Journal of Theoretical Biology 408 (2016) 198–212

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Circular code motifs in genomes of eukaryotes

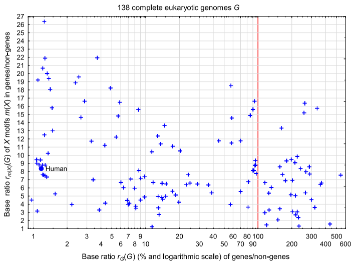
Karim El Soufi, Christian J. Michel *

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

HIGHLIGHTS

- Large circular code motifs in genomes of eukaryotes.
- Ratio of circular code motifs in genes and non-gene regions about 8.
- Circular code information in non-gene regions for translation.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 February 2016
Received in revised form
16 June 2016
Accepted 16 July 2016
Available online 19 July 2016

Keywords:
Circular code motifs
Circular code
Bijective transformation circular code
Permuted circular code
Genomes of eukaryotes

ABSTRACT

A set X of 20 trinucleotides was identified in genes of bacteria, eukaryotes, plasmids and viruses, which has in average the highest occurrence in reading frame compared to its two shifted frames (Michel, 2015; Arquès and Michel, 1996). This set X has an interesting mathematical property as X is a circular code (Arquès and Michel, 1996). Thus, the motifs from this circular code X, called X motifs, have the property to always retrieve, synchronize and maintain the reading frame in genes. In this paper, we develop several statistical analyzes of X motifs in 138 available complete genomes of eukaryotes in which genes as well as non-gene regions are examined. Large X motifs (with lengths of at least 15 consecutive trinucleotides of X and compositions of at least 10 different trinucleotides of X among 20) have the highest occurrence in genomes of eukaryotes compared to its 23 large bijective motifs, its two large permuted motifs and large random motifs. The largest X motifs identified in eukaryotic genomes are presented, e.g. an X motif in a non-gene region of the genome *Solanum pennellii* with a length of 155 trinucleotides (465 nucleotides) and an expectation $\mathbb{E} = 10^{-71}$. In the human genome, the largest X motif occurs in a non-gene region of the chromosome 13 with a length of 36 trinucleotides and an expectation $\mathbb{E} = 10^{-11}$. X motifs in non-gene regions of genomes could be evolutionary relics of primitive genes using the circular code for translation. However, the proportion of X motifs (with lengths of at least 10 consecutive trinucleotides of X and compositions of at least 5 different trinucleotides of X among 20) in genes/non-genes of the 138 complete eukaryotic genomes is about 8. Thus, the X motifs occur preferentially in genes, as expected from the previous works of 20 years.

1. Introduction

In 1996, a statistical analysis of occurrence frequencies of the 64 trinucleotides $\{AAA, \ldots, TTT\}$ in the three frames of genes of prokaryotes and eukaryotes showed that the trinucleotides are not uniformly distributed in these three frames (Arquès and Michel, 1996). By excluding the four periodic trinucleotides $\{AAA, CCC, GGG, TTT\}$ and by assigning each trinucleotide to a preferential frame (frame of its highest occurrence frequency), three subsets $X = X_0$, X_1 and X_2 of 20 trinucleotides each are found in the frames 0 (reading frame), 1 (frame 0 shifted by one nucleotide in the 5′ direction, i.e. to the right) and 2 (frame 0 shifted

* Corresponding author.
E-mail addresses: kelsoufi@unistra.fr (K. El Soufi), c.michel@unistra.fr (C.J. Michel).

http://dx.doi.org/10.1016/j.jtbi.2016.07.022
0022-5193/© 2016 Elsevier Ltd. All rights reserved.

by two nucleotides in the 5′ direction) in genes of both prokaryotes and eukaryotes. This set X contains the 20 following trinucleotides (Arquès and Michel, 1996):

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

These 20 trinucleotides of X are overrepresented in the reading frame of genes, as compared to their frequencies in the two shifted frames (Arquès and Michel, 1996). The two sets X_1 and X_2 can be deduced from X by a circular permutation (see below). These three trinucleotide sets present several strong mathematical properties, particularly the fact that X is a maximal C^3 self-complementary trinucleotide circular code (Arquès and Michel, 1996).

In 2012, in addition to the circular code X in genes (DNA and mRNA), a second major step of this circular code theory is revealed by the identification of X motifs, i.e. motifs from the circular code X, in tRNAs of prokaryotes and eukaryotes (Michel, 2012, 2013) and rRNAs of prokaryotes (16S) and eukaryotes (18S), in particular in the ribosome decoding center where the universally conserved nucleotides G530, A1492 and A1493 are included in X motifs (Michel, 2012; El Soufi and Michel, 2014, 2015). A 3D visualization of X motifs in the ribosome shows several spatial configurations involving mRNA X motifs, tRNA X motifs and rRNA X motifs (Michel, 2012; El Soufi and Michel, 2014, 2015). These results led to the concept of a possible translation (framing) code based on the circular code which was proposed in Michel (2012). The tRNAs and rRNAs are short non-gene regions, in average between 71 to 91 nucleotides for the tRNAs of prokaryotes and eukaryotes (Sections 2.4.1 and 2.4.2 in Michel, 2013), around 1500 nucleotides for the 16S rRNAs and around 1800 nucleotides for the 18S rRNAs (Table 1 in El Soufi and Michel, 2015). The X motifs in tRNAs and rRNAs have short lengths, up to 7 trinucleotides in 16S of prokaryotes (Table 2 in El Soufi and Michel, 2015), up to 5 trinucleotides in 18S of eukaryotes (Table 3 in El Soufi and Michel, 2015) and up to 7 trinucleotides in Ala-tRNA of *T. thermophilus* and Tyr-tRNA of *E. coli* (Tables 4a,t in El Soufi and Michel, 2015). These X motifs of short lengths retrieve the reading frame. Indeed, it was proved that X motifs of lengths greater than 4 trinucleotides always retrieve the reading frame, by definition of a circular code (Arquès and Michel, 1996).

In 2015, by quantifying the approach used in 1996 for identifying a preferential frame for each trinucleotide and by applying a massive statistical analysis of gene taxonomic groups, the circular code X is strengthened in genes of prokaryotes (7,851,762 genes, 2,481,566,882 trinucleotides) and eukaryotes (1,662,579 genes, 824,825,761 trinucleotides), and now also identified in genes of plasmids (237,486 genes, 68,244,356 trinucleotides) and viruses (184,344 genes, 45,688,798 trinucleotides) (Michel, 2015).

New properties of this circular code theory are identified here with several statistical analyzes of X motifs in 138 available complete eukaryotic genomes containing 91,421,182,030 bases with 3,133,622,680 bases for the genes (3.4%) and 88,287,559,350 bases for the non-gene regions (96.6%).

2. Method

2.1. Recall

We recall the basic definitions of complementary map $\mathcal{C}$, permutation map $\mathcal{P}$, code, trinucleotide code, trinucleotide circular code and self-complementary trinucleotide circular code in order to understand the concept of X motifs, i.e. motifs from the circular code X (Eq. (1)). The "advanced" definitions of maximal trinucleotide circular code, C^3 trinucleotide circular code and C^3 self-complementary trinucleotide circular code are given in Michel (2012, 2013) and El Soufi and Michel (2014, 2015).

Notation 1. The letters (or nucleotides or bases) define the genetic alphabet $A_4 = \{A, C, G, T\}$. The set of non-empty words (words, respectively) over A_4 is denoted by A_4^+ (A_4^*, respectively). The set of the 64 words of length 3 (trinucleotides or triletters) on A_4 is denoted by $A_4^3 = \{AAA,\ldots,TTT\}$. Let $x_1\bullet\bullet\bullet x_n$ be the concatenation of the words x_i for $i = 1, \ldots, n$, the symbol "$\bullet$" being the concatenation operator.

There are two important biological maps involved in codes in genes on A_4.

Definition 1. The *nucleotide complementarity map* $\mathcal{C}: A_4 \rightarrow A_4$ is defined by $\mathcal{C}(A) = T$, $\mathcal{C}(C) = G$, $\mathcal{C}(G) = C$, $\mathcal{C}(T) = A$. According to the property of the complementary and antiparallel double helix, the *trinucleotide complementarity map* $\mathcal{C}: A_4^3 \rightarrow A_4^3$ is defined by $\mathcal{C}(l_0\bullet l_1\bullet l_2) = \mathcal{C}(l_2)\bullet\mathcal{C}(l_1)\bullet\mathcal{C}(l_0)$ for all $l_0, l_1, l_2 \in A_4$, e.g. $\mathcal{C}(ACG) = CGT$. By extension to a trinucleotide set S, the *set complementarity map* $\mathcal{C}: \mathbb{P}\left(A_4^3\right) \rightarrow \mathbb{P}\left(A_4^3\right)$, $\mathbb{P}$ being the set of all subsets of A_4^3, is defined by $\mathcal{C}(S) = \left\{v : u, v \in A_4^3,\ u \in S,\ v = \mathcal{C}(u)\right\}$, i.e. a complementary trinucleotide set $\mathcal{C}(S)$ is obtained by applying the complementarity map $\mathcal{C}$ to all its trinucleotides, e.g. $\mathcal{C}\left(\{ACG, AGT\}\right) = \{ACT, CGT\}$.

