(m(\bar{X},\mathbb{H})\big) = \frac{1}{3}+\frac{2}{3}+\frac{3}{3}+\frac{3}{3}+\frac{3}{3}+\frac{4}{3}+\frac{3}{3}+\frac{1}{3}+\frac{4}{3}+\frac{1}{3}+\frac{3}{3}+\frac{1}{3} = \frac{29}{3} \approx 9.67$, $Nns\big(m(\bar{X},\mathbb{H})\big) = 3 \times 12 - \frac{29}{3} = \frac{79}{3} \approx 26.33$, $Os\big(m(\bar{X},\mathbb{H})\big) = 3$ (three synonymous substitutions: CCG $(P) \longrightarrow CCT$ (P), GGC $(G) \longrightarrow GGG$ (G) and GTG $(V) \longrightarrow GTA$ (V) and $Ons\big(m(\bar{X},\mathbb{H})\big) = 1$ (one non-synonymous substitution: CGA $(R) \longrightarrow CCA$ (P)), $Ps\big(m(\bar{X},\mathbb{H})\big) = \frac{3}{\frac{29}{3}} = \frac{9}{29} \approx 0.31$ and $Pns\big(m(\bar{X},\mathbb{H})\big) = \frac{1}{\frac{79}{3}} = \frac{3}{79} \approx 0.04$.

2.6. Codon substitution matrix of X motifs and random motifs

We define a codon (trinucleotide) substitution matrix $\mathbf{A}\big(m(\mathcal{S},\mathbb{R})\big), \mathcal{S} \in \{X, R\}$ for studying X motifs and R random motifs, in the reference genes $s_1 = \mathbb{R}$ in all the gene multiple alignments $s_1, s_2, \ldots, s_n$. The codon substitution matrix $\mathbf{A}\big(m(\mathcal{S},\mathbb{R})\big) = \big[a_{ij}\big]_{1\le i\le 64, 1\le j\le 64}$ of size 64×64 (square matrix) where the 64 rows and the 64 columns are associated with the 64 codons B^3, has element $a_{ij} = Nb(\mathcal{S}[j] \to B^3[i])$ in row i and column j referring to the number of substitutions of codon $\mathcal{S}[j]$ (jth codon of $\mathcal{S}$) of the motifs m (in the reference genes $\mathbb{R}$) by the aligned codon $B^3[i]$ (ith codon of B^3) of the $n-1$ genes $s_2, \ldots, s_n$.

We define the normalized matrix $\mathbf{B}\big(m(\mathcal{S},\mathbb{R})\big) = \big[b_{ij}\big]_{1\leq i\leq 64, 1\leq j\leq 64}$ with element $b_{ij} = a_{ij}/a_{.j}$ where $a_{.j} = \sum_{k=1}^{64} a_{kj}$ $(a_{.j} \neq 0)$, for $1 \leq i \leq 64$ and $1 \leq j \leq 64$, such that it is stochastic in column. The column normalization, rather than a full matrix normalization, allows the codons to be compared whatever the codon usage.

Remark 3. The elements a_{ii} of **A** and b_{ii} of **B** can be indexed either by numbers or by the codons B^3.

Remark 4. The diagonal elements a_{ii} of **A** and b_{ii} of **B** can be different from 0.

Remark 5. For a given code $\mathcal{S}$, i.e. the circular code X or a given random code R, with $Card(\mathcal{S}) = 20$ codons $\mathcal{S}$ (Section 2.2), the matrices **A** and **B** have 20 non-empty codon columns and $64 - 20 = 44$ empty codon columns. However, the 20 codon columns of **A** and **B** vary with each (different) random code R, which obviously differ from the 20 codon columns of the circular code X.

Example 9. An example of construction of the matrices **A** and **B** is given from the alignment in Table 3. The first codon column leads to the submatrix **A** given in Table 4. The procedure is iterated for each codon column and leads to the matrix **A** given in Table 5. The normalized matrix **B** is given in Table 6.

s_1	GAG	GAC	ATC	CTG	GAC	CTG	AAC	CAG
s_2	GAC	GAC	ATC	CCA	GGC	CTG	AGT	CAG
s_3	GAA	GAC	ATC	CCG	GGC	CCA	CAT	CAC
s_4	GAG	GAC	ATC	CGG	GGC	CTG	AGC	CCG

Table 3. Example of a multiple alignment of four genes where $s_1 = \mathbb{R}$ is the reference gene.

A	GAG	
GAA	1	$GAG \longrightarrow GAA$
GAC	1	$GAG \longrightarrow GAC$
GAG	1	$GAG \longrightarrow GAG$

Table 4. Codon substitution submatrix **A** of the first codon column from example of Table 3.

A	AAC	ATC	CAG	CTG	GAC	GAG
AGC	1					
AGT	1					
ATC		3				
CAC			1			
CAG			1			
CAT	1					
CCA				2		
CCG			1	1		
CGG				1		
CTG				2		
GAA						1
GAC					3	1
GAG						1
GGC					3	

Table 5. Codon substitution matrix **A** from example of Table 3. The remaining codon rows and columns equal to 0 are not shown.

B	*AAC*	*ATC*	*CAG*	*CTG*	*GAC*	*GAG*
AGC	1/3					
AGT	1/3					
ATC		1				
CAC			1/3			
CAG			1/3			
CAT	1/3					
CCA				1/3		
CCG			1/3	1/6		
CGG				1/6		
CTG				1/3		
GAA						1/3
GAC					1/2	1/3
GAG						1/3
GGC					1/2	
Sum	1	1	1	1	1	1

Table 6. Normalized matrix **B** from example of Table 3. The remaining codon rows and columns equal to 0 are not shown.

2.7. Amino acid conservation parameter of X motifs and random motifs

We define a simple statistical parameter for analysing the conservation of X motifs and R random motifs for the 12 amino acids $\mathcal{X}$ (2) coded by the circular code X, in the reference genes $s_1 = \mathbb{R}$ in all the gene multiple alignments $s_1, s_2, \ldots, s_n$.

Definition 16. The percentage $Paac(m(\mathcal{S},\mathbb{R}),p)$ of conservation of X codons per amino acid p (peptide component) coded by all the motifs $m = \big(m(\mathcal{S},\mathbb{R})\big)$, $\mathcal{S} \in \{X, R\}$ for studying X motifs and R random motifs, in the reference genes $s_1 = \mathbb{R}$ in all the gene multiple alignments $s_1, s_2, \ldots, s_n$, is equal to

$$Paac(m(\mathcal{S},\mathbb{R}),p) = \frac{1}{Card(\mathcal{g}^{-1}(p) \cap \mathcal{S})} \sum_{\substack{i,j \in \mathcal{g}^{-1}(p) \\ i,j \in \mathcal{S}}} b_{ij}\big(m(\mathcal{S},\mathbb{R})\big)$$

where $p \in \mathcal{X} = \{A, D, E, F, G, I, L, N, Q, T, V, Y\}$ (2), $b_{ij}\big(m(\mathcal{S},\mathbb{R})\big)$ is the element of the normalized matrix **B** of $\mathcal{S}$ defined in Section 2.6 and the inverse genetic code map $\mathcal{g}^{-1}$ defined in Definition 3.

Definition 17. The mean percentage $\bar{P}aac(m(\mathcal{S},\mathbb{R}),\mathcal{X})$ of conservation of X codons in the 12 amino acids $\mathcal{X}$ (2) coded by all the motifs $m = \big(m(\mathcal{S},\mathbb{R})\big)$, $\mathcal{S} \in \{X, R\}$, in the reference genes $s_1 = \mathbb{R}$ in all the gene multiple alignments $s_1, s_2, \ldots, s_n$, is equal to

$$\bar{P}aac(m(\mathcal{S},\mathbb{R}),\mathcal{X}) = \frac{1}{Card(\mathcal{X})} \sum_{p \in \mathcal{X}} Paac(m(\mathcal{S},\mathbb{R}),p)$$

where $p \in \mathcal{X} = \{A, D, E, F, G, I, L, N, Q, T, V, Y\}$ (2) and $Paac(m(\mathcal{S},\mathbb{R}),p)$ defined in Definition 16.

Remark 6. The mean percentage $\bar{P}aac$ give the same statistical weight for each amino acid.

Definition 18. To achieve a strong statistical significance, we use the information from the 100 random codes R $(R_1, \ldots, R_{100})$, and not only one random code, the percentage $Paac(\overline{m}(R,\mathbb{R}),p)$ of conservation of X codons per amino acid p coded by the R mean random motifs $\overline{m}(R,\mathbb{R})$ in the reference genes $s_1 = \mathbb{R}$ in all the gene multiple alignments $s_1, s_2, \ldots, s_n$, is equal to

$$Paac(\overline{m}(R,\mathbb{R}),p)=\frac{1}{\sum_{k=1}^{100}\delta_k}\sum_{k=1}^{100}Paac(m(R_k,\mathbb{R}),p)$$

where $p\in\mathcal{X}=\{A,D,E,F,G,I,L,N,Q,T,V,Y\}$ (2), $Paac(m(R_k,\mathbb{R}),p)$ defined in Definition 16 and $\delta_k=1$ if $\mathcal{g}^{-1}(p)\cap R_k\neq\emptyset$ (i.e. the random code R_k can code the amino acid p) and $\delta_k=0$ otherwise.

Remark 7. For the R mean random motifs $\overline{m}(R,\mathbb{R})$, we only analyse in the mean matrix $\overline{\mathbf{B}}$ the trinucleotides coding the 12 amino acids $\mathcal{X}$ of the circular code X.

Definition 19. The mean percentage $\bar{P}aac(\overline{m}(R,\mathbb{R}),\mathcal{X})$ of conservation of X codons in the 12 amino acids $\mathcal{X}$ (2) coded by the R mean random motifs $\overline{m}(R,\mathbb{R})$ in the reference genes $s_1=\mathbb{R}$ in all the gene multiple alignments $s_1,s_2,\ldots,s_n$, is equal to

$$\bar{P}aac(\overline{m}(R,\mathbb{R}),\mathcal{X})=\frac{1}{Card(\mathcal{X})}\sum_{p\in\mathcal{X}}Paac(\overline{m}(R,\mathbb{R}),p)$$

where $p\in\mathcal{X}=\{A,D,E,F,G,I,L,N,Q,T,V,Y\}$ (2) and $Paac(\overline{m}(R,\mathbb{R}),p)$ defined in Definition 18.

2.8. Data

In order to increase the significance of the results, two classes of independent alignments are investigated. The first class of alignments is based on four mammal genomes, which represent highly evolved species and closely related genomes. The second class of alignments is built from nine yeast genomes, which represent the simplest eukaryotes and are more divergent. The human genome for the first class and the *Saccharomyces cerevisiae* genome for the second class are taken as reference genomes as they are well documented model organisms.

2.8.1. Mammal gene alignments

From the mammalian gene multiple alignments available on the UCSC site (https://bds.mpi-cbg.de/hillerlab/ 144VertebrateAlignment_CESAR/, Sharma and Hiller, 2017), we have used genes from four well annotated genomes. Table 7 shows some summary statistics of the four selected mammal genomes.