Definition 2. The *trinucleotide circular permutation map* $\mathcal{P}: A_4^3 \rightarrow A_4^3$ is defined by $\mathcal{P}\left(l_0\bullet l_1\bullet l_2\right) = l_1\bullet l_2\bullet l_0$ for all $l_0, l_1, l_2 \in A_4$, e.g. $\mathcal{P}(ACG) = CGA$. The 2nd iterate of $\mathcal{P}$ is denoted $\mathcal{P}^2$, e.g. $\mathcal{P}^2(ACG) = GAC$. By extension to a trinucleotide set S, the *set circular permutation map* $\mathcal{P}: \mathbb{P}\left(A_4^3\right) \rightarrow \mathbb{P}\left(A_4^3\right)$ is defined by $\mathcal{P}(S) = \left\{v : u, v \in A_4^3,\ u \in S,\ v = \mathcal{P}(u)\right\}$, i.e. a permuted trinucleotide set $\mathcal{P}(S)$ is obtained by applying the circular permutation map $\mathcal{P}$ to all its trinucleotides, e.g. $\mathcal{P}\left(\{ACG, AGT\}\right) = \{CGA, GTA\}$ and $\mathcal{P}^2\left(\{ACG, AGT\}\right) = \{GAC, TAG\}$.

Definition 3. A set $S \subset A_4^+$ of words is a *code* if, for each $x_1,\ldots,x_n, y_1,\ldots,y_m \in S$, $n, m \geq 1$, the condition $x_1\bullet\bullet\bullet x_n = y_1\bullet\bullet\bullet y_m$ implies $n = m$ and $x_i = y_i$ for $i = 1, \ldots, n$.

Definition 4. As the set $A_4^3 = \{AAA,\ldots,TTT\}$ is a code, its non-empty subsets are codes and called *trinucleotide codes* X.

Definition 5. A trinucleotide code $X \subset A_4^3$ is *circular* if, for each $x_1,\ldots,x_n, y_1,\ldots,y_m \in X$, $n, m \geq 1$, $r \in A_4^*$, $s \in A_4^+$, the conditions $sx_2\bullet\bullet\bullet x_n r = y_1\bullet\bullet\bullet y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ (empty word) and $x_i = y_i$ for $i = 1, \ldots, n$.

For simplification and without loss of generality, we present the properties of circular codes with the trinucleotide circular codes, i.e. circular codes constituted of triletters.

Definition 6. An X circular code motif (X motif in brief) of a trinucleotide circular code X is a word written on a circle (the next letter after the last letter of the X motif being the first letter) which has a unique decomposition (factorization) into trinucleotides of X.

Fig. 1 gives a graphical representation of the trinucleotide circular code definition.

Example 1. The trinucleotide code $Y = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ is not circular. Indeed, the word $w = ATGGCCCTA$, for example, written on a circle, can be factorized into trinucleotides of Y according to two different ways: $ATG\bullet GCC\bullet CTA$ and $AAT\bullet GGC\bullet CCT$ (Fig. 2).

Example 2. The trinucleotide code $X = \{AAT, ATG, CCT, CTA, GCC, GTC\}$ obtained by replacing the trinucleotide GGC of Y by GTC, is

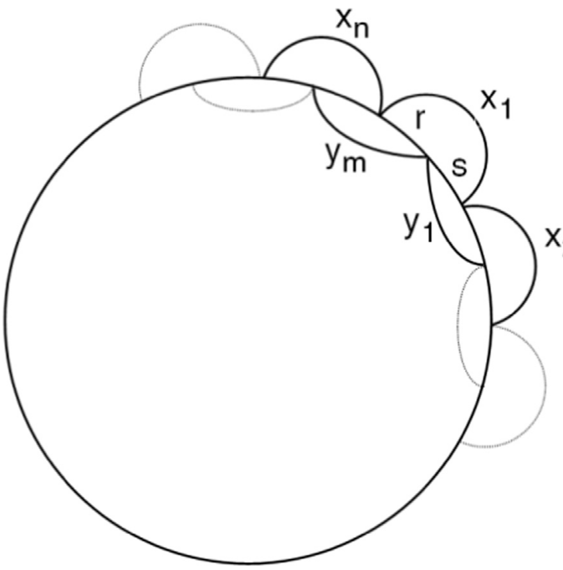


Fig. 1. A graphical representation of the trinucleotide circular code definition (Definition 5). A trinucleotide code X is circular if any word generated by a concatenation of trinucleotides of X and written on a circle has a unique decomposition into trinucleotides of X.

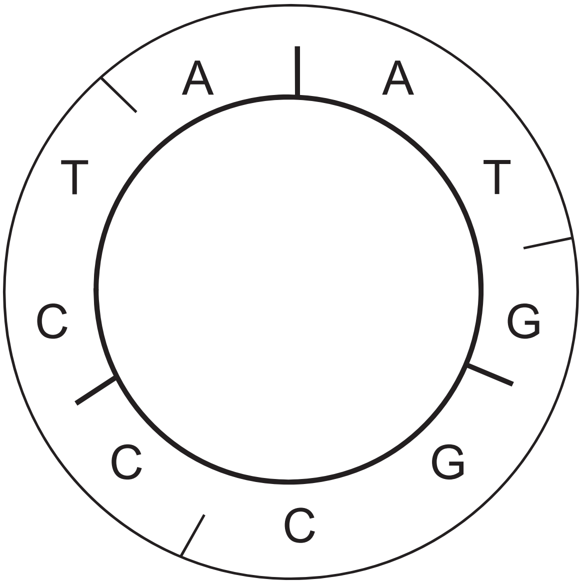


Fig. 2. The trinucleotide code Y={*AAT*, *ATG*, *CCT*, *CTA*, *GCC*, *GGC*} is not circular as the word w = *ATGGCCCTA* written on a circle can be factorized into trinucleotides of Y according to two different ways: *ATG•GCC•CTA* and *AAT•GGC•CCT*.

circular as there never exists words written on a circle with two decompositions, such as w for Y.

The proofs to decide that a code is circular or not are based on the flower automaton (Arquès and Michel, 1996), the necklaces 5*LDCN* (Letter Diletter Continued Necklace) (Pirillo, 2003) and *n*LDCCN (Letter Diletter Continued Closed Necklace) with $n \in \{2, 3, 4, 5\}$ (Michel and Pirillo, 2010), and the graph theory (Fimmel et al., 2016).

Remark 1. A trinucleotide code X containing either one periodic permuted trinucleotide $P_4^3 = \{AAA, CCC, GGG, TTT\}$ or two non-periodic permuted trinucleotides $\{t, \mathcal{P}(t)\}$ for a trinucleotide $t \in A_4^3 \setminus P_4^3$ cannot be circular. In both cases, there are words written on a circle without unique decomposition. Thus, the two trinucleotide codes A_4^3 and $A_4^3 \setminus P_4^3$ are not circular.

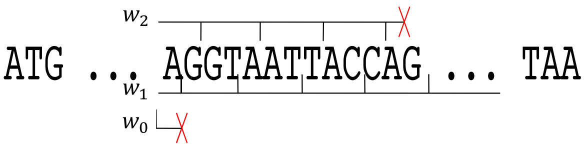


Fig. 3. Retrieval of the reading frame of the word w=...*AGGTAATTACCAG*... of the trinucleotide circular code X (Eq. (1)). Among the three possible factorizations w_0, w_1 and w_2, only one factorization w_1 into trinucleotides of X is possible leading to ...*A•GGT•AAT•TAC•CAG•*... Thus, the first letter A of w is the 3rd letter of a trinucleotide of X.

Definition 7. A trinucleotide circular code $X \subset A_4^3$ is *self-complementary* if, for each $t \in X$, $\mathcal{C}(t) \in X$.

The fundamental property of a trinucleotide circular code X is the ability to always retrieve the reading (original or constructed) frame of any word generated with X. The reading frame in a word is retrieved after the reading of a certain number of letters (nucleotides), called the window of X. The length of this window for retrieving the reading frame is the letter length of the longest ambiguous words which can be read in at least two frames, plus one letter.

Example 3. Suppose that the word w = ...*AGGTAATTACCAG*... has been constructed with the trinucleotide circular code X (Eq. (1)) (Fig. 3). By definition of a circular code, the construction of this word w is unique. Thus, we can decide unambiguously if the first nucleotide of w, i.e. *A*, is the 1st, the 2nd or the 3rd nucleotide of a trinucleotide of X? By trying the three possible factorizations (frames) w_0, w_1 and w_2 (w_1 and w_2 being w_0 shifted by one and two nucleotides, respectively) into trinucleotides of X, only one factorization, i.e. w_1, is possible. Thus, the first nucleotide A of w is the 3rd nucleotide of a trinucleotide of X. Indeed, the factorization w_1 leads to the trinucleotides *NNA*, *GGT*, *AAT*, *TAC* and *CAG* (*N* being any appropriate letter of X) which belong to X (Eq. (1)). The factorizations w_0 and w_2 are impossible as no trinucleotide of X starts with the prefix *AG* (Eq. (1)). This case occurs immediately for w_0 and after 11 letters for w_2 (Fig. 3). Thus, the unique factorization of w is w_1 = ...*A•GGT•AAT•TAC•CAG•*... This word w can be located anywhere in a sequence of X, i.e. the sequence of X does not require a start codon, a stop codon or any frame signal to retrieve the reading frame. The word w' = *AGGTAATTACCA* (w without the last *G*) with a length of 12 nucleotides is ambiguous as it has two factorizations w_1 and w_2 into trinucleotides of X (Fig. 3). This word w' is called an ambiguous word of X. By definition of a circular code, all the ambiguous words are finite words. The word w', taken as an illustration example here, is one of the four longest ambiguous words of X (Fimmel et al., 2016). Thus, the window length l to retrieve the construction frame of a word of a circular code X is the letter length of the longest ambiguous words w', plus one letter. With the trinucleotide circular code X (Eq. (1)), $l = 12 + 1 = 13$ nucleotides (Arquès and Michel, 1996).