Genome name	Identification	Number of genes	Nucleotide length of genes
Canis lupus familiaris	*canFam3*	21,137	34,379,490
Homo sapiens	*hg38* (ℍ)	22,352	36,808,167
Mus musculus	*mm10* (𝕄)	20,178	33,519,381
Tupaia belangeri	*tupBel1*	18,485	23,387,559

Table 7. Genomes of four mammals.

H. sapiens (*hg38*, ℍ) is taken as the reference genome and is present in each of the 22,352 gene alignments. One or two corresponding genes from the three other species may be missing, in which case the corresponding genes are replaced by gaps in the alignment.

2.8.2. Yeast gene alignments

For the yeast multiple alignments, the protein sequences of nine different yeasts and the localization of the corresponding nucleic acid sequence on the chromosomes (Table 8) are obtained from the NCBI Genbank (https://www.ncbi.nlm.nih.gov/genbank/).

Genome name	Identification	Number of genes	Nucleotide length of genes
Debaryomyces hansenii	*Dh*	6288	7,506,066
Kluyveromyces lactis	*Kl* (𝕃)	5085	7,729,998
Kuraishia capsulata	*Kc*	5989	6,911,424
Lodderomyces elongisporus	*Le*	5799	7,110,237
Meyerozyma guilliermondii	*Mg*	5920	6,633,972
Saccharomyces cerevisiae	*Sc* (ℂ)	6008	8,246,529
Scheffersomyces stipitis	*Ss*	5818	6,991,422
Schizosaccharomyces pombe	*Sp*	4980	5,614,506
Yarrowia lipolytica	*Yl*	6472	6,762,072

Table 8. Genomes of nine yeasts.

S. cerevisiae (Sc, ℂ) is taken as the reference genome. A BLAST (Altschul *et al.*, 1997) database of all protein sequences of these nine organisms is created. For all protein sequences of *S. cerevisiae*, a BLAST search in this database is performed. Then, the protein alignments containing from 2 to 9 sequences are obtained using ClustalW (Thompson *et al.*, 1994). The corresponding nucleic sequence alignments are created by localizing each amino acid on the genome. The BLAST searches, alignments and some data analyses were performed using our in-house software platform Gscope (R. Ripp, unpublished, details in Section 2.9).

2.9. Software development

In a nucleic sequence alignment, the X and R random motifs are localized in the genes using a program developed in the Java language (El Soufi and Michel, 2017). The program takes optional parameters that define the minimum cardinality c (in trinucleotides) and the length l (in trinucleotides) of the X and R motifs searched. The X and R motifs verify Definition 6 (cardinality $c \geq 4$ trinucleotides and with any length $l \geq c \geq 4$ trinucleotides). Although the X and R motifs are contiguous in the gene sequences, gaps may be inserted during the alignment process.

Gscope is an integrated platform allowing the analysis of all kinds of genomic data. It is written in Tcl/Tk and runs under all operating systems. It is specifically designed to perform high throughput analyses. Gscope includes the tools necessary to create the basic data, analysis tools and visualization interfaces. It also allows the creation of SQL relational databases and the querying and display of the available information through a web based interface (Wscope).

3. Results

The results presented below are based on basic frequency statistics and their biological significance is clear. In order to evaluate the statistical significance of the different results presented below, we chose an approach that involved comparing the results obtained for the X motifs with those obtained for R random motifs generated by 100 (different) random codes R. This approach avoids the problems associated with defining statistical hypotheses about the nucleotide composition, the length and the random model of the different regions of the genome. The main disadvantage of our approach is the additional computational resources required to obtain the results for the 100 random codes.

This section is divided into two main parts. In the first part, we estimate the evolutionary conservation of X motifs in two large-scale sets of genes from mammal and yeast species (~20,000 and ~6000 genes, respectively). In the second part, we evaluate the potential functional activity of the X motifs, by correlating them with existing experimental data.

3.1. Evolutionary conservation of X motifs in mammal and yeast genes

3.1.1. Enrichment of X motifs in mammal and yeast genes

We first investigated the occurrence number and codon length of X motifs in the two sets of genes. Figure 3 and Figure 4 show a very strong enrichment of X motifs in both mammal and yeast genes compared to the R random motifs from the 100 (different) random codes R. The number of X motifs in mammal genes is equal to 173,390, compared to a mean number of 60,330 R motifs. This difference is significant according to a one-sided Student's t-test with value $p \approx 10^{-82}$. The number of X motifs in yeast genes is equal to 35,833, compared to a mean number of 15,853 R motifs. Again, this difference is significant according to a one-sided Student's t-test with value $p \approx 10^{-75}$. This result is an additional and strong confirmation of the enrichment of X motifs in genes previously observed in the yeast *S. cerevisiae* (Michel *et al.*, 2017).

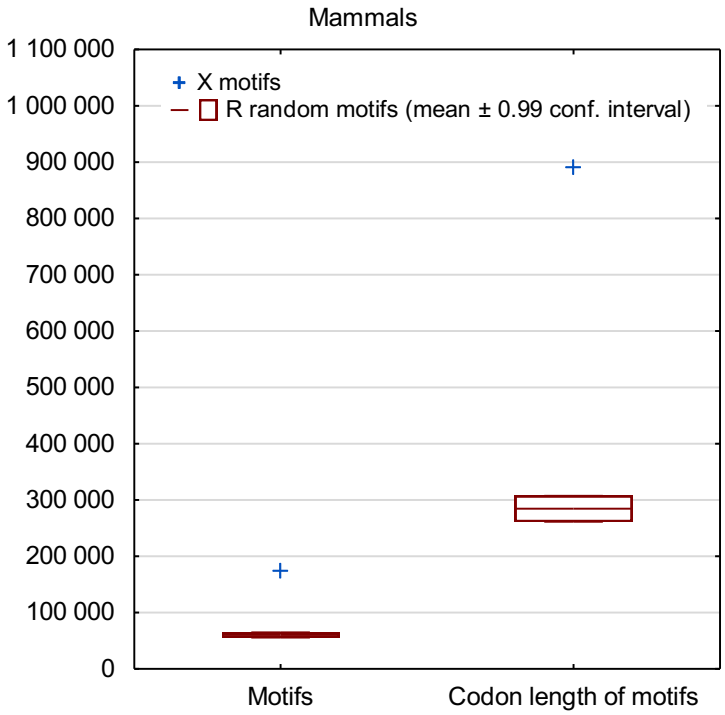

Figure 3. Comparison of the number of *X* and *R* random motifs and their codon lengths in mammalian genes. The number of *X* motifs is represented with a blue cross. The distribution of the *R* random motifs from the 100 random codes *R* is indicated by boxplots representing the mean and ±0.99 confidence interval.

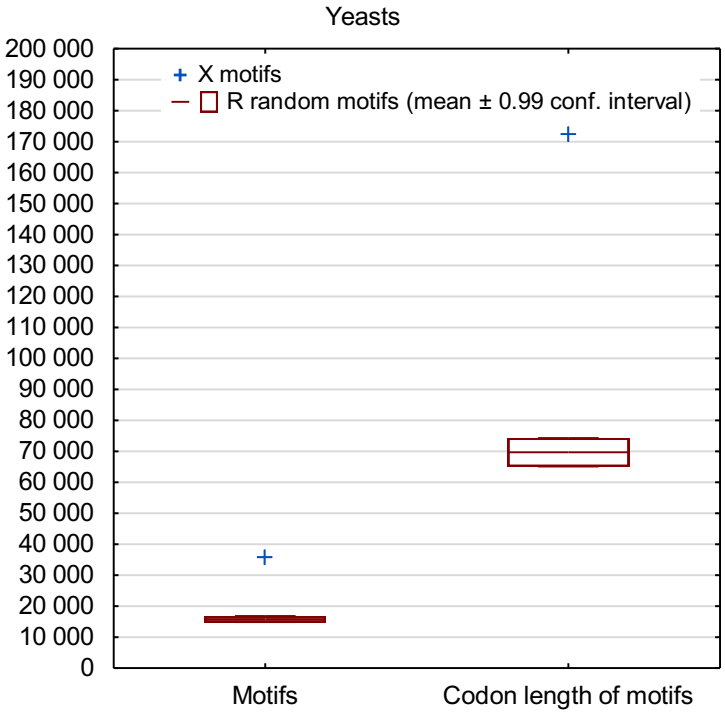

Figure 4. Comparison of the number of *X* and *R* random motifs and their codon lengths in the yeast genes. The number of *X* motifs is represented with a blue cross. The distribution of the *R* random motifs from the 100 random codes *R* is indicated by boxplots representing the mean and ±0.99 confidence interval.

3.1.2. Positional conservation of X motifs in mammal and yeast genes

We then used the multiple alignments corresponding to the 22,352 mammal genes and the 6008 yeast genes, to calculate the positional conservation scores $Ppc(m)$ (Definition 10; $m = \left(m(\mathcal{S}, \mathbb{R})\right), \mathcal{S} \in \{X, R\}$) for the X motifs and the R random motifs from the 100 (different) random codes R, shown in Figure 5 and Figure 6, respectively. The positional conservation score Ppc measures the number of motifs that are found in the same columns in a given multiple alignment. For both mammals and yeasts, the number of X motifs with the highest positional conservation score $Ppc = 1$ was higher than the number of R motifs. In contrast, the number of X motifs with the lowest positional conservation score $Ppc < 0.25$ was much lower than the number of R motifs. A one sample Wilcoxon signed rank indicated that the X motifs and the R motifs have significantly different medians with two-sided values $p = 0.031$ for the mammals and $p = 0.016$ for the yeasts.

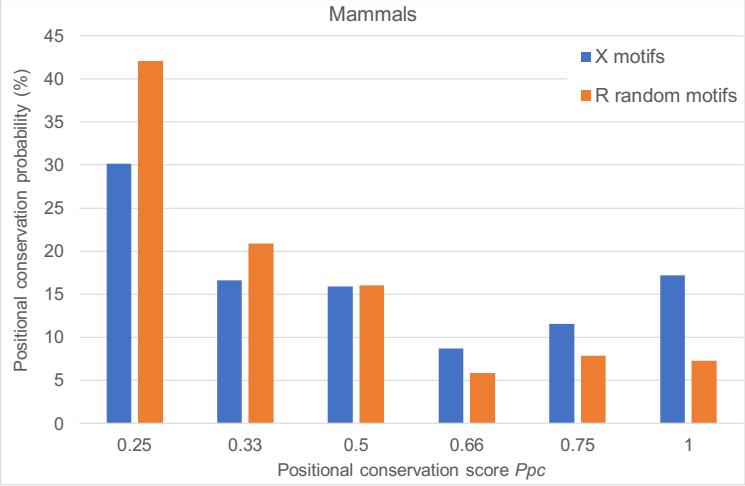

Figure 5. Positional conservation probability (%) of X motifs and R random motifs in the mammal gene multiple alignments with respect to the human reference genes $\mathbb{H}$ as a function of the score $Ppc(m)$ (Definition 10; $m = \left(m(\mathcal{S}, \mathbb{H})\right)$, $\mathcal{S} \in \{X, R\}$) varying from 0 (no conservation in the alignment) to 1 (highest conservation in the alignment).

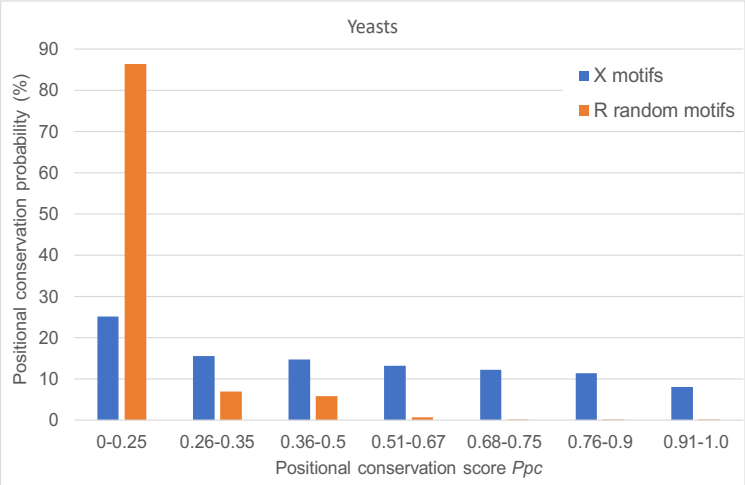

Figure 6. Positional conservation probability (%) of X motifs and R random motifs in the yeast gene multiple alignments with respect to the *S. cerevisiae* reference genes $\mathbb{C}$ as a function of the score $Ppc(m)$ (Definition 10; $m = \left(m(\mathcal{S}, \mathbb{C})\right)$, $\mathcal{S} \in \{X, R\}$) varying from 0 (no conservation in the alignment) to 1 (highest conservation in the alignment).

We conclude that X motifs are more likely to be preserved in the same position in the orthologous genes of mammals and yeasts than R random motifs.

3.1.3. Sequence conservation of X motifs in mammal and yeast genes

In the previous section, we showed that X motifs tend to occur at the same positions in orthologous genes from different organisms. To investigate the level of sequence conservation within the X motifs that are found at the same position, we computed several classical pairwise alignment parameters that were defined in Section 2.5.

We first calculated the percentage $Pid(m)$ (Definition 11; $m = \left(m(\mathcal{S}, \mathbb{R})\right)$, $\mathcal{S} \in \{X, \bar{X}\}$), of identical nucleotides in the aligned X motifs and compared it to the percentage of identical nucleotides in the aligned non-X motifs (alignment columns with no X motifs). For this initial analysis, we selected two organisms from each of the mammal and yeast gene sets. For the mammals, 14,681 gene pairwise alignments containing both human $\mathbb{H}$ and mouse $\mathbb{M}$ genes were used, and for the yeasts, 1088 gene pairwise alignments containing both *S. cerevisiae* $\mathbb{C}$ and *K. lactis* $\mathbb{L}$ genes were used. The Pid observed in X motifs was 87.44% for $\mathbb{H}$-$\mathbb{M}$ alignments, and 59.88% for $\mathbb{C}$-$\mathbb{L}$ alignments. In comparison, the Pid observed in non-X motifs was 77.56% for $\mathbb{H}$-$\mathbb{M}$ alignments, and 53.94% for $\mathbb{C}$-$\mathbb{L}$ alignments. For $\mathbb{H}$-$\mathbb{M}$ alignments ($n = 14{,}681$), a $\mathcal{X}^2$ test shows a strongly significant difference between the Pid values of X motifs (87.44%) and non-X motifs (77.56%) with one-sided value $p \approx 10^{-110}$. For $\mathbb{C}$-$\mathbb{L}$ alignments ($n = 1088$), a $\mathcal{X}^2$ test shows a significant difference between the Pid values of X motifs (59.88%) and non-X motifs (53.94%) with a one-sided value $p \approx 0.005$. Thus, the sequences of X motifs are generally more conserved, i.e. evolve more slowly, than the remainder of the gene alignments.

This increased conservation of X motifs indicates that their sequences are maintained during natural selection, and may reflect differences in the strengths of positive selection or purifying selection. To understand the relative contributions of these different modes of selection better, we calculated the ratio $Pns(m)/Ps(m)$ (Definition 15; $m = \left(m(\mathcal{S}, \mathbb{R})\right)$, $\mathcal{S} \in \{X, \bar{X}\}$), of non-synonymous to synonymous substitutions for $\mathbb{H}$-$\mathbb{M}$ and $\mathbb{C}$-$\mathbb{L}$ alignments (Table 9 and Table 10). This ratio is commonly used to infer purifying ($Pns/Ps < 1$) or positive ($Pns/Ps > 1$) selection in genes. It is important to note that a non-synonymous substitution implies a change in the amino acid in the translated protein, while a synonymous substitution only changes the codon: the original codon is replaced by another codon coding for the same amino acid.

$\mathbb{H}$-$\mathbb{M}$ alignment	Nns	Ns	Ons	Os	Pns	Ps	Pns/Ps
X motifs	1,611,224	480,358	99,670	184,643	0.06	0.38	0.16
Non-X motifs	19,772,931	8,225,136	1,524,889	2,572,797	0.08	0.31	0.25

Table 9. Comparison of non-synonymous and synonymous substitutions for X motifs and non-X motifs in pairs of aligned genes in human $\mathbb{H}$ and mouse $\mathbb{M}$. $Nns(m)$, $Ns(m)$ respectively, are the potential numbers of non-synonymous, synonymous respectively, sites for the motifs $m = \left(m(\mathcal{S}, \mathbb{H})\right)$, $\mathcal{S} \in \{X, \bar{X}\}$, (Definition 13). $Ons(m)$, $Os(m)$ respectively, are the observed numbers of non-synonymous, synonymous respectively, substitutions of the motifs m (Definition 14). $Pns(m)$, $Ps(m)$ respectively, are

the percentages of non-synonymous, synonymous respectively, substitutions of the motifs m (Definition 15).

$\mathbb{C}$-$\mathbb{L}$ alignment	Nns	Ns	Ons	Os	Pns	Ps	Pns/Ps
X motifs	369,426	93,981	103,766	80,081	0.28	0.85	0.33
Non-X motifs	5,310,908	1,973,266	1,580,781	1,362,399	0.30	0.69	0.43

Table 10. Comparison of non-synonymous and synonymous substitutions for X motifs and non-X motifs in pairs of aligned genes in *S. cerevisiae* $\mathbb{C}$ and *K. lactis* $\mathbb{L}$. $Nns(m)$, $Ns(m)$ respectively, are the potential numbers of non-synonymous, synonymous respectively, sites for the motifs $m = \big(m(\mathcal{S}, \mathbb{C})\big)$, $\mathcal{S} \in \{X, \bar{X}\}$, (Definition 13). $Ons(m)$, $Os(m)$ respectively, are the observed numbers of non-synonymous, synonymous respectively, substitutions of the motifs m (Definition 14). $Pns(m)$, $Ps(m)$ respectively, are the percentages of non-synonymous, synonymous respectively, substitutions of the motifs m (Definition 15).

When we compare the rates of non-synonymous and synonymous substitutions, Pns and Ps respectively, for $\mathbb{H}$-$\mathbb{M}$ than $\mathbb{C}$-$\mathbb{L}$, the values of Pns and Ps are obviously lower in both X motifs and non-X motifs, due to the smaller phylogenetic distance between $\mathbb{H}$ and $\mathbb{M}$. In other words, $\mathbb{H}$ and $\mathbb{M}$ are more closely related than $\mathbb{C}$ and $\mathbb{L}$, so we would expect less substitutions, both synonymous and non-synonymous. Also, the values of Pns are lower than Ps for X motifs and non-X motifs in both sets of genes. Again, this is expected since non-synonymous substitutions have a larger effect on the translated protein and occur less often than synonymous substitutions.

Importantly, we observe significantly lower ratios Pns/Ps of non-synonymous to synonymous substitutions in X motifs than in non-X motifs, suggesting more evolutionary constraints on X motifs. Furthermore, while the proportion Pns of non-synonymous substitutions is lower between X motifs and non-X motifs, the proportion Ps of synonymous substitutions is higher in X motifs than in non-X motifs. This result motivated the studies presented in the next section, which were designed to analyse in more detail the specific selective constraints in X motifs.

3.1.4. Synonymous substitutions of trinucleotides in X motifs

Given a multiple alignment of orthologous gene sequences, our goal is to determine whether trinucleotides in X motifs are conserved beyond what would be expected by chance if they were evolving only under the selective pressure on the amino acid they encode. Therefore, we chose to consider only those positions in the alignment with a conserved amino acid, i.e. involving only synonymous substitutions.

We first calculated the codon substitution matrices $\mathbf{A}(m)$ (defined in Section 2.6) for the X motifs in all the mammal and yeast gene multiple alignments. These two matrices $\mathbf{A}(m)$ are shown in Appendix (Table 19 and Table 20). We then normalized the columns of $\mathbf{A}(m)$ to produce the two normalized codon substitution matrices $\mathbf{B}(m)$ (defined in Section 2.6) for the X motifs, and extracted the rows and columns of the matrices $\mathbf{B}(m)$ that correspond to the synonymous substitutions of the X codons, as shown in Appendix (Table 21 and Table 22). We also calculated the equivalent submatrices $\mathbf{B}(m)$ for the R random motifs. Finally, we calculated the percentages $Paac(m(X, \mathbb{R}), p)$ (Definition 16) of conservation

of X codons per amino acid $p \in \mathcal{X}$ in Table 21 and Table 22, for the two mammal and yeast gene multiple alignments, as summarized in Table 11 and Table 12. In addition, we provide in these Table 11 and Table 12, the mean percentages $\bar{P}aac(m(X,\mathbb{R}),\mathcal{X})$ (Definition 17) of conservation of the 12 amino acids $\mathcal{X}$ (2) for mammals and yeasts. For comparison, the values of $Paac(\bar{m}(R,\mathbb{R}),p)$ (Definition 18) and $\bar{P}aac(\bar{m}(R,\mathbb{R}),\mathcal{X})$ (Definition 19) calculated for the R mean random motifs from the 100 random codes R are reported. As these $Paac$ and $\bar{P}aac$ values for R motifs are mean values, their distributions with a ±0.99 confidence interval are shown in Figure 7 and Figure 8.

	Mean	A	D	E	F	G	I	L	N	Q	T	V	Y
$m(X,\mathbb{H})$	78.1	66.1	89.4	90.3	78.9	77.3	84.9	78.3	85.6	80.7	63.5	65.7	76.1
$\bar{m}(R,\mathbb{H})$	67.6	60.2	73.0	76.8	77.7	68.1	68.0	67.1	70.5	71.3	58.1	62.6	74.6

Table 11. For the mammal gene multiple alignments with respect to the human reference genes $s_1 = \mathbb{H}$, mean percentage $\bar{P}aac(m(X,\mathbb{H}),\mathcal{X})$ (Definition 17) of conservation of X codons for the 12 amino acids $\mathcal{X}$ (2) coded by the X motifs and percentages $Paac(m(X,\mathbb{H}),p)$ (Definition 16) of conservation per amino acid $p \in \mathcal{X}$ (first row). Mean percentage $\bar{P}aac(\bar{m}(R,\mathbb{H}),\mathcal{X})$ (Definition 19) of conservation of X codons for the 12 amino acids $\mathcal{X}$ (2) coded by the R mean random motifs (from the 100 random codes R) and percentages $Paac(\bar{m}(R,\mathbb{H}),p)$ (Definition 18) of conservation per amino acid $p \in \mathcal{X}$ (second row).