The trinucleotide set X (Eq. (1)) coding the reading frame in genes is a maximal (20 trinucleotides) C^3 self-complementary (property $X = \mathcal{C}(X)$) trinucleotide circular code. The set $X_1 = \mathcal{P}(X)$ containing the 20 following trinucleotides

$$X_1 = \{AAG, ACA, ACG, ACT, AGC, AGG, ATA, ATG, CCA, CCG, GCG, GTG, TAG, TCA, TCC, TCG, TCT, TGC, TTA, TTG\} \quad (2)$$

and the set $X_2 = \mathcal{P}^2(X)$ containing the 20 following trinucleotides

$$X_2 = \{AGA, AGT, CAA, CAC, CAT, CCT, CGA, CGC, CGG, CGT, CTA, CTT, GCA, GCT, GGA, TAA, TAT, TGA, TGG, TGT\} \quad (3)$$

are also maximal trinucleotide circular codes (property C^3).

For the first time, we study here X circular code motifs (X motifs in brief) of the trinucleotide circular code X (Eq. (1)) in eukaryotic genomes. It is important to remind the reader of these two concepts: (i) the circular code X, which is a set of 20 trinucleotides (Eq. (1)); and (ii) X motifs which are words obtained (constructed, generated) with the circular code X. For example, *AAC•AAT* (a concatenation of the 1st and 2nd trinucleotides of X) and *TTC•TAC•AAC* (a concatenation of the 20th, 19th and 1st trinucleotides of X) are X motifs while *TTC•TAC•AAG* is not an X motif.

2.2. Definition of X motifs $m(X)$

The X motifs $m(X) = w_1 w_2 \ldots w_n$ with $w_i \in X$, $1 \le i \le n$, studied in eukaryotic genomes are defined by two parameters: their trinucleotide length and their trinucleotide cardinality (composition)

$$\begin{cases} n = l\left(m(X)\right) \\ \mathrm{Card}\left(\{w_1\} \cup \{w_2\} \cup \ldots \cup \{w_n\}\right) = \mathrm{Card}\left(\left\{w\left(m(X)\right)\right\}\right). \end{cases} \tag{4}$$

The particular class of large X motifs $m(X)$ studied is defined by the two conditions on their trinucleotide length and their trinucleotide cardinality

$$\begin{cases} l\left(m(X)\right) \ge 15 \text{ trinucleotides} \\ \mathrm{Card}\left(\left\{w\left(m(X)\right)\right\}\right) \ge 10 \text{ trinucleotides}. \end{cases} \tag{5}$$

Thus, the large X motifs $m(X)$ with lengths of at least 15 consecutive trinucleotides of X and compositions of at least 10 different trinucleotides of X differ from trinucleotide repeats. The latter is a particular case of tandem repeats where one trinucleotide or a very few number of different trinucleotides are concatenated in a series.

2.3. Definition of 23 bijective motifs

2.3.1. Bijective transformation circular codes

There are 23 bijective transformation circular codes $\Pi(X) = \left\{\pi_1(X), \ldots, \pi_{23}(X)\right\}$ of the maximal C^3 self-complementary trinucleotide circular code $X = \pi_0(X)$ (Table 1). The notation of bijective transformations used here is based on the notation of Michel and Seligmann (2014) which relies on (i) the transcript data identified from the human mitochondrial genome by Seligmann (2013a, 2013b); and (ii) the biological function of the polymerase. These biological observations suggest that bijective transformations of RNA transcripts using only two bases are simpler than bijective transformations of three bases which are also simpler than bijective transformations of four bases. Another notation of bijective transformations of circular codes is also proposed by Fimmel et al. (2013, page 225–226) in a study of circular codes based on group theory.

2.3.1.1. Partition into symmetric and asymmetric bijective transformation circular codes. The 23 bijective transformation circular codes $\Pi(X)$ of X can be partitioned into nine symmetric bijective transformation circular codes $\Pi_{\mathcal{S}}(X) = \left\{\pi_1(X), \ldots, \pi_9(X)\right\}$ and 14 asymmetric bijective transformation circular codes $\Pi_{\mathcal{A}}(X) = \left\{\pi_{10}(X), \ldots, \pi_{23}(X)\right\}$ (Table 1). The number $N(n, p)$ of bijective transformation circular codes at p letters among n letters is (obviously) equal to

$$N(n, p) = \frac{n!}{(n-p)!p}.$$

Table 1

The maximal C^3 self-complementary trinucleotide circular code $X = \pi_0(X)$ and its 23 bijective transformation circular codes $\Pi(X) = \left\{\pi_1(X), \ldots, \pi_{23}(X)\right\}$: the six symmetric bijective transformation circular codes $\Pi_{\mathcal{S},2}(X) = \left\{\pi_1(X), \pi_2(X), \pi_3(X), \pi_4(X), \pi_5(X), \pi_6(X)\right\}$ at 2 letters, the three symmetric bijective transformation circular codes $\Pi_{\mathcal{S},2,2}(X) = \left\{\pi_7(X), \pi_8(X), \pi_9(X)\right\}$ of two disjoint transformations at 2 letters, the eight asymmetric bijective transformation circular codes $\Pi_{\mathcal{A},3}(X) = \left\{\pi_{10}(X), \pi_{11}(X), \pi_{12}(X), \pi_{13}(X), \pi_{14}(X), \pi_{15}(X), \pi_{16}(X), \pi_{17}(X)\right\}$ at 3 letters and the six asymmetric bijective transformation circular codes $\Pi_{\mathcal{A},4}(X) = \left\{\pi_{18}(X), \pi_{19}(X), \pi_{20}(X), \pi_{21}(X), \pi_{22}(X), \pi_{23}(X)\right\}$ at 4 letters. The seven bijective transformations $\left\{\pi_3(X), \pi_4(X), \pi_7(X), \pi_8(X), \pi_9(X), \pi_{19}(X), \pi_{21}(X)\right\}$, in bold, are maximal C^3 self-complementary trinucleotide circular codes.