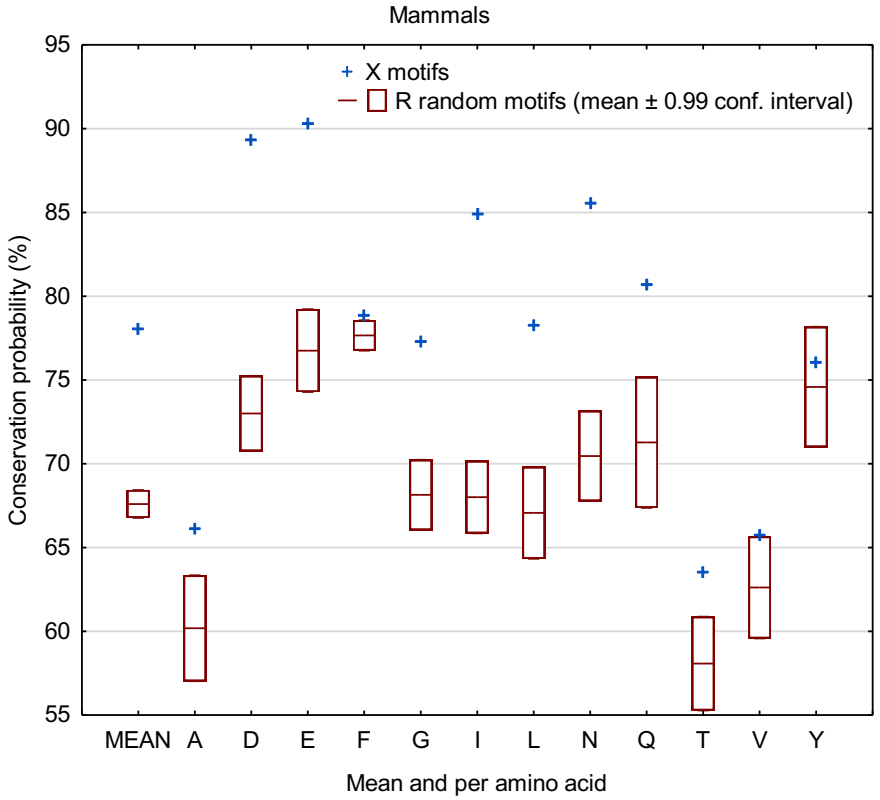

Figure 7 (associated with Table 11). For the mammal gene multiple alignments with respect to the human reference genes $s_1 = \mathbb{H}$, mean percentage $\bar{P}aac(m(X,\mathbb{H}),\mathcal{X})$ (Definition 17) of conservation of X codons for the 12 amino acids $\mathcal{X}$ (2) coded the X motifs and percentages $Paac(m(X,\mathbb{H}),p)$ (Definition 16) of conservation per amino acid $p \in \mathcal{X}$ (blue cross). The distribution of the R mean random motifs (from the 100 random codes R) is indicated by boxplots representing the mean percentage $\bar{P}aac(\bar{m}(R,\mathbb{H}),\mathcal{X})$ (Definition 19) and percentages $Paac(\bar{m}(R,\mathbb{H}),p)$ (Definition 18) with a ±0.99 confidence interval.

A new and strong property is identified with the X motifs of the circular code X. The average percentage ($\bar{P}aac$) conservation of X codons is significantly higher in X motifs than the conservation observed in the R mean random motifs in the mammal gene alignments (one-sided Student's t-test with $p \approx 10^{-55}$) (Table 11 and Figure 7). Furthermore, this is true for 11 out of 12 amino acids (percentage $Paac$). For the amino acid Y, the conservation of X codons in X motifs is higher than in R motifs although the difference is not significant at 0.99. This new property can be formalized simply by the following inequalities:

$$\begin{cases} \bar{P}aac(m(X,\mathbb{H}),\mathcal{X}) > \bar{P}aac(\bar{m}(R,\mathbb{H}),\mathcal{X}) \\ Paac(m(X,\mathbb{H}),p) > Paac(\bar{m}(R,\mathbb{H}),p) \quad \forall p \in \mathcal{X} \end{cases}.$$

	Mean	A	D	E	F	G	I	L	N	Q	T	V	Y
$m(X,\mathbb{C})$	29.3	16.1	45.3	41.8	29.9	40.4	39.3	10.6	33.5	13.6	14.9	33.6	33.0
$\bar{m}(R,\mathbb{C})$	22.3	19.1	26.9	25.2	29.8	27.1	20.3	21.8	22.3	20.9	15.6	18.2	33.0

Table 12. For the yeast gene multiple alignments with respect to the *S. cerevisiae* reference genes $s_1 = \mathbb{C}$, mean percentage $\bar{P}aac(m(X,\mathbb{C}),\mathcal{X})$ (Definition 17) of conservation of X codons for the 12 amino acids $\mathcal{X}$ (2) coded by the X motifs and percentages $Paac(m(X,\mathbb{C}),p)$ (Definition 16) of conservation per amino acid $p \in \mathcal{X}$ (first row). Mean percentage $\bar{P}aac(\bar{m}(R,\mathbb{C}),\mathcal{X})$ (Definition 19) of conservation of X codons for the 12 amino acids $\mathcal{X}$ (2) coded by the R mean random motifs (from the 100 random codes R) and percentages $Paac(\bar{m}(R,\mathbb{C}),p)$ (Definition 18) of conservation per amino acid $p \in \mathcal{X}$ (second row).

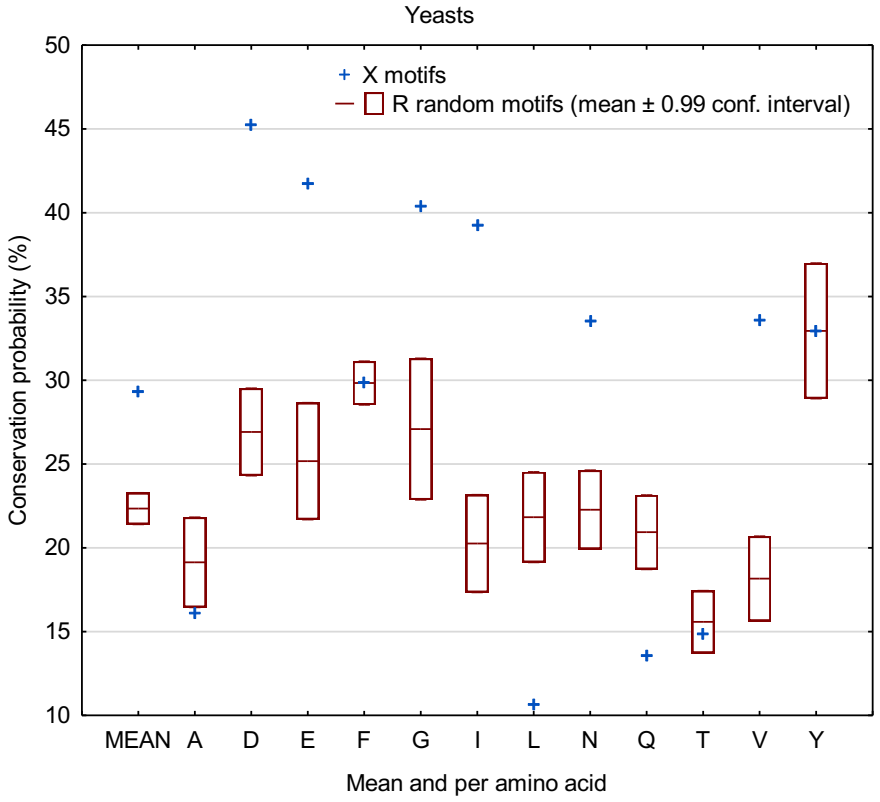

Figure 8 (associated with Table 12). For the yeast gene multiple alignments with respect to the *S. cerevisiae* reference genes $s_1 = \mathbb{C}$, mean percentage $\bar{P}aac(m(X,\mathbb{C}),\mathcal{X})$ (Definition 17) of conservation of X codons for the 12 amino acids $\mathcal{X}$ (2) coded by the X motifs and percentages $Paac(m(X,\mathbb{C}),p)$ (Definition 16) of conservation per amino acid $p \in \mathcal{X}$ (blue cross). The distribution of the R mean random motifs (from the 100 random codes R) is indicated by boxplots representing the mean

percentage $\bar{P}aac(\bar{m}(R,\mathbb{C}),\mathcal{X})$ (Definition 19) and percentages $Paac(\bar{m}(R,\mathbb{C}),p)$ (Definition 18) with a ±0.99 confidence interval.

The average percentage ($\bar{P}aac$) conservation of X codons is significantly higher in X motifs than in the R mean random motifs in the yeast gene alignments (one-sided Student's t-test with $p \approx 10^{-35}$) (Table 12 and Figure 8). For 6 out of 12 amino acids, the conservation (percentage $Paac$) of X codons in X motifs is higher than in R motifs. In contrast, for the amino acids A, L, Q and T, the conservation is lower than in the R motifs. For the amino acids F and L, the conservation is similar to the R motifs. This property can also be summarized by the following inequalities:

$$\begin{cases} \bar{P}aac(m(X,\mathbb{C}),\mathcal{X}) > \bar{P}aac(\bar{m}(R,\mathbb{C}),\mathcal{X}) \\ Paac(m(X,\mathbb{C}),p) > Paac(\bar{m}(R,\mathbb{C}),p) \quad \forall p \in \mathcal{X}\backslash\{A,L,Q,T\} \end{cases}.$$

The conservation values of the motifs observed in the yeast alignments is lower than in the human alignments. This is not surprising since it is known that the yeasts diverged much earlier (more synonymous and non-synonymous substitutions) than the mammals included in this study. This evolutionary diversity in yeasts may also explain the exception with the four amino acids observed with this simple statistical parameter $Paac$. It should also be stressed that the identified conservation property of X codons in X motifs with respect to the amino acids is independent of the codon usage, the GC content, the nucleotide composition, the length of genes, etc.

3.1.5. A new hypothesis of evolution of the genetic code: union of circular codes associated with each amino acid

The statistical analyses performed in the previous sections show that the 20 trinucleotides of the circular code X (1) are strongly linked to the amino acids they encode, thus, leading to a partition of the 20 trinucleotides of X into 12 trinucleotide classes, each trinucleotide class being associated with an amino acid $p \in \mathcal{X}$ (2). This property leads us to propose that the extant genetic code may result from a union of circular codes: the subcodes of the circular code X (1) associated with each amino acid (Table 13 and Figure 9). Remember that a subcode of a circular code, is also circular. Interestingly, classes of circular codes with the strongest constraints of reading frame retrieval, i.e. the strong comma-free and comma-free codes (Theorem 3 and Theorem 2), can code an amino acid. Using these theorems, we determine the circular class of each trinucleotide code involved in Table 13 (an initial approach developed in Michel, 2014, Section 3.4, Table 6).

The evolution of the genetic code may thus have started from the circular codes with the strongest constraints, i.e. the strong comma-free codes and the comma-free codes with motifs retrieving the reading frame after the reading of 2 and 3 nucleotides, i.e. a nucleotide length of a codon or anticodon. It is tempting to suggest that these circular codes may have emerged independently in different "primitive soups". However, such strongly constrained coding systems may not have been viable in the long term. By relaxing the constraints, they may have evolved to circular codes having flexible motifs for retrieving the reading frame after the reading of at most 13 nucleotides, and to non-circular codes

without the ability to retrieve the reading frame. Among the 12 amino acids $\mathcal{X}$ (2) coded by the circular code X (1), 10 amino acids are coded by strong comma-free codes and 2 amino acids E_X and L_X of X, by comma-free codes (Table 13 and Figure 9). In the extant genetic code, only 3 amino acids D, N and Q are still coded by strong comma-free codes, 6 amino acids A, E, I, T, V and Y, by comma-free codes, 1 amino acid L, by a circular code, and 2 amino acids F and G, by simple codes (not circular). The union of circular codes allows to extend the amino acid coding. For example, the union of the strong comma-free code $Q_X = \{CAG\}$ of X and the strong comma-free code $\{CAA\}$ leads to the strong comma-free code $Q = \{CAA, CAG\}$ of the genetic code, etc. Obviously, the union of 2 comma-free codes does not imply that the resulting code is comma-free, see for example the case of the amino acid L (Table 13). The 8 remaining amino acids could have been generated by mutations in circular codes. The extant genetic code is a code from a mathematical point of view (Definition 3), however it is not circular, i.e. it does not have the ability to retrieve the reading frame in genes after its circularity property loss.

AA	Circular code X	Class	Union	Class	Genetic code	Class
Asn	$N_X = \{AAC, AAT\}$	*SCF*			$N = \{AAC, AAT\}$	*SCF*
Asp	$D_X = \{GAC, GAT\}$	*SCF*			$D = \{GAC, GAT\}$	*SCF*
Gln	$Q_X = \{CAG\}$	*SCF*	$\{CAA\}$	*SCF*	$Q = \{CAA, CAG\}$	*SCF*
Glu	$E_X = \{GAA, GAG\}$	*CF*			$E = \{GAA, GAG\}$	*CF*
Phe	$F_X = \{TTC\}$	*SCF*	$\{TTT\}$	*NC*	$F = \{TTC, TTT\}$	*NC*
Tyr	$Y_X = \{TAC\}$	*SCF*	$\{TAT\}$	*CF*	$Y = \{TAC, TAT\}$	*CF*
Ile	$I_X = \{ATC, ATT\}$	*SCF*	$\{ATA\}$	*CF*	$I = \{ATA, ATC, ATT\}$	*CF*
Ala	$A_X = \{GCC\}$	*SCF*	$\{GCA, GCG, GCT\}$	*CF*	$A = \{GCA, GCC, GCG, GCT\}$	*CF*
Gly	$G_X = \{GGC, GGT\}$	*SCF*	$\{GGA, GGG\}$	*NC*	$G = \{GGA, GGC, GGG, GGT\}$	*NC*
Thr	$T_X = \{ACC\}$	*SCF*	$\{ACA, ACG, ACT\}$	*CF*	$T = \{ACA, ACC, ACG, ACT\}$	*CF*
Val	$V_X = \{GTA, GTC, GTT\}$	*SCF*	$\{GTG\}$	*CF*	$V = \{GTA, GTC, GTG, GTT\}$	*CF*
Leu	$L_X = \{CTC, CTG\}$	*CF*	$\{CTA, CTT, TTA, TTG\}$	*CF*	$L = \{CTA, CTC, CTG, CTT, TTA, TTG\}$	*C*

Table 13. Classes of codes (non-circular *NC*, circular *C*, comma-free *CF*, strong comma-free *SCF*) of the 12 amino acids $\mathcal{X}$ (2) with respect to the circular code X (1) and the universal genetic code.

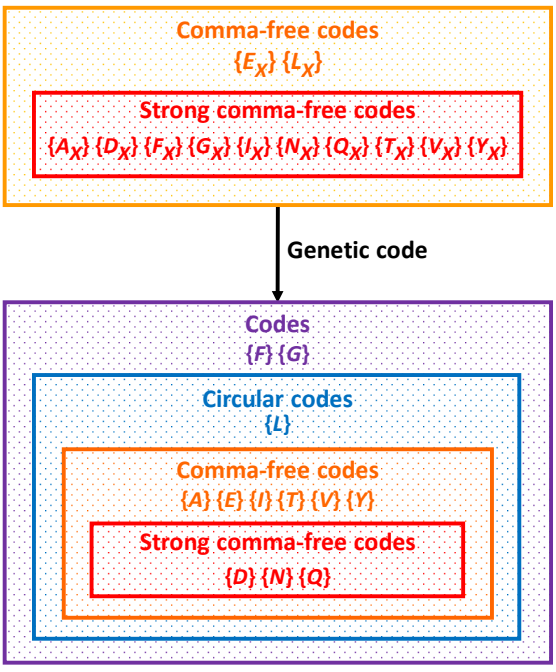

Figure 9 (associated with Table 13). Evolution of the genetic code by union of circular codes associated with each amino acid from the circular code X (1).

3.2. Functionality of *X* motifs in extant genomes

In the previous section, we identified specific evolutionary constraints, suggesting that the *X* motifs in the genomes included in this study have evolved under purifying selection. Indeed, the nucleotides in the *X* motifs display a considerable excess of synonymous substitutions compared to the non-*X* motifs. Furthermore, the average conservation of codons in *X* motifs is significantly higher than expected if the substitution process was random. These results suggest a possible functional role of *X* motifs, presumably as elements of the complex genome decoding system. In order to investigate the potential effects of *X* motifs on the translation of protein-coding genes, we compared the frequency of *X* motifs in the genes of the four mammalian and nine yeast species with existing experimental data on protein expression and protein production. We will show that the experimental data can generally be explained by circular code motifs, i.e. motifs having the property of reading frame retrieval.

3.2.1. Dicodons associated with reduced protein production are not located in *X* motifs

Recently, experimental studies in *S. cerevisiae* (Gamble *et al.*, 2016) were performed to investigate the effects of different codons on translation efficiency. The authors measured the expression levels of more than 35,000 synthetic protein variants in which three adjacent codons of the coding sequence were randomized. No individual codons had consistent effects on gene expression. However, 17 pairs of adjacent codons (called in the following dicodons) were identified that, when they were present in-frame in the coding sequence, reduced the expression level of the genes. This list is recalled in Table 14. In this list, we identified the codons belonging to the circular code *X*.

Dicodon	Class	Dicodon	Class
AGGCGA	*NN*	*CGAGCG*	*NN*
AGGCGG	*NN*	*CTCCCG*	*XN*
ATACGA	*NN*	*CTGATA*	*XN*
ATACGG	*NN*	*CTGCCG*	*XN*
CGAATA	*NN*	*CTGCGA*	*XN*
CGACCG	*NN*	*GTACCG*	*XN*
CGACGA	*NN*	*GTACGA*	*XN*
CGACGG	*NN*	*GTGCGA*	*NN*
CGACTG	*NX*		

Table 14. List of the 17 dicodons that reduced the expression level of the genes (Gamble *et al.*, 2016) (1st and 3rd columns). Class of the dicodons according to its codons belonging to the circular code *X* (symbol *X*) or not (symbol *N*) (2nd and 4th columns).

Surprisingly, none of these 17 dicodons are composed of two *X* codons meaning that they cannot be located in a *X* motif.

3.2.2. Correlation of *X* motifs with dicodons associated with low and high protein production

Following the work of Gamble *et al.* (2016), Diambre (2017) performed a statistical analysis of dicodon usage frequencies over two sets of proteins: a low protein abundance (PA) set and a high PA set, from

nine diverse organisms including three prokaryotes, one plant, one yeast (*S. cerevisiae*), and two multicellular eukaryotes and two mammals. The working hypothesis was that sequences encoding abundant proteins should be optimized, in the sense of translation efficiency. He found an important bias of dicodon usage depending on PA and determined which dicodons where statistically associated with low or high abundance. These usage preferences cannot be explained by the frequency usage of the single codons. The statistical analysis of coding sequences of nine organisms reveals that in many cases dicodon preferences are shared between related organisms.

Dicodon	Class	Dicodon	Class
AAAATA	*NN*	*CAGAAA*	*XN*
AATGCA	*XN*	*GAAAGT*	*XN*
AATTGG	*XN*	*GAACTA*	*XN*
AGTAAG	*NN*	*GCATTT*	*NN*
AGTGTG	*NN*	*TATAAA*	*NN*
ATAGGT	*NX*	*TATCCG*	*NN*
ATTAAA	*XN*	*TTTCAG*	*NX*
CAAAGT	*NN*	*TTTTTT*	*NN*

Table 15. List of the 16 dicodons with low protein abundance (Diambra, 2017) (1st and 3rd columns). Class of the dicodons according to its codons belong to the circular code *X* (symbol *X*) or not (symbol *N*) (2nd and 4th columns).

In addition to the 17 previous dicodons (identified in *S. cerevisiae*), this study identified 16 new dicodons (Table 15) associated with low protein abundance in a number of different organisms. Again, these 16 dicodons cannot be located in a *X* motif. Thus, there are 33 low abundance dicodons that support the circular code theory.

Furthermore, the study revealed 40 dicodons shared between different organisms and preferentially used by high abundance proteins (Table 16). Importantly, 27 of these 40 dicodons (67.5%) are potentially in *X* motifs.

Dicodon	Class	Dicodon	Class	Dicodon	Class	Dicodon	Class
AACAAC	*XX*	*ACCTTC*	*XX*	*GACACC*	*XX*	*GTCACC*	*XX*
AACAAG	*XN*	*ATCAAC*	*XX*	*GACTAC*	*XX*	*GTCATC*	*XX*
AACACC	*XX*	*ATCAAG*	*XN*	*GATGCT*	*XN*	*GTTGCC*	*XX*
AAGTCC	*NN*	*ATCACC*	*XX*	*GCCAAC*	*XX*	*TACAAC*	*XX*
ACCAAC	*XX*	*ATCATC*	*XX*	*GCCAAG*	*XN*	*TACAAG*	*XN*
ACCAAG	*XN*	*ATTGCC*	*XX*	*GCCACC*	*XX*	*TCCACC*	*NX*
ACCACC	*XX*	*CCACCA*	*NN*	*GCCATC*	*XX*	*TTCAAC*	*XX*
ACCATC	*XX*	*CGTCGT*	*NN*	*GCCGCC*	*XX*	*TTCAAG*	*XN*
ACCATT	*XX*	*GACAAC*	*XX*	*GGTGTC*	*XX*	*TTCACC*	*XX*
ACCGCC	*XX*	*GACAAG*	*XN*	*GTCAAG*	*XN*	*TTCATC*	*XX*

Table 16. List of the 40 dicodons with high protein abundance (Diambra, 2017) (1st, 3rd, 5th and 7th columns). Class of the dicodons according to its codons belong to the circular code *X* (symbol *X*) or not (symbol *N*) (2nd, 4th, 6th and 8th columns).

3.2.3. Classification of genes as low or high abundance according to the circular code theory

The experimental and statistical results of Gamble *et al.* (2016) and Diambre (2017) (Table 14, Table 15 and Table 16) can be summarized in the following Table 17.