$X = \pi_0(X)$	*AAC*	*AAT*	*ACC*	*ATC*	*ATT*	*CAG*	*CTC*	*CTG*	*GAA*	*GAC*	*GAG*	*GAT*	*GCC*	*GGC*	*GGT*	*GTA*	*GTC*	*GTT*	*TAC*	*TTC*
$\pi_1(X)$:(*A*, *C*)	*CCA*	*CCT*	*CAA*	*CTA*	*CTT*	*ACG*	*ATA*	*ATG*	*GCC*	*GCA*	*GCG*	*GCT*	*GAA*	*GGA*	*GGT*	*GTC*	*GTA*	*GTT*	*TCA*	*TTA*
$\pi_2(X)$:(*A*, *G*)	*GGC*	*GGT*	*GCC*	*GTC*	*GTT*	*CGA*	*CTC*	*CTA*	*AGG*	*AGC*	*AGA*	*AGT*	*ACC*	*AAC*	*AAT*	*ATG*	*ATC*	*ATT*	*TGC*	*TTC*
$\pi_3(X)$:(*A*, *T*)	***TTC***	***TTA***	***TCC***	***TAC***	***TAA***	***CTG***	***CAC***	***CAG***	***GTT***	***GTC***	***GTG***	***GTA***	***GCC***	***GGC***	***GGA***	***GAT***	***GAC***	***GAA***	***ATC***	***AAC***
$\pi_4(X)$:(*C*, *G*)	***AAG***	***AAT***	***AGG***	***ATG***	***ATT***	***GAC***	***GTG***	***GTC***	***CAA***	***CAG***	***CAC***	***CAT***	***CGG***	***CCG***	***CCT***	***CTA***	***CTG***	***CTT***	***TAG***	***TTG***
$\pi_5(X)$:(*C*, *T*)	*AAT*	*AAC*	*ATT*	*ACT*	*ACC*	*TAG*	*TCT*	*TCG*	*GAA*	*GAT*	*GAG*	*GAC*	*GTT*	*GGT*	*GGC*	*GCA*	*GCT*	*GCC*	*CAT*	*CCT*
$\pi_6(X)$:(*G*, *T*)	*AAC*	*AAG*	*ACC*	*AGC*	*AGG*	*CAT*	*CGC*	*CGT*	*TAA*	*TAC*	*TAT*	*TAG*	*TCC*	*TTC*	*TTG*	*TGA*	*TGC*	*TGG*	*GAC*	*GGC*
$\pi_7(X)$:(*A*, *C*)(*G*, *T*)	***CCA***	***CCG***	***CAA***	***CGA***	***CGG***	***ACT***	***AGA***	***AGT***	***TCC***	***TCA***	***TCT***	***TCG***	***TAA***	***TTA***	***TTG***	***TGC***	***TGA***	***TGG***	***GCA***	***GGA***
$\pi_8(X)$:(*A*, *G*)(*C*, *T*)	***GGT***	***GGC***	***GTT***	***GCT***	***GCC***	***TGA***	***TCT***	***TCA***	***AGG***	***AGT***	***AGA***	***AGC***	***ATT***	***AAT***	***AAC***	***ACG***	***ACT***	***ACC***	***CGT***	***CCT***
$\pi_9(X)$:(*A*, *T*)(*C*, *G*)	***TTG***	***TTA***	***TGG***	***TAG***	***TAA***	***GTC***	***GAG***	***GAC***	***CTT***	***CTG***	***CTC***	***CTA***	***CGG***	***CCG***	***CCA***	***CAT***	***CAG***	***CAA***	***ATG***	***AAG***
$\pi_{10}(X)$:(*A*, *C*, *G*)	*CCG*	*CCT*	*CGG*	*CTG*	*CTT*	*GCA*	*GTG*	*GTA*	*ACC*	*ACG*	*ACA*	*ACT*	*AGG*	*AAG*	*AAT*	*ATC*	*ATG*	*ATT*	*TCG*	*TTG*
$\pi_{11}(X)$:(*A*, *C*, *T*)	*CCT*	*CCA*	*CTT*	*CAT*	*CAA*	*TCG*	*TAT*	*TAG*	*GCC*	*GCT*	*GCG*	*GCA*	*GTT*	*GGT*	*GGA*	*GAC*	*GAT*	*GAA*	*ACT*	*AAT*
$\pi_{12}(X)$:(*A*, *G*, *C*)	*GGA*	*GGT*	*GAA*	*GTA*	*GTT*	*AGC*	*ATA*	*ATC*	*CGG*	*CGA*	*CGC*	*CGT*	*CAA*	*CCA*	*CCT*	*CTG*	*CTA*	*CTT*	*TGA*	*TTA*
$\pi_{13}(X)$:(*A*, *G*, *T*)	*GGC*	*GGA*	*GCC*	*GAC*	*GAA*	*CGT*	*CAC*	*CAT*	*TGG*	*TGC*	*TGT*	*TGA*	*TCC*	*TTC*	*TTA*	*TAG*	*TAC*	*TAA*	*AGC*	*AAC*
$\pi_{14}(X)$:(*A*, *T*, *C*)	*TTA*	*TTC*	*TAA*	*TCA*	*TCC*	*ATG*	*ACA*	*ACG*	*GTT*	*GTA*	*GTG*	*GTC*	*GAA*	*GGA*	*GGC*	*GCT*	*GCA*	*GCC*	*CTA*	*CCA*
$\pi_{15}(X)$:(*A*, *T*, *G*)	*TTC*	*TTG*	*TCC*	*TGC*	*TGG*	*CTA*	*CGC*	*CGA*	*ATT*	*ATC*	*ATA*	*ATG*	*ACC*	*AAC*	*AAG*	*AGT*	*AGC*	*AGG*	*GTC*	*GGC*
$\pi_{16}(X)$:(*C*, *G*, *T*)	*AAG*	*AAC*	*AGG*	*ACG*	*ACC*	*GAT*	*GCG*	*GCT*	*TAA*	*TAG*	*TAT*	*TAC*	*TGG*	*TTG*	*TTC*	*TCA*	*TCG*	*TCC*	*CAG*	*CCG*
$\pi_{17}(X)$:(*C*, *T*, *G*)	*AAT*	*AAG*	*ATT*	*AGT*	*AGG*	*TAC*	*TGT*	*TGC*	*CAA*	*CAT*	*CAC*	*CAG*	*CTT*	*CCT*	*CCG*	*CGA*	*CGT*	*CGG*	*GAT*	*GGT*
$\pi_{18}(X)$:(*A*, *C*, *G*, *T*)	*CCG*	*CCA*	*CGG*	*CAG*	*CAA*	*GCT*	*GAG*	*GAT*	*TCC*	*TCG*	*TCT*	*TCA*	*TGG*	*TTG*	*TTA*	*TAC*	*TAG*	*TAA*	*ACG*	*AAG*
$\pi_{19}(X)$:(*A*, *C*, *T*, *G*)	***CCT***	***CCG***	***CTT***	***CGT***	***CGG***	***TCA***	***TGT***	***TGA***	***ACC***	***ACT***	***ACA***	***ACG***	***ATT***	***AAT***	***AAG***	***AGC***	***AGT***	***AGG***	***GCT***	***GGT***
$\pi_{20}(X)$:(*A*, *G*, *C*, *T*)	*GGT*	*GGA*	*GTT*	*GAT*	*GAA*	*TGC*	*TAT*	*TAC*	*CGG*	*CGT*	*CGC*	*CGA*	*CTT*	*CCT*	*CCA*	*CAG*	*CAT*	*CAA*	*AGT*	*AAT*
$\pi_{21}(X)$:(*A*, *G*, *T*, *C*)	***GGA***	***GGC***	***GAA***	***GCA***	***GCC***	***AGT***	***ACA***	***ACT***	***TGG***	***TGA***	***TGT***	***TGC***	***TAA***	***TTA***	***TTC***	***TCG***	***TCA***	***TCC***	***CGA***	***CCA***
$\pi_{22}(X)$:(*A*, *T*, *C*, *G*)	*TTG*	*TTC*	*TGG*	*TCG*	*TCC*	*GTA*	*GCG*	*GCA*	*ATT*	*ATG*	*ATA*	*ATC*	*AGG*	*AAG*	*AAC*	*ACT*	*ACG*	*ACC*	*CTG*	*CCG*
$\pi_{23}(X)$:(*A*, *T*, *G*, *C*)	*TTA*	*TTG*	*TAA*	*TGA*	*TGG*	*ATC*	*AGA*	*AGC*	*CTT*	*CTA*	*CTC*	*CTG*	*CAA*	*CCA*	*CCG*	*CGT*	*CGA*	*CGG*	*GTA*	*GGA*

Note: If $p = n$ then $N(n, n) = (n - 1)!$.

The nine symmetric bijective transformation circular codes $\Pi_{\mathcal{S}}(X)$ can again be partitioned into:

(1) $N(4,2) = 6$ symmetric bijective transformation circular codes $\Pi_{\mathcal{S},2}(X)$ at 2 letters

$$\Pi_{\mathcal{S},2}(X) = \left\{ \pi_1(X){:}(A, C), \pi_2(X){:}(A, G), \pi_3(X){:}(A, T), \pi_4(X){:}(C, G), \pi_5(X){:}(C, T), \pi_6(X){:}(G, T) \right\}$$

where $\pi_i(X){:}(l_1, l_2)$ is the ith bijective transformation in the lexicographical order of the letter $l_1 \in A_4$ into the letter $l_2 \in A_4$, $l_2 \neq l_1$, and reciprocally;

(2) $N(4,2)/2 = 3$ symmetric bijective transformation circular codes $\Pi_{\mathcal{S},2,2}(X)$ of two disjoint transformations at 2 letters

$$\Pi_{\mathcal{S},2,2}(X) = \left\{ \pi_7(X){:}(A, C)(G, T), \pi_8(X){:}(A, G)(C, T), \pi_9(X){:}(A, T)(C, G) \right\}$$

where $\pi_i(X){:}(l_1, l_2)(l_3, l_4)$ is the ith bijective transformation in the lexicographical order of the letter $l_1 \in A_4$ into the letter $l_2 \in A_4$, $l_2 \neq l_1$, and reciprocally, and of the letter $l_3 \in A_4$, $l_3 \neq l_2 \neq l_1$, into the letter $l_4 \in A_4$, $l_4 \neq l_3 \neq l_2 \neq l_1$, and reciprocally.

The 14 asymmetric bijective transformation circular codes $\Pi_{\mathcal{A}}(X)$ can also be partitioned into:

(1) $N(4,3) = 8$ asymmetric bijective transformation circular codes $\Pi_{\mathcal{A},3}(X)$ at 3 letters

$$\Pi_{\mathcal{A},3}(X) = \left\{ \pi_{10}(X){:}(A, C, G), \pi_{11}(X){:}(A, C, T), \pi_{12}(X){:}(A, G, C), \pi_{13}(X){:}(A, G, T), \pi_{14}(X){:}(A, T, C), \pi_{15}(X){:}(A, T, G), \pi_{16}(X){:}(C, G, T), \pi_{17}(X){:}(C, T, G) \right\}$$

where $\pi_i(X){:}(l_1, l_2, l_3)$ is the ith bijective transformation in the lexicographical order of the letter $l_1 \in A_4$ into the letter $l_2 \in A_4$, $l_2 \neq l_1$, the letter l_2 into the letter $l_3 \in A_4$, $l_3 \neq l_2 \neq l_1$, and the letter l_3 into the letter l_1;

(2) $N(4,4) = 6$ asymmetric bijective transformation circular codes $\Pi_{\mathcal{A},4}(X)$ at 4 letters

$$\Pi_{\mathcal{A},4}(X) = \left\{ \pi_{18}(X){:}(A, C, G, T), \pi_{19}(X){:}(A, C, T, G), \pi_{20}(X){:}(A, G, C, T), \pi_{21}(X){:}(A, G, T, C), \pi_{22}(X){:}(A, T, C, G), \pi_{23}(X){:}(A, T, G, C) \right\}$$

where $\pi_i(X){:}(l_1, l_2, l_3, l_4)$ is the ith bijective transformation in the lexicographical order of the letter $l_1 \in A_4$ into the letter $l_2 \in A_4$, $l_2 \neq l_1$, the letter l_2 into the letter $l_3 \in A_4$, $l_3 \neq l_2 \neq l_1$, the letter l_3 into the letter $l_4 \in A_4$, $l_4 \neq l_3 \neq l_2 \neq l_1$, and the letter l_4 into the letter l_1.

Note that the transformations at 1 ($X = \pi_0(X)$), 2, 3 and 4 letters are the transformations of order 1, 2, 3 and 4, respectively, according to the notation in Fimmel et al. (2013, page 225–226).

2.3.1.2. Partition into complementary and non-complementary bijective transformation circular codes. The 23 bijective transformation circular codes $\Pi(X)$ of X can also be partitioned into seven self-complementary bijective transformation circular codes $\Pi_{\mathcal{C}}(X) = \left\{ \pi_3(X), \pi_4(X), \pi_7(X), \pi_8(X), \pi_9(X), \pi_{19}(X), \pi_{21}(X) \right\}$ and 16 non self-complementary bijective transformation circular codes $\Pi_{\bar{\mathcal{C}}}(X) = \Pi(X) \setminus \Pi_{\mathcal{C}}(X)$ of X (Table 1).

2.3.1.3. Recall of the main properties of the 23 bijective transformation circular codes $\Pi(X)$

Proposition 1. The 23 bijective transformation circular codes $\Pi(X)$ of X are C^3.

Proof. By letter invariance, $\Pi(X)$ belongs to the set of the 221,328 C^3 trinucleotide circular codes (Michel, unpublished) or by Proposition 3 in Michel and Pirillo (2010) or by Theorem 1 in Fimmel et al. (2014).