	XX	$\{NN,NX,XN\}$	Total
Low abundance protein	0	33	33
High abundance protein	27	13	40
Total	27	46	73

Table 17. Contingency table of low/high abundance protein and presence/absence of dicodons XX (deduced from Table 14, Table 15 and Table 16).

A $\mathcal{X}^2$ test shows a strongly significant relation between the presence/absence of dicodons XX and protein abundancy with a one sided value $p \approx 10^{-9}$ (Table 17). The following probabilities can be easily deduced from Table 17:

$$P(\text{Low abundance protein} \mid XX) = 0/33 = 0\%,$$

$$P(\text{High abundance protein} \mid XX) = 27/40 = 67.5\%.$$

Thus, the presence-absence of XX dicodons in a gene is an important and new factor in the classification of genes as low or high abundance.

3.2.4. Presence of X motifs in wild type genes and genes optimized to increase expression

The SGDB database (Wu *et al.*, 2007) contains gene expression data for genes that have been experimentally re-engineered to increase gene expression. Generally, this is achieved by replacing codons in the wild type gene with optimal codons for the expression system (i.e. replace rare codons with the most frequently used codons in the organism). We only considered the re-engineered genes that did not involve non-synonymous changes. Thus, we analysed 42 re-engineered genes that had increased expression and 4 re-engineered genes had no significant increase in expression. We searched for X motifs and R random motifs (from the 100 random codes) in the wild type genes and the genes optimized for gene expression. Then, we calculated the mean number and the mean nucleotide length of X and R motifs per sequence (Table 18).

		Mean number of X motifs	Mean number of R motifs	Mean length of X motifs	Mean length of R motifs
42 genes with increased expression	Wild type	5.4	3.6	86.1	53.7
	Optimized gene	11.2	3.7	188.6	58.2
3 genes with no increased expression	Wild type	5.3	2.6	80.0	35.8
	Optimized gene	5.0	3.8	80.0	55.6

Table 18. Mean number and mean nucleotide length of X and R random motifs per wild type gene and per optimized gene from the SGDB database (Wu *et al.*, 2007).

For the re-engineered genes that did not present an increase expression, we observe a non-significant difference in the mean number ($5.0 - 5.3 = -0.3$, one tailed Wilcoxon test with value $p = 0.50$) and no

difference in the mean length ($80.0 - 80.0 = 0$) of X motifs between the optimized genes and the wild type genes. These differences are also not significant for R motifs ($3.8 - 2.6 = 1.2$ and $55.6 - 35.8 = 19.8$, respectively, data not shown). In contrast, for the re-engineered genes that resulted in increased expression, the optimized genes have significantly more X motifs ($11.2 - 5.4 = 5.8$, one tailed Wilcoxon test with value $p \approx 10^{-6}$) and the X motifs covered a larger proportion of the genes ($188.6 - 86.1 = 102.5$, one tailed Wilcoxon test with value $p \approx 10^{-6}$) for most genes (Figure 10). These differences are not observed with the R motifs (one tailed Wilcoxon test, p values equal to 0.24 and 0.12, respectively). Thus, this important result suggest a potential new strategy for the efficient gene optimization.

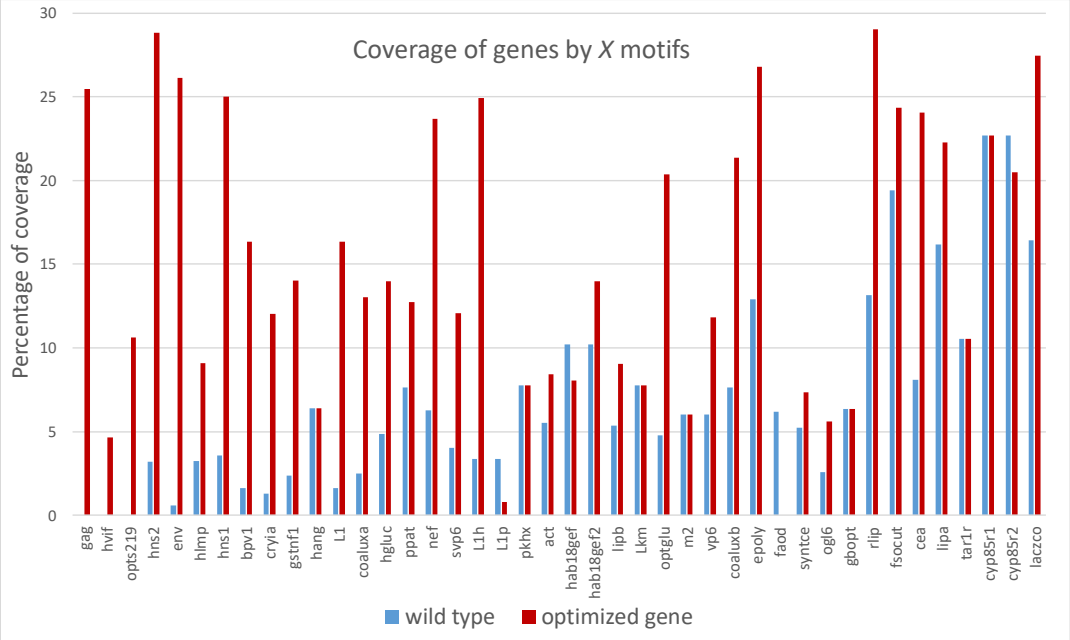

Figure 10. Percentage coverage (total length of X motifs divided by the total length of genes) of 42 wild type and optimized genes by X motifs.

4. Conclusion

The work described in this paper addressed two questions: are X motifs conserved during evolution? and do they continue to play a functional role in the processes of genome decoding and protein synthesis?

We performed a large scale study involving the complete genomes of four mammals and nine yeast species. The organisms chosen represent a large phylogenetic distribution, and a wide variety of gene structures, ranging from the simple, single exon genes of *S. cerevisiae* to the highly complex intron/exon structure of human genes. To avoid any bias towards a specific sequence alignment algorithm or evolutionary model, the multiple alignments of the gene sequences were obtained by two different methods. First, high quality mammal gene alignments were obtained from a previous independent study (Sharma and Hiller, 2017) of genome annotation methods. Second, multiple alignments of the yeast genes were constructed using a simple protein alignment method (ClustalW, Thompson *et al.*, 1994).

Furthermore, well characterized, well annotated genomes (Human and *S. cerevisiae*) were chosen to ensure high quality gene models.

In a preliminary analysis, we identified a strongly significant enrichment of X motifs (number and length) in both mammal and yeast genes, confirming our previous findings in *S. cerevisiae* (Michel *et al.*, 2017). We then calculated a number of different measures of evolutionary conservation, and showed that the X motifs are more conserved than the rest of the gene sequences, with a lower ratio Pns/Ps of non-synonymous to synonymous substitutions, indicative of purifying selection. These results were found to hold in both the mammal and yeast gene alignments. We then performed a more in-depth investigation of the synonymous substitutions in X motifs. At this stage, we modelled the evolutionary processes at the codon level, since it is important to account precisely for the protein-coding constraints on each nucleotide site. We demonstrated that the sequence conservation observed in the X motifs is the result of two types of selective pressure. The first type is the pressure to maintain the amino acids of the proteins encoded by the genes. The second type of selective pressure applies only to X motifs and highlights a new conservation property of X motifs per amino acid, which led us to propose a novel hypothesis for the evolution of the genetic code as a union of circular codes associated with each amino acid.

The increased conservation of X motifs and the specific evolutionary constraints suggest that X motifs may represent an additional, overlapping function within the protein-coding regions of genomes. Indeed, the genetic code establishes the rules to translate the 64 possible codons into the 20 amino acids and a stop signal. It is well known that the genetic code is degenerate and that different synonymous codons encoding the same amino acid are not used with the same frequency in different species. The genetic code also contains information that influences the rate and efficiency of translation, although the mechanisms of codon-mediated regulation are still not clear (Brule and Greyhack, 2017). Many recent studies have been performed to try to explain the different codon usages observed and their effects on translation. In particular, it has been shown that the efficiency of translating a particular codon is influenced by the nature of the immediately adjacent flanking codons (Gamble *et al.*, 2016; Diambra, 2017; Chevance and Hughes, 2017). These studies are mostly based on statistical and/or experimental analyses of gene sequences without being related to the results to a theoretical model. Here, we have investigated the pertinence of the circular code theory to explain the observations.

For example, in two related studies (Gamble *et al.*, 2016; Diambra, 2017), a total of 33 dicodons were found to be associated with low protein abundance, and 40 dicodons associated with high abundance proteins. We identified a significant correlation between the protein abundance level and dicodons belonging to the circular code X. To further investigate this link between the presence of X motifs in a gene and the expression level, we compared a set of re-engineered genes (that had been experimentally optimized for increased expression, using synonymous substitutions to replace rare codons with more frequent ones) with the original wild type genes. Again, we identified a significant correlation between the number and length of X motifs and protein expression levels. These results, taken together, suggest

that increasing the proportion of X motifs in a gene may represent an important new strategy for their efficient optimization.

The molecular mechanisms underlying the functional correlations observed here remain to be elucidated. However, it has been observed previously that short X motifs have also been conserved in many transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) (Michel, 2012, 2013; El Soufi and Michel, 2014, 2015). In particular, the universally conserved nucleotides A1492, A1493 G530 in the ribosome decoding center are located in short X motifs. Given the self-complementary property of the circular code X, it is possible that there is some kind of interaction between the X motifs in the protein-coding genes and the ribosomal X motifs in order to maintain the correct reading frame during translation.

The results presented here indicate that the circular code motifs may explain how the choice of different synonymous codons within genes and between species, known as codon usage bias, impacts nucleic acid stability, protein levels, structure, function and evolution. Further investigation, and *in vitro* or *in vivo* experimental validation, will be required to refine our hypothesis for the evolution of the genetic code and the proposed functional role in the translation of genes.

REFERENCES

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 125, 3389-3402.

Arquès D.G., Michel C.J. (1996). A complementary circular code in the protein coding genes. *Journal of Theoretical Biol*ogy 182, 45-58.

Brule C.E., Grayhack E.J. (2017). Synonymous codons: choose wisely for expression. *Trends in Genetics* 33, 283-297.

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 U.S.A.*, 1614896114.

Diambra L.A. (2017). Differential bicodon usage in lowly and highly abundant proteins. *PeerJ* 5:e3081.

El Houmami N., Seligmann H. (2017). Evolution of nucleotide punctuation marks: from structural to linear signals. *Frontiers in Genetics* 8, 36.

El Soufi K., Michel C.J. (2014). Circular code motifs in the ribosome decoding center. *Computational Biology and Chemistry* 52, 9-17.

El Soufi K., Michel C.J. (2015). Circular code motifs near the ribosome decoding center. *Computational Biology and Chemistry* 59, 158-176.

El Soufi K., Michel C.J. (2017). Unitary circular code motifs in genomes of eukaryotes. *Biosystems* 153, 45-62.

Fimmel E., Michel C.J., Strüngmann L. (2016). *n*-Nucleotide circular codes in graph theory. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 20150058.

Fimmel E., Michel C.J., Strüngmann L. (2017). Strong comma-free codes in genetic information. *Bulletin of Mathematical Biology* 79, 1796-1819.

Fimmel E., Strüngmann L. (2018). Mathematical fundamentals for the noise immunity of the genetic code. *Biosystems* 164, 186-198.

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.

Michel C.J. (2008). A 2006 review of circular codes in genes. *Computer and Mathematics with Applications* 55, 984-988.

Michel C.J. (2012). Circular code motifs in transfer and 16S ribosomal RNAs: A possible translation code in genes. *Computational Biology and Chemistry* 37, 24-37.

Michel C.J. (2013). Circular code motifs in transfer RNAs. *Computational Biology and Chemistry* 45, 17-29.

Michel C.J. (2014). A genetic scale of reading frame coding. *Journal of Theoretical Biology* 355, 83-94.

Michel C.J. (2015). The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. *Journal of Theoretical Biology* 380, 156-177.

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, 1-16.

Michel C.J., Nguefack Ngoune V., 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.

Nei M., Gojobori T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology Evolution* 3, 418-26.

Seligmann H., Pollock DD. (2004). The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA and Cell Biology* 23, 701-705.

Seligmann H. (2011). Error compensation of tRNA misacylation by codon-anticodon mismatch prevents translational amino acid misinsertion. *Computational Biology and Chemistry* 35, 81-95.

Seligmann H., Warthi G. (2017). Genetic Code Optimization for Cotranslational Protein Folding: Codon Directional Asymmetry Correlates with Antiparallel Betasheets, tRNA Synthetase Classes. *Computational and Structural Biotechnology Journal* 15, 412-424.

Sharma V., Hiller M. (2017). Increased alignment sensitivity improves the usage of genome alignments for comparative gene annotation. *Nucleic Acids Research* 45, 8369-8377.

Thompson J.D., Higgins D.G., Gibson T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*.22, 4673-4680.

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 Research* 35, D76-D79.

APPENDIX 1:

		N	N	T	I	I	Q	L	L	E	D	E	D	A	G	G	V	V	V	Y	F
		AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC
K	AAA	521	569	81	33	22	378	11	23	1449	126	494	125	21	39	21	35	10	7	14	3
N	AAC	7493	420002	767	186	53	92	36	10	105	1728	200	569	199	471	129	3	47	16	178	19
K	AAG	650	744	109	31	37	2148	15	96	423	130	2316	133	45	40	42	23	9	9	14	3
N	AAT	1636	556175	200	40	194	131	8	7	147	477	221	1539	58	112	275	5	8	46	53	5
T	ACA	110	103	5453	143	139	61	11	124	147	16	112	15	387	20	14	264	63	56	3	17
T	ACC	905	268	67464	1434	369	22	126	50	26	149	45	37	3797	184	44	27	516	103	24	79
T	ACG	57	64	5203	126	99	61	8	120	22	9	98	14	292	24	10	59	42	31	1	5
T	ACT	204	698	12412	333	1202	23	43	22	30	48	27	94	943	60	100	28	143	327	8	23
R	AGA	108	101	33	22	13	169	3	20	244	33	108	38	15	94	58	47	13	5	3	2
S	AGC	4025	1342	1741	344	112	64	47	8	57	460	68	151	412	3307	669	5	119	38	61	30
R	AGG	104	123	80	17	22	602	17	86	69	15	297	25	19	113	99	16	12	2	5	2
S	AGT	1020	3570	328	72	297	49	9	10	49	136	69	392	70	724	1386	7	35	54	18	4
I	ATA	16	16	93	2981	2224	23	79	339	53	4	35	5	36	6	3	1421	236	137	4	28
I	ATC	236	45	1429	9570	819908	5	1531	226	11	47	8	10	413	90	16	252	5416	909	32	522
M	ATG	75	58	370	854	862	120	178	2474	28	13	164	11	132	13	12	302	223	212	3	54
I	ATT	63	167	286	1636	654789	8	241	149	12	9	9	32	96	32	42	140	869	2577	7	94
Q	CAA	95	61	9	3	4	14988	56	182	909	71	339	43	12	18	15	13	2	1	46	28
H	CAC	798	320	52	24	4	1695	360	79	55	412	113	116	51	164	31	2	18	5	1241	82
Q	CAG	130	131	24	8	7	148989	119	1289	566	136	2483	110	32	43	25	13	9	15	102	12
H	CAT	226	653	17	9	24	1223	92	49	49	114	93	279	21	34	77	3	4	20	331	30
P	CCA	5	10	59	10	9	672	122	1091	65	4	33	6	119	7	7	31	8	13	3	17
P	CCC	72	38	740	52	29	222	685	289	11	30	15	11	788	49	18	9	88	18	59	108
P	CCG	4	4	48	7	3	721	90	1000	16	3	66	6	68	10	3	7	3	4	3	15
P	CCT	17	47	233	21	89	157	246	284	12	12	14	48	279	14	19	7	17	49	22	40
R	CGA	15	8	1	1	1	608	21	69	63	10	30	3	0	30	9	9	1	1	8	5
R	CGC	72	37	27	7	3	413	283	62	8	65	21	7	36	279	38	0	9	3	90	45
R	CGG	17	13	3	2	4	3737	48	427	22	7	137	6	5	37	20	5	20	9	12	4
R	CGT	25	99	3	7	11	240	58	29	3	14	11	31	8	57	77	1	3	11	29	12
L	CTA	10	3	10	72	49	130	2150	13178	16	0	6	0	15	2	2	227	33	23	3	97
L	CTC	42	15	126	1521	351	89	74979	10241	22	17	18	14	158	33	10	41	887	155	74	1850
L	CTG	12	19	28	340	242	1221	11984	163737	14	5	101	4	59	9	7	187	181	116	14	403
L	CTT	11	28	28	268	913	79	10926	6044	3	8	7	19	32	11	17	22	140	427	20	297
E	GAA	177	163	35	6	4	579	13	22	80704	1875	25592	1845	91	118	96	154	24	39	27	10
D	GAC	2014	716	147	56	30	170	17	6	2386	90508	3351	23213	659	1402	334	16	181	52	173	15
E	GAG	244	226	30	16	15	2591	21	86	27179	3295	152612	3028	158	207	174	57	52	46	26	11
D	GAT	605	1944	61	15	44	94	8	7	1769	20719	2606	70821	141	396	893	17	46	147	47	2
A	GCA	24	32	265	60	35	52	15	145	508	83	395	59	5329	69	74	696	127	128	2	13
A	GCC	170	75	3897	416	127	27	131	68	128	621	225	177	86827	1100	199	132	1969	392	25	111
A	GCG	11	13	248	49	28	91	11	149	145	44	485	34	4952	88	49	145	135	105	2	10
A	GCT	62	135	912	86	303	16	27	39	62	150	126	365	17318	280	437	106	432	1191	5	22
G	GGA	19	43	17	7	8	71	8	15	1247	125	434	102	70	4063	2444	126	25	19	2	3
G	GGC	452	183	195	91	27	47	31	14	110	1304	197	389	940	70299	10181	19	321	69	47	22
G	GGG	39	66	30	11	10	239	6	97	334	115	1478	123	119	4951	3298	75	40	38	10	4
G	GGT	91	343	44	27	59	24	8	1	91	330	125	862	200	10850	24402	16	61	199	17	4
V	GTA	1	5	28	212	145	12	27	261	142	9	69	11	128	12	4	12620	1378	974	1	31
V	GTC	50	26	577	6081	1244	19	937	173	34	153	57	51	2028	234	50	1824	47252	6691	29	481
V	GTG	17	13	171	989	767	139	188	2165	102	23	498	26	562	57	49	8757	6074	4909	2	71
V	GTT	18	49	113	869	3036	7	127	96	20	43	30	138	437	86	195	1159	5514	24443	10	71
*	TAA	3	2	1	0	0	29	1	4	43	2	15	2	1	0	0	2	0	0	41	2
Y	TAC	161	63	21	34	6	86	82	21	16	170	35	55	29	39	11	2	22	4	68411	941
*	TAG	2	5	0	1	1	207	3	35	12	1	89	1	0	1	0	0	0	0	28	1
Y	TAT	55	209	13	7	26	76	26	17	16	73	26	169	5	3	32	0	11	10	16017	273
S	TCA	16	10	134	4	5	74	23	203	44	3	37	5	152	6	3	54	15	8	15	75
S	TCC	108	26	1174	57	25	30	187	52	13	51	27	23	1690	51	7	8	108	26	261	725
S	TCG	7	12	103	12	8	71	10	151	10	4	39	5	138	7	4	9	6	2	9	51
S	TCT	20	83	325	23	67	28	49	66	6	14	20	51	470	37	54	11	32	57	103	258
*	TGA	1	1	0	0	0	27	1	12	14	3	10	1	1	12	4	4	1	0	3	3
C	TGC	55	21	51	25	3	36	83	20	9	46	9	19	54	338	56	1	26	5	566	269
W	TGG	3	6	5	2	8	431	11	197	9	6	46	3	1	28	6	4	2	1	25	45
C	TGT	35	72	20	12	24	55	20	16	13	12	29	33	9	109	195	4	5	11	185	93
L	TTA	0	2	10	43	39	20	291	1886	16	2	10	1	5	1	0	204	19	28	15	382
F	TTC	18	10	72	505	105	13	1918	278	10	25	4	6	114	37	2	26	397	83	1042	92119
L	TTG	7	1	23	75	80	130	918	15979	14	2	43	0	35	7	6	97	59	41	17	591
F	TTT	15	22	23	86	350	15	366	234	4	10	7	12	29	8	8	18	80	197	287	16156

Table 19. For the mammal gene multiple alignments with respect to the human reference genes $s_1 = \mathbb{H}$, codon substitution matrix $\mathbf{A}(m(X, \mathbb{H}))$ of X motifs $m(X, \mathbb{H})$ (Section 2.6). For each codon, the encoded amino acid is given.