Proposition 2. The seven bijective transformation circular codes $\Pi_{\mathcal{C}}(X) = \{\pi_3(X), \pi_4(X), \pi_7(X), \pi_8(X), \pi_9(X), \pi_{19}(X), \pi_{21}(X)\}$ are C^3 self-complementary.

Proof. By letter invariance for the complementarity map $\mathcal{C}$, $\Pi_{\mathcal{C}}(X)$ belongs to the set of the 216 C^3 self-complementary trinucleotide circular codes (Arquès and Michel, 1996) or by Proposition 3 in Michel and Pirillo (2010) or by Theorem 2 in Fimmel et al. (2014).

Proposition 3. The probability PrRFC (Definition 2.2.1 in Michel, 2014) of reading frame coding (RFC) of the 23 bijective transformation circular codes $\Pi(X)$ of X are obviously all equal to the probability PrRFC = 81.3% of X (Section 2.2.2.(vi) in Michel (2014)).

2.3.2. Definition of 23 bijective motifs $m(\Pi(X))$

The 23 bijective motifs $m(\Pi(X))$ are obtained from the 23 bijective transformation circular codes $\Pi(X)$ of the maximal C^3 self-complementary trinucleotide circular code X. For comparison with the large X motifs $m(X)$, the large bijective motifs $m(\Pi(X))$ must also satisfy the two conditions of Eq. (5), i.e. the length $l(m(\Pi(X))) \geq 15$ trinucleotides (at least 15 consecutive trinucleotides of $\Pi(X)$) and the cardinality $\mathrm{Card}\left(\left\{ w\left(m(\Pi(X))\right)\right\}\right) \geq 10$ trinucleotides (composition of at least 10 different trinucleotides of $\Pi(X)$).

2.4. Definition of two permuted motifs $m(X_1)$ and $m(X_2)$

The two permuted motifs $m(X_1)$ and $m(X_2)$ are obtained from the permuted circular codes $X_1 = \mathcal{P}(X)$ (Eq. (2)) and $X_2 = \mathcal{P}^2(X)$ (Eq. (3)), respectively, by applying the permutation map $\mathcal{P}$ to the maximal C^3 self-complementary trinucleotide circular code X. For comparison with the large X motifs $m(X)$, the large permuted motifs $m(X_1)$ and $m(X_2)$ must also satisfy the two conditions of Eq. (5), i.e. the lengths $l(m(X_1))$, $l(m(X_2)) \geq 15$ trinucleotides and the cardinalities $\mathrm{Card}\left(\left\{ w\left(m(X_1)\right)\right\}\right)$, $\mathrm{Card}\left(\left\{ w\left(m(X_2)\right)\right\}\right) \geq 10$ trinucleotides.

2.5. Definition of random motifs $m(R)$

The X motifs $m(X)$, $m(\Pi(X))$, $m(X_1)$ and $m(X_2)$ are generated from the maximal circular codes X, $\Pi(X)$, X_1 and X_2, respectively. All these circular codes have 20 trinucleotides with the same total numbers of nucleotides, i.e. 15 A, 15 C, 15 G, 15 T. Furthermore, by definition of a circular code, they have neither a periodic trinucleotide $P_4^3 = \{AAA, CCC, GGG, TTT\}$ nor two non-periodic permuted trinucleotides $\{t, \mathcal{P}(t)\}$ (Remark 1).

In order to have an evaluation of the statistical significance of occurrence numbers of the large X motifs $m(X)$, $m(\Pi(X))$, $m(X_1)$ and $m(X_2)$, 30 random codes R are generated with respect to the four necessary conditions of maximal circular codes: (i) a random code R with a number of trinucleotides equal to 20; (ii) a random code R without a periodic trinucleotide P_4^3; (iii) a random code R without two non-periodic permuted trinucleotides $\{t, \mathcal{P}(t)\}$;

and (iv) a random code R containing the same total numbers of nucleotides (15 A, 15 C, 15 G, 15 T). Then, a random code R of trinucleotides randomly chosen in A_4^3 is generated satisfying the four previous conditions (i), (ii), (iii) and (iv). The large random motifs $m(R)$ of a random trinucleotide code R must also satisfy the two conditions of Eq. (5), i.e. the length $l\left(m(R)\right) \geq 15$ trinucleotides and the cardinality $\mathrm{Card}\left(\left\{w\left(m(R)\right)\right\}\right) \geq 10$ trinucleotides.

2.6. Occurrence number of large X motifs $m(X)$, $m\left(\Pi(X)\right)$, $m(X_1)$, $m(X_2)$ and $m(R)$ in the genomes of eukaryotes

The occurrence numbers $N\left(m(X)\right)$ of large X motifs $m(X)$, $N\left(m\left(\Pi(X)\right)\right)$ of large bijective motifs $m\left(\Pi(X)\right)$, $N\left(m(X_1)\right)$ of large permuted motifs $m(X_1)$, $N\left(m(X_2)\right)$ of large permuted motifs $m(X_2)$ and $N\left(m(R)\right)$ of large random motifs $m(R)$ are computed in the eukaryotic genomes according to the following algorithm.

The algorithm searches for motifs in a DNA sequence with lengths greater than or equal to the parameter minsize and returns a list containing all motifs found in the sequence. Each motif has a start, an end and a frame according to the sequence, a length in trinucleotides and a cardinality in trinucleotides. This algorithm allows the retrieval of the maximum number of motifs in a sequence because it eliminates the issue of overlapping motifs in different frames. It is also suitable for multi-threading which greatly accelerate the search procedure.

```
Read sequence
INIT X AS a trinucleotide circular code
INIT minsize AS the minimum size of motifs
INIT shift
FOR EACH frame
    CASE frame OF
        0: set shift to 0
        1: set shift to 2
        2: set shift to 1
    ENDCASE
    INIT motif AS empty
    FOR EACH trinucl. in sequence starting from shift AS tri
        IF X contains tri THEN
            IF motif is empty THEN Set motif to tri
            ELSE Concatenate tri to motif
        ELSE
            IF motif length is larger than minsize THEN
                Add motif to list of motifs
            Set motif to empty
        ENDIF
    ENDFOR
ENDFOR
```

2.7. Expectation of the occurrence number of an X motif $m(X)$ in a DNA sequence

The expectation $\mathbb{E}\left[N\left(m_{\mathcal{G}_{Chr}}(X)\right)\right]$ of the occurrence number $N\left(m_{\mathcal{G}_{Chr}}(X)\right)$ of an X motif $m_{\mathcal{G}_{Chr}}(X)$ in a chromosome Chr of a genome $\mathcal{G}$ can easily be calculated with the Bernoulli model thank to equation:

$$\mathbb{E}\left[N\left(m_{\mathcal{G}_{Chr}}(X)\right)\right] = \left(N(\mathcal{G}_{Chr})-3l+1\right)\left(\frac{20}{64}\right)^l \tag{6}$$

where $N(\mathcal{G}_{Chr})$ is the total base number (size) of the chromosome Chr in $\mathcal{G}$, $l = l\left(m_{\mathcal{G}_{Chr}}(X)\right)$ is the trinucleotide length of $m_{\mathcal{G}_{Chr}}(X)$ and the term $\frac{20}{64}$ is the occurrence probability of a trinucleotide X (X has 20 trinucleotides among 64). Remember that any X motif $m(X)$ of length greater than four trinucleotides cannot overlap by definition of a circular code. Thus, the large X motifs $m_{\mathcal{G}_{Chr}}(X)$ with lengths $l \geq 15$ trinucleotides (Eq. (5)) cannot overlap.

2.8. Proportion of X motifs $m(X)$ in genes and non-gene regions of the eukaryotic genomes

The statistical analysis of X motifs $m(X)$ in a genome is based on two simple ratios: a base ratio of genes/non-genes for characterizing the base proportion of genes in a genome and a base ratio of X motifs in genes/non-genes for analyzing the base proportion of X motifs $m(X)$ in genes and non-gene regions of a genome.

The base ratio $r_G(\mathcal{G})$ of genes/non-genes in a genome $\mathcal{G}$ is defined as follows

$$r_G(\mathcal{G}) = \frac{N(\mathcal{G}_G)}{N(\mathcal{G}_{\bar{G}})} \tag{7}$$

where $N(\mathcal{G}_G)$ is the total base number of genes $\mathcal{G}_G$ in a given genome $\mathcal{G}$ and $N(\mathcal{G}_{\bar{G}})$ is the total base number of non-gene regions $\mathcal{G}_{\bar{G}}$ in $\mathcal{G}$ with $\mathcal{G} = \mathcal{G}_G \bigcup \mathcal{G}_{\bar{G}}$. The numbers $N(\mathcal{G}_G)$ and $N(\mathcal{G}_{\bar{G}})$ for the 138 studied complete eukaryotic genomes $\mathcal{G}$ are given in Appendix A.

Remark 2. $N(\mathcal{G}_G)+N(\mathcal{G}_{\bar{G}}) = N(\mathcal{G})$ where $N(\mathcal{G})$ is the total base number (size) of a genome $\mathcal{G}$ (also given in Appendix A).

Remark 3. When $r_G(\mathcal{G})<1$, the total base number $N(\mathcal{G}_G)$ of all genes $\mathcal{G}_G$ in a genome $\mathcal{G}$ is less than the total base number $N(\mathcal{G}_{\bar{G}})$ of all non-gene regions $\mathcal{G}_{\bar{G}}$ in $\mathcal{G}$, and conversely when $r_G(\mathcal{G})>1$.

Example 4. With the genome $\mathcal{G} =$ *Anolis carolinensis*, $N(\mathcal{G}_G) = 16670366$ and $N(\mathcal{G}_{\bar{G}}) = 1064974225$ (see Appendix A), then $r_G(\mathcal{G}) = 1.6\%$.

In order to study a greater variety of X motifs $m(X)$, i.e. not necessary large, the two length and cardinality (composition) conditions defined in Eq. (5) are relaxed. Thus, the X motifs $m(X)$ studied in this genome analysis are based on the two conditions

$$\begin{cases} l\left(m(X)\right) \geq 10 \text{ trinucleotides} \\ \mathrm{Card}\left(\left\{w\left(m(X)\right)\right\}\right) \geq 5 \text{ trinucleotides.} \end{cases} \tag{8}$$

The base ratio $r_{m(X)}(\mathcal{G})$ of X motifs in genes/non-genes in a genome $\mathcal{G}$ is defined as follows

$$r_{m(X)}(\mathcal{G}) = \frac{P\left(m_{\mathcal{G}_G}(X)\right)}{P\left(m_{\mathcal{G}_{\bar{G}}}(X)\right)} \tag{9}$$

where the probability $P\left(m_{\mathcal{G}_G}(X)\right) = \frac{N\left(m_{\mathcal{G}_G}(X)\right)}{N(\mathcal{G}_G)}$ is the total base number $N\left(m_{\mathcal{G}_G}(X)\right)$ of X motifs $m(X)$ in the genes $\mathcal{G}_G$ of a genome $\mathcal{G}$ divided by the total base number $N(\mathcal{G}_G)$ of genes $\mathcal{G}_G$ in $\mathcal{G}$ (see Eq. (7)), and the probability $P\left(m_{\mathcal{G}_{\bar{G}}}(X)\right) = \frac{N\left(m_{\mathcal{G}_{\bar{G}}}(X)\right)}{N(\mathcal{G}_{\bar{G}})}$ is the total base number $N\left(m_{\mathcal{G}_{\bar{G}}}(X)\right)$ of X motifs $m(X)$ in the non-gene regions $\mathcal{G}_{\bar{G}}$ of $\mathcal{G}$ divided by the total base number $N(\mathcal{G}_{\bar{G}})$ of non-gene regions $\mathcal{G}_{\bar{G}}$ in $\mathcal{G}$ (see Eq. (7)).

Remark 4. A ratio $r_{m(X)}(\mathcal{G}) = 1$ means that the proportion of X motifs $m(X)$ in genes $\mathcal{G}_G$ and non-genes $\mathcal{G}_{\bar{G}}$ is identical in the genome $\mathcal{G}$. A ratio $r_{m(X)}(\mathcal{G})<1$ means that there is a preferential occurrence of X motifs $m(X)$ in non-genes $\mathcal{G}_{\bar{G}}$ of $\mathcal{G}$. Conversely, a

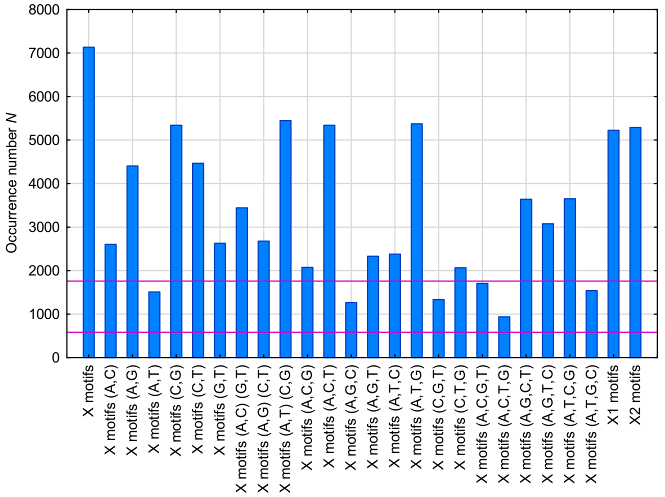


Fig. 4. Occurrence numbers $N\left(m(X)\right)$ of large X motifs $m(X)$, $N\left(m\left(\Pi(X)\right)\right)$ of its 23 large bijective motifs $m\left(\Pi(X)\right)$, $N\left(m(X_1)\right)$ and $N\left(m(X_2)\right)$ of its two large permuted motifs $m(X_1)$ and $m(X_2)$, respectively, in the 138 complete eukaryotic genomes (see Appendix A). All these 26 classes of large motifs have lengths $l \geq 15$ trinucleotides and cardinality (composition) Card ≥ 10 trinucleotides (Eq. (5)). The top horizontal line (1760) and the bottom horizontal line (582) represent the confidence interval at 99% (student t test by assuming a normal distribution of the population) of the mean occurrence number $\bar{N}\left(m(R)\right) = 1171$ (standard deviation $\sigma\left(m(R)\right) = 1170$) of large random motifs $m(R)$ from $\text{Card}(R) = 30$ random codes in the 138 eukaryotic genomes. The large X motifs $m(X)$ have the highest occurrence. The six large bijective motifs $m\left(\pi_2(X){:}(A, G)\right)$, $m\left(\pi_4(X){:}(C, G)\right)$, $m\left(\pi_5(X){:}(C, T)\right)$, $m\left(\pi_9(X){:}(A, T)(C, G)\right)$, $m\left(\pi_{11}(X){:}(A, C, T)\right)$ and $m\left(\pi_{15}(X){:}(A, T, G)\right)$, and the two large permuted motifs $m(X_1)$ and $m(X_2)$ have occurrence numbers greater than $\bar{N}\left(m(R)\right) + 2.75\sigma\left(m(R)\right) \approx 4400$.

ratio $r_{m(X)}(\mathcal{G}) > 1$ means that there is a preferential occurrence of X motifs $m(X)$ in genes $\mathcal{G}_G$ of $\mathcal{G}$.

The numbers $N\left(m_{\mathcal{G}_G}(X)\right)$ and $N\left(m_{\mathcal{G}_{\bar{G}}}(X)\right)$ of X motifs $m(X)$ in genes $\mathcal{G}_G$ and non-gene regions $\mathcal{G}_{\bar{G}}$, respectively, of a genome $\mathcal{G}$ are computed according to the following simple algorithm. We laid markers on the genomic sequence $\mathcal{G}$. The first marker labels the nucleotide in $\mathcal{G}$ that belongs to an X motif $m(X)$ and the second marker notes if this nucleotide belongs to a gene $\mathcal{G}_G$ (GenBank keyword CDS). The number $N\left(m_{\mathcal{G}_G}(X)\right)$ of X motifs $m(X)$ in genes $\mathcal{G}_G$ is obtained by counting the number of nucleotides that have two markers. The number $N\left(m_{\mathcal{G}_{\bar{G}}}(X)\right)$ of X motifs $m(X)$ in non-genes $\mathcal{G}_{\bar{G}}$ is the number of nucleotides that have one maker. Note that if an X motif $m(X)$ overlaps a non-gene region $\mathcal{G}_{\bar{G}}$ (5′ region) and a gene $\mathcal{G}_G$, or a gene $\mathcal{G}_G$ and a non-gene region $\mathcal{G}_{\bar{G}}$ (3′ region), its nucleotides are split accordingly.

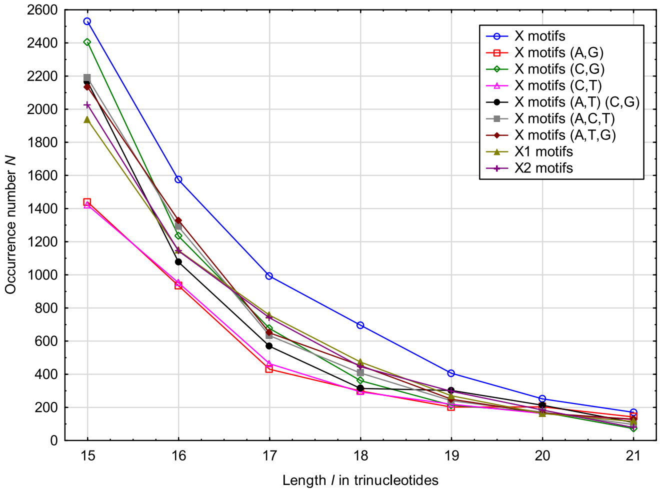


Fig. 5. Occurrence numbers $N\left(m(X)\right)$ of large X motifs $m(X)$, $N\left(m\left(\Pi(X)\right)\right)$ of its six large bijective motifs $m\left(\pi_2(X){:}(A, G)\right)$, $m\left(\pi_4(X){:}(C, G)\right)$, $m\left(\pi_5(X){:}(C, T)\right)$, $m\left(\pi_9(X){:}(A, T)(C, G)\right)$, $m\left(\pi_{11}(X){:}(A, C, T)\right)$ and $m\left(\pi_{15}(X){:}(A, T, G)\right)$, $N\left(m(X_1)\right)$ and $N\left(m(X_2)\right)$ of its two large permuted motifs $m(X_1)$ and $m(X_2)$ (greater than $\bar{N}\left(m(R)\right) + 2.75\sigma\left(m(R)\right) \approx 4400$, see Fig. 4) as a function of their lengths l varying from 15 to 21 trinucleotides in the 138 complete eukaryotic genomes (see Appendix A). All these classes of large motifs have cardinalities (composition) Card ≥ 10 trinucleotides (Eq. (5)). The large X motifs $m(X)$ have the highest occurrence for all trinucleotide lengths.