		N	N	T	I	I	Q	L	L	E	D	E	D	A	G	G	V	V	V	Y	F
		AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC
K	AAA	1235	1785	475	304	592	769	107	187	2233	890	1083	1517	357	284	432	250	221	409	277	187
N	AAC	985	210627	507	238	394	477	82	136	1752	1456	764	2375	319	418	805	209	193	346	300	245
K	AAG	1512	2008	612	364	595	959	117	227	2653	991	1280	1665	467	320	551	289	289	579	298	243
N	AAT	7143	10014	516	288	532	490	90	150	1931	1473	878	2699	319	453	821	251	186	338	304	214
T	ACA	574	900	1923	320	462	215	70	105	771	344	420	638	278	141	251	267	283	451	130	178
T	ACC	622	880	3844	405	545	240	81	136	889	418	402	690	345	131	318	301	405	605	186	163
T	ACG	281	471	929	175	263	119	36	67	365	201	190	307	132	55	119	142	115	237	77	83
T	ACT	796	1084	3676	392	657	255	90	149	1055	454	475	805	381	170	353	351	402	661	207	171
R	AGA	491	677	171	140	225	276	50	94	622	233	334	471	119	114	211	123	116	172	155	115
S	AGC	581	823	274	118	185	160	42	62	699	391	338	649	187	179	278	120	95	180	103	104
R	AGG	174	218	61	60	77	129	21	26	255	95	120	184	38	37	60	26	40	61	47	44
S	AGT	698	1041	405	149	216	208	54	84	816	509	379	818	305	247	432	126	127	261	119	112
I	ATA	157	287	150	2179	3768	97	227	406	305	125	152	229	105	55	101	684	665	1091	137	284
I	ATC	287	331	292	7454	10862	131	350	601	398	192	221	327	214	75	127	1072	1659	2450	248	538
M	ATG	371	552	256	908	1677	214	357	776	590	199	245	421	201	100	161	408	402	751	275	644
I	ATT	356	516	391	8368	15453	194	471	869	640	222	316	429	260	109	190	1608	2023	3469	343	715
Q	CAA	1011	1519	380	228	459	3493	94	183	2517	752	1100	1405	330	220	413	210	177	359	223	171
H	CAC	535	664	125	89	177	185	43	50	472	241	247	414	95	85	120	72	82	110	411	188
Q	CAG	793	1139	284	227	357	2678	87	159	1835	561	894	1012	224	174	268	202	121	256	185	154
H	CAT	640	928	167	116	227	270	60	73	622	323	309	583	92	127	181	118	96	143	454	225
P	CCA	416	637	216	156	290	185	52	73	767	388	330	672	261	115	271	127	131	239	103	109
P	CCC	253	383	133	89	174	94	34	62	484	239	242	382	179	76	165	92	87	172	65	67
P	CCG	127	252	60	38	91	58	14	34	238	134	109	230	75	40	82	53	52	73	47	39
P	CCT	381	598	190	151	268	159	59	66	737	364	365	677	264	107	245	126	124	245	122	118
R	CGA	193	292	69	63	106	157	30	43	258	134	156	204	53	54	87	46	47	85	68	51
R	CGC	113	166	36	32	76	77	21	23	154	95	90	127	39	30	46	24	23	43	35	22
R	CGG	99	136	46	39	53	64	3	19	120	66	72	110	23	22	50	22	19	38	42	18
R	CGT	211	299	75	55	106	127	22	39	319	160	133	250	51	54	102	47	42	82	68	43
L	CTA	105	147	58	353	595	68	391	797	174	74	89	132	65	23	45	181	170	279	98	241
L	CTC	127	187	128	734	1021	89	573	1274	264	98	140	163	97	54	92	259	294	508	177	451
L	CTG	231	368	135	555	883	132	431	1076	390	177	190	298	141	83	111	253	230	410	144	379
L	CTT	230	338	166	959	1513	118	808	1749	368	166	182	270	150	76	105	397	418	709	248	572
E	GAA	1790	2638	701	450	770	1160	137	241	23870	3467	9443	6507	673	493	810	429	404	652	352	246
D	GAC	1796	2674	382	226	343	460	64	103	4714	9211	2002	14480	308	431	701	166	147	288	210	145
E	GAG	1246	1802	459	316	550	861	103	174	15702	2460	6558	4270	451	314	554	285	256	493	250	219
D	GAT	2463	3664	533	306	438	616	119	151	6351	11838	2842	21898	413	569	929	267	237	413	275	228
A	GCA	481	651	364	284	445	207	69	114	879	348	426	707	2571	281	519	321	318	536	131	169
A	GCC	528	719	447	285	482	216	88	158	937	393	447	627	4330	303	796	319	408	610	166	166
A	GCG	180	283	116	114	193	103	38	61	333	135	167	267	946	114	233	142	112	199	53	79
A	GCT	681	986	584	431	711	305	101	187	1339	590	621	926	5124	437	979	463	486	851	202	229
G	GGA	654	980	189	110	226	160	47	52	717	499	401	905	350	3102	8881	96	114	154	134	116
G	GGC	490	745	149	67	128	112	40	47	505	370	291	655	219	1922	4921	68	66	104	73	89
G	GGG	194	318	67	47	59	43	12	21	216	161	109	262	104	829	2096	28	27	51	29	37
G	GGT	807	1045	212	134	201	146	57	54	769	548	371	889	517	4173	21358	83	127	209	116	143
V	GTA	166	238	193	808	1464	86	113	221	357	115	174	235	166	43	94	1160	1414	2377	119	214
V	GTC	237	340	360	1648	2406	104	146	286	438	168	220	253	356	77	133	1897	3774	5672	158	304
V	GTG	276	344	337	1577	2623	165	181	346	528	187	231	320	337	104	171	2142	2734	4596	199	337
V	GTT	334	517	505	2487	4180	177	237	431	771	264	321	494	500	127	199	3025	5325	9415	260	481
*	TAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	TAC	338	479	166	270	444	151	121	179	410	179	193	358	90	72	121	141	154	303	10095	1871
*	TAG	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Y	TAT	326	453	144	267	451	119	116	172	420	204	212	374	83	67	130	159	161	252	6896	1430
S	TCA	654	1071	451	173	272	254	61	96	918	488	428	883	410	217	454	167	133	287	151	149
S	TCC	635	951	558	149	249	201	46	74	849	470	402	760	558	228	424	165	166	248	170	125
S	TCG	547	828	418	136	206	223	42	88	714	406	397	683	351	187	365	176	106	209	119	105
S	TCT	981	1513	725	239	414	322	64	118	1430	740	636	1272	724	312	620	240	203	414	223	213
*	TGA	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
C	TGC	73	85	79	98	170	22	20	37	73	37	29	44	141	31	69	120	115	173	63	75
W	TGG	85	125	50	86	164	24	52	71	115	51	62	80	49	23	47	60	45	108	473	481
C	TGT	83	157	151	176	230	36	33	79	85	63	43	82	248	51	86	189	247	343	69	141
L	TTA	194	318	154	855	1465	112	894	1970	332	167	181	252	146	52	110	411	374	694	197	575
F	TTC	219	340	133	522	759	102	242	407	271	136	146	266	131	79	107	246	282	484	1598	11263
L	TTG	384	587	316	1793	2826	231	1667	4327	615	266	279	436	264	101	190	745	740	1265	426	1171
F	TTT	299	453	162	579	1064	130	301	505	397	200	188	336	149	100	155	306	272	557	1725	10020

Table 20. For the yeast gene multiple alignments with respect to the *S. cerevisiae* reference genes $s_1 = \mathbb{C}$, codon substitution matrix $\mathbf{A}(m(X, \mathbb{C}))$ of X motifs $m(X, \mathbb{C})$ (Section 2.6). For each codon, the encoded amino acid is given.

	A			
	GCA	GCC	GCG	GCT
GCA				
GCC		66.1		
GCG				
GCT				
Sum		66.1		
Mean		66.1		

	D		
	GAC	GAT	
GAC	72.9	22.0	
GAT	16.7	67.1	
Sum	89.6	89.1	
Mean	89.4		

	E		
	GAA	GAG	
GAA	67.3	13.0	
GAG	22.7	77.7	
Sum	90.0	90.7	
Mean	90.3		

	F		
	TTC	TTT	
TTC	78.9		
TTT			
Sum	78.9		
Mean	78.9		

	G			
	GGA	GGC	GGG	GGT
GGA				
GGC		69.6		21.9
GGG				
GGT		10.7		52.4
Sum		80.3		74.3
Mean		77.3		

	I		
	ATA	ATC	ATT
ATA			
ATC	73.1	22.4	
ATT	12.5	61.8	
Sum	85.6	84.2	
Mean	84.9		

	L					
	CTA	CTC	CTG	CTT	TTA	TTG
CTA						
CTC		68.1	4.6			
CTG		10.9	73.0			
CTT						
TTA						
TTG						
Sum		79.0	77.6			
Mean		78.3				

	N		
	AAC	AAT	
AAC	71.1	22.2	
AAT	15.5	62.3	
Sum	86.6	84.5	
Mean	85.6		

	Q		
	CAA	CAG	
CAA			
CAG		80.7	
Sum		80.7	
Mean		80.7	

	T			
	ACA	ACC	ACG	ACT
ACA				
ACC		63.5		
ACG				
ACT				
Sum		63.5		
Mean		63.5		

	V			
	GTA	GTC	GTG	GTT
GTA	42.7	1.9		2.1
GTC	6.2	64.2		14.8
GTG				
GTT	3.9	7.5		53.9
Sum	52.8	73.6		70.9
Mean		65.7		

	Y	
	TAC	TAT
TAC	76.1	
TAT		
Sum	76.1	
Mean	76.1	

Table 21. For the mammal gene multiple alignments with respect to the human reference genes $s_1 = \mathbb{H}$, codon substitution submatrices of $\mathbf{B}(m(X, \mathbb{H}))$ (in %) of X motifs $m(X, \mathbb{H})$ (Section 2.6) for the 12 amino acids $p \in \mathcal{X} = \{A, D, E, F, G, I, L, N, Q, T, V, Y\}$ coded by the circular code X (1).

	A			
	GCA	*GCC*	*GCG*	*GCT*
GCA				
GCC		16.1		
GCG				
GCT				
Sum		16.1		
Paac		16.1		

	D		
	GAC	*GAT*	
GAC	19.9	17.9	
GAT	25.6	27.1	
Sum	45.4	45.1	
Paac	45.3		

	E		
	GAA	*GAG*	
GAA	26.3	23.6	
GAG	17.3	16.4	
Sum	43.6	39.9	
Mean	41.8		

	F		
	TTC	*TTT*	
TTC	29.9		
TTT			
Sum	29.9		
Mean	29.9		

	G			
	GGA	*GGC*	*GGG*	*GGT*
GGA				
GGC		10.1		9.1
GGG				
GGT	21.9		39.7	
Sum	32.0		48.8	
Mean	40.4			

	I		
	ATA	*ATC*	*ATT*
ATA			
ATC		18.5	16.3
ATT		20.7	23.1
Sum		39.2	39.4
Mean		39.3	

	L					
	CTA	*CTC*	*CTG*	*CTT*	*TTA*	*TTG*
CTA						
CTC		5.6	6.2			
CTG		4.2	5.2			
CTT						
TTA						
TTG						
Sum		9.9	11.4			
Mean			10.6			

	N		
	AAC	*AAT*	
AAC	20.6	16.2	
AAT	15.0	15.3	
Sum	35.6	31.5	
Mean	33.5		

	Q		
	CAA	*CAG*	
CAA			
CAG		13.6	
Sum		13.6	
Mean		13.6	

	T			
	ACA	*ACC*	*ACG*	*ACT*
ACA				
ACC		14.9		
ACG				
ACT				
Sum		14.9		
Mean		14.9		

	V			
	GTA	*GTC*	*GTG*	*GTT*
GTA	5.2	5.0		5.0
GTC	8.4	13.3		11.9
GTG				
GTT	13.5	18.8		19.7
Sum	27.1	37.2		36.6
Mean		33.6		

	Y	
	TAC	*TAT*
TAC	33.0	
TAT		
Sum	33.0	
Mean	33.0	

Table 22. For the yeast gene multiple alignments with respect to the *S. cerevisiae* reference genes $s_1 = \mathbb{C}$, codon substitution submatrices of $\mathbf{B}(m(X,\mathbb{C}))$ (in %) of X motifs $m(X,\mathbb{C})$ (Section 2.6) for the 12 amino acids $p \in \mathcal{X} = \{A, D, E, F, G, I, L, N, Q, T, V, Y\}$ coded by the circular code X (1).