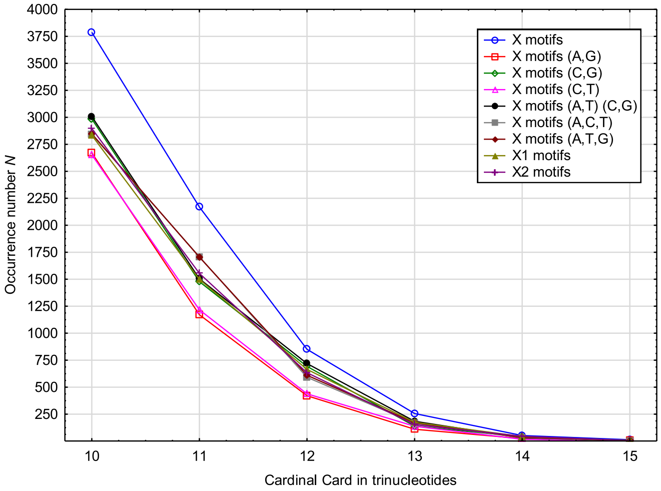


Fig. 6. Occurrence numbers $N\left(m(X)\right)$ of large X motifs $m(X)$, $N\left(m\left(\Pi(X)\right)\right)$ of its six large bijective motifs $m\left(\pi_2(X){:}(A, G)\right)$, $m\left(\pi_4(X){:}(C, G)\right)$, $m\left(\pi_5(X){:}(C, T)\right)$, $m\left(\pi_9(X){:}(A, T)(C, G)\right)$, $m\left(\pi_{11}(X){:}(A, C, T)\right)$ and $m\left(\pi_{15}(X){:}(A, T, G)\right)$, $N\left(m(X_1)\right)$ and $N\left(m(X_2)\right)$ of its two large permuted motifs $m(X_1)$ and $m(X_2)$ (greater than $\bar{N}\left(m(R)\right) + 2.75\sigma\left(m(R)\right) \approx 4400$, see Fig. 4) as a function of their cardinality (composition) Card varying from 10 to 15 trinucleotides in the 138 complete eukaryotic genomes (see Appendix A). All these classes of large motifs have lengths $l \geq 15$ trinucleotides (Eq. (5)). The large X motifs $m(X)$ have the highest occurrence for all trinucleotide cardinalities.

2.9. Genomic data

Using bioperl, we were able to retrieve all the eukaryotic chromosome sequences from the RefSeq database (GenBank keyword Reference Sequence). The RefSeq is a curated non-redundant sequence database of genomes. We took one species from each genus and only complete genomic molecules (GenBank keyword NC), excluding alternate assembly. One strain from each species is considered. Complete genomes $\mathcal{G}$ with total numbers $N\left(\mathcal{G}_{\bar{G}}\right) < 400000$ bases of non-gene regions $\mathcal{G}_{\bar{G}}$ are eliminated (in order to avoid several data with null values). Six such genomes are eliminated: *Cryptomonas paramecium* ($N\left(\mathcal{G}_{\bar{G}}\right) = 82348$ bases), *Encephalitozoon cuniculi* ($N\left(\mathcal{G}_{\bar{G}}\right) = 357485$ bases), *Encephalitozoon hellem* ($N\left(\mathcal{G}_{\bar{G}}\right) = 245811$ bases), *Encephalitozoon intestinalis* ($N\left(\mathcal{G}_{\bar{G}}\right)$ =230782 bases), *Encephalitozoon romaleae* ($N\left(\mathcal{G}_{\bar{G}}\right) = 215619$ bases) and *Nitzschia* ($N\left(\mathcal{G}_{\bar{G}}\right) = 14661$ bases). This led to 138 eukaryotic genomes. After filtering the database, we retrieved the Genbank file for each chromosome which allowed us to extract the coordinates of its genes (GenBank keyword CDS).

Thus, 138 complete genomes of eukaryotes are extracted from GenBank (http://www.ncbi.nlm.nih.gov/genbank/, April 2016). The genome list and the total base numbers $N\left(\mathcal{G}_G\right)$ of genes $\mathcal{G}_G$ and $N\left(\mathcal{G}_{\bar{G}}\right)$ of non-gene regions $\mathcal{G}_{\bar{G}}$, and their sum $N(\mathcal{G}) = N\left(\mathcal{G}_G\right) + N\left(\mathcal{G}_{\bar{G}}\right)$ for the 138 complete eukaryotic genomes $\mathcal{G}$ are given in Appendix A. This genome information represents a total of 91,421,182,030 bases with 3,133,622,680 bases for the genes (3.4%) and 88,287,559,350 bases for the non-gene regions (96.6%). The human chromosome list and the total base numbers

Table 2
The top 20 largest X motifs $m_{\mathcal{G}_{Chr}}(X)$ with cardinality (composition) Card ≥10 trinucleotides (Eq. (5)) in the chromosomes $\mathcal{G}_{Chr}$ of the 138 complete eukaryotic genomes $\mathcal{G}$ (see Appendix A) in descending order of trinucleotide lengths $l \geq 45$. The 1st and 2nd columns give the genome $\mathcal{G}$ and its chromosome number $\mathcal{G}_{Chr}$, respectively, the 3rd column gives its base size $N(\mathcal{G}_{Chr})$, the 4th and 5th columns indicate the start and end positions of the largest X motif $m_{\mathcal{G}_{Chr}}(X)$ in the chromosome $\mathcal{G}_{Chr}$, the 6th column gives the trinucleotide length l of the largest X motif $m_{\mathcal{G}_{Chr}}(X)$, the 7th column indicates its expectation $\mathbb{E}$ (Eq. (6)), and the last column mentions if the largest X motif $m_{\mathcal{G}_{Chr}}(X)$ belongs (Yes) or not (No) to a gene.

Genome $\mathcal{G}$	$\mathcal{G}_{Chr}$	Size $N(\mathcal{G}_{Chr})$ (in bases)	Start position	End position	Length l of X motifs (in trinucleotides)	Expectation $\mathbb{E}$ (Eq. (6))	In gene
Solanum pennellii	3	75414019	36982714	36983178	155	10^{-71}	No
Salmo salar	15	103963436	16024777	16025130	118	10^{-52}	No
Salmo salar	15	103963436	17850373	17850726	118	10^{-52}	No
Monodelphis domestica	2	541556283	513328228	513328533	102	10^{-43}	No
Solanum lycopersicum	8	65866657	30359989	30360276	96	10^{-41}	No
Monodelphis domestica	4	435153693	290107123	290107407	95	10^{-40}	No
Plasmodium falciparum	11	2038337	872956	873216	87	10^{-38}	Yes
Equus caballus	28	46177339	35484817	35485047	77	10^{-32}	No
Bombus terrestris	14	11649563	11165956	11166153	66	10^{-27}	Yes
Sorghum bicolor	4	68034345	38474677	38474856	60	10^{-23}	No
Felis catus	3	140925898	2211844	2212020	59	10^{-22}	No
Cynoglossus semilaevis	9	19616557	14919031	14919192	54	10^{-20}	No
Plasmodium knowlesi	13	2200295	1265167	1265322	52	10^{-20}	Yes
Mus musculus	1	195471971	74368813	74368968	52	10^{-18}	Yes
Micromonas sp.	12	1084119	530353	530496	48	10^{-19}	Yes
Dictyostelium discoideum	2	8484197	1796161	1796304	48	10^{-18}	Yes
Apis mellifera	4	12718334	12440101	12440241	47	10^{-17}	No
Salmo salar	19	82978132	46877047	46877184	46	10^{-16}	No
Bombus terrestris	15	11467329	3286219	3286353	45	10^{-16}	No
Camelina sativa	10	25316904	13177546	13177680	45	10^{-16}	No

$N(\mathcal{H}_{Chr_G})$ of genes Chr_G and $N(\mathcal{H}_{Chr_{\bar{G}}})$ of non-gene regions $Chr_{\bar{G}}$, and their sum $N(\mathcal{H}_{Chr}) = N(\mathcal{H}_{Chr_G})+N(\mathcal{H}_{Chr_{\bar{G}}})$ for the 24 chromosomes $\mathcal{H}_{Chr}$ of the complete human genome $\mathcal{G} = \mathcal{H} = Homo\ sapiens$ are given in Appendix B.

3. Results

3.1. Occurrence of large random motifs $m(R)$ in the genomes of eukaryotes

The mean number $\bar{N}(m(R)) = \frac{1}{\text{Card}(R)} \sum_{j=1}^{\text{Card}(R)} N(m(R_j))$ and its standard deviation $\sigma(m(R))$ of large random motifs $m(R)$ from $\text{Card}(R) = 30$ random codes are determined in the 138 eukaryotic genomes. The computation leads to $\bar{N}(m(R)) = 1171$ and $\sigma(m(R)) = 1170$. By assuming a normal distribution of the population, a student t test gives a confidence interval at 99% for the mean $\bar{N}(m(R))$ equal to $[582,1760]$ (represented in Fig. 4). Note that the number of random codes R generated was limited to 30 as their statistical analysis in the 138 eukaryotic genomes (91,421,182,030 bases) takes several days.

3.2. Highest occurrence of large X motifs $m(X)$ in the genomes of eukaryotes compared to its 23 large bijective motifs $m(\Pi(X))$ and its two large permuted motifs $m(X_1)$ and $m(X_2)$

Fig. 4 shows the occurrence numbers $N(m(X))$ of large X motifs $m(X)$, $N(m(\Pi(X)))$ of its 23 large bijective motifs $m(\Pi(X))$, $N(m(X_1))$ and $N(m(X_2))$ of its two large permuted motifs $m(X_1)$ and $m(X_2)$, respectively, in the 138 complete eukaryotic genomes. All these 26 classes of large motifs have lengths $l \geq 15$ trinucleotides and cardinality (composition) $\text{Card} \geq 10$ trinucleotides (Eq. (5)). The large X motifs $m(X)$ have the highest occurrence with $N(m(X)) = 7133$ compared to all the 25 other classes of large motifs $m(\Pi(X))$, $m(X_1)$ and $m(X_2)$ in genomes of eukaryotes. Eight large motifs also occur significantly with numbers greater than $\bar{N}(m(R))+2.75\sigma(m(R)) \approx 4400$ (where $\bar{N}(m(R))$ and $\sigma(m(R))$ are given in Section 3.1). They are in descending fashion: $m(\pi_9(X){:}(A,T)(C,G))$ with $N(m(\pi_9(X))) = 5447$, $m(\pi_{15}(X){:}(A,T,G))$ with $N(m(\pi_{15}(X))) = 5374$, $m(\pi_4(X){:}(C,G))$ with $N(m(\pi_4(X))) = 5341$, $m(\pi_{11}(X){:}(A,C,T))$ with $N(m(\pi_{11}(X))) = 5341$, $m(X_2)$ with $N(m(X_2)) = 5289$, $m(X_1)$ with $N(m(X_1)) = 5223$, $m(\pi_5(X){:}(C,T))$ with $N(m(\pi_5(X))) = 4466$ and $m(\pi_2(X){:}(A,G))$ with $N(m(\pi_2(X))) = 4404$ (Fig. 4). Note that $\pi_2(X)$, $\pi_4(X)$ and $\pi_5(X)$ are symmetric bijective transformation circular codes $\Pi_{\mathcal{S},2}(X)$ at 2 letters, $\pi_9(X)$ is a symmetric bijective transformation circular code $\Pi_{\mathcal{S},2,2}(X)$ of two disjoint transformations at 2 letters, and $\pi_{11}(X)$ and $\pi_{15}(X)$ are asymmetric bijective transformation circular codes $\Pi_{\mathcal{A},3}(X)$ at 3 letters (Section 2.3.1.1). Note also that $\pi_4(X)$ and $\pi_9(X)$ are C^3 self-complementary trinucleotide circular codes (Section 2.3.1.2 and Proposition 2).

The six motifs $m(\pi_3(X){:}(A,T))$, $m(\pi_{12}(X){:}(A,G,C))$, $m(\pi_{16}(X){:}(C,G,T))$, $m(\pi_{18}(X){:}(A,C,G,T))$, $m(\pi_{19}(X){:}(A,C,T,G))$ and $m(\pi_{23}(X){:}(A,T,G,C))$ occur randomly ($N(m(\pi_i(X))) \in [582,1760]$, $i = 3,12,16,18,19,23$, see Section 3.1) and the four motifs $m(\pi_{10}(X){:}(A,C,G))$, $m(\pi_{13}(X){:}(A,G,T))$, $m(\pi_{14}(X){:}(A,T,C))$ and $m(\pi_{17}(X){:}(C,T,G))$ have low occurrences $(2000<N(m(\pi_i(X)))<2400$, $i = 10,13,14,17)$ (Fig. 4).

Figs. 5 and 6 strengthen the previous results. Indeed, Fig. 5 shows that the large X motifs $m(X)$ with cardinality $\text{Card} \geq 10$ trinucleotides (Eq. (5)) have the highest occurrence compared to all the 25 other classes of large motifs $m(\Pi(X))$, $m(X_1)$ and $m(X_2)$ (with cardinalities $\text{Card} \geq 10$ trinucleotides) for all lengths l from

Table 3

Largest X motifs $m_{\mathcal{H}_{Chr}}(X)$ with cardinality (composition) Card ≥ 10 trinucleotides (Eq. (5)) and expectation $\mathbb{E}<1$ (Eq. (6)) in the chromosomes $\mathcal{H}_{Chr}$ of the human genome $\mathcal{G}=\mathcal{H}=$ *Homo sapien*. The 1st and 2nd columns give the human chromosome number $\mathcal{H}_{Chr}$ and its base size $N(\mathcal{H}_{Chr})$, respectively, the 3rd column shows the largest X motifs $m_{\mathcal{H}_{Chr}}(X)$ with cardinality Card ≥ 10 trinucleotides and expectation $\mathbb{E}<1$, the 4th and 5th columns indicate the start and end positions of the largest X motif $m_{\mathcal{H}_{Chr}}(X)$ in the chromosome $\mathcal{H}_{Chr}$, the 6th column gives the trinucleotide length l of the largest X motif $m_{\mathcal{H}_{Chr}}(X)$, the 7th column indicates its expectation $\mathbb{E}$, and the last column mentions if the largest X motif $m_{\mathcal{H}_{Chr}}(X)$ belongs (Yes) or not (No) to a gene.

$\mathcal{H}_{Chr}$	Size $N(\mathcal{H}_{Chr})$ (in bases)	Largest X motifs $m_{\mathcal{H}_{Chr}}(X)$ in the human chromosomes $\mathcal{H}_{Chr}$	Start position	End position	Length l of X motifs (in trinucleotides)	Expectation $\mathbb{E}$ (Eq. (6))	In gene
1	248956422	GAG,GAG,GAG,CTG,CTG,GCC,CAG,CTG,GAG,GAG,TAC,GAG,CAG,GTC,ATC,CTG,GAC,TTC, CAG,TTC,AAC,CTG,GAG,GCC,ACC	3763375	3763449	25	5.9×10^{-5}	Yes
2	242193529	GTC,GAT,GAG,CAG,AAT,GCC,CAG,ACC,CAG,GAG,CAG,GAG,GGC,TTC,GTC,CTG,GGC,CTC	233449984	233450037	18	2.0×10^{-1}	Yes
4	190214555	GCC,ATC,ATT,ATC,ATT,ATC,ATC,CTC,ACC,TTC,ATC,ATT,AAT,AAC,CTG,GGC,CAG,GGT	42018853	42018906	18	1.5×10^{-1}	No
5	181538259	GAA,ATC,TTC,ATC,ATT,ACC,CTC,ACC,GCC,GCC,ATC,ATT,GAC,CTG,GTT,AAT,GTT	133306903	133306953	17	4.7×10^{-1}	No
7	159345973	ATC,ACC,CAG,GAT,GAA,GAT,GGT,CTC,ACC,CTG,CTC,ATT,GAG,GAT,GCC,GGT,GGT	30452806	30452856	17	4.1×10^{-1}	Yes
8	145138636	ACC,GTC,ACC,AAC,CTG,TTC,ATC,CTC,AAC,CTG,GCC,ATC,GCC,GAC,GAG,CTC,TTC	52940113	52940163	17	3.8×10^{-1}	Yes
9	138394717	GGT,CTC,CAG,GCC,AAT,GTC,ATT,GAC,GTC,ACC,ATC,ATC,GCC,ATC,ACC,ATC,ATT,ACC	95705686	95705739	18	1.1×10^{-1}	No
11	135086622	GAT,GAT,GCC,ACC,ACC,CTC,TAC,CTG,CAG,AAC,AAC,CAG,ATC,AAC,AAC,GCC,GGC,ATC	64116508	64116561	18	1.1×10^{-1}	Yes
13	114364328	AAT,GAG,GAC,ACC,ACC,CAG,GGC,ATC,GCC,AAC,GAG,GAA,GCC,GCC,CAG,GGC,ATC,GCC, GAG,GAC,GCC,ATC,CAG,GGC,ATC, GCC,AAC,GAG,GAG,GTT,GCC,CAG,GGC,ATC,GCC,AAT	18235684	18235791	36	7.5×10^{-11}	No
14	107043718	GCC,CAG,GAC,GAC,GAG,GGT,CTG,CTG,GAC,AAC,TTC,GTC,ACC,TTC,TTC,ATT	99716146	99716193	16	8.9×10^{-1}	Yes
15	101991189	GGC,GAA,GAA,GGT,GAA,GAT,GAA,GAG,GAT,GAA,GAT,CTG,GCC,CTC,GGT,GAC,CAG,GTA	68208355	68208408	18	8.2×10^{-2}	Yes
17	83257441	CTG,CTG,GTT,GAA,GTT,GTC,AAT,GAT,GAC,GCC,AAT,GAA,GAG,GTT,GAG,GGT,GAA,GAA	63944680	63944733	18	6.7×10^{-2}	Yes
18	80373285	ATC,GAG,CAG,AAT,GCC,ACC,AAC,ACC,TTC,CTG,GTC,TAC,ACC,GAG,GAG,GAC	49583566	49583613	16	6.6×10^{-1}	Yes
19	58617616	GAA,ACC,AAC,CAG,GTC,CTC,ATC,AAC,ATT,GGC,CTG,CTG,CTC,CTG,GCC,TTC	13959991	13960038	16	4.8×10^{-1}	Yes
20	64444167	TAC,CTG,GCC,CAG,GTC,CAG,GGT,GAC,GTT,GAC,CTC,GTT,GTA,CTC,CAG,GCC	62362396	62362443	16	5.3×10^{-1}	No
22	50818468	CAG,GTT,GAA,GAA,GTT,GTA,GTT,GCC,GGT,GAT,GAT,AAT,CAG,GAC,CTG,CAG,CAG	50505760	50505810	17	1.3×10^{-1}	Yes
X	156040895	CTC,CAG,GTA,GAG,GGC,ATT,GAG,CAG,CTC,AAT,GAT,GTC,AAC,GAG,GAC,CTG,GTT,GTC	39981361	39981414	18	1.3×10^{-1}	No

Table 4
Base ratio $r_G(\mathcal{G})$ (Eq. (7) in %) of genes/non-genes and base ratio $r_{m(X)}(\mathcal{G})$ (Eq. (9)) of X motifs $m(X)$ of length $l \geq 10$ trinucleotides and cardinality (composition) Card ≥ 5 trinucleotides (Eq. (8)) in genes/non-genes of the 138 complete eukaryotic genomes $\mathcal{G}$ (see Appendix A).

Genome $\mathcal{G}$	$r_G(\mathcal{G})$ (%)	$r_{m(X)}(\mathcal{G})$	Genome $\mathcal{G}$	$r_G(\mathcal{G})$ (%)	$r_{m(X)}(\mathcal{G})$	Genome $\mathcal{G}$	$r_G(\mathcal{G})$ (%)	$r_{m(X)}(\mathcal{G})$
Anolis carolinensis	1.6	5.3	*Esox lucius*	5.5	12.2	*Ovis aries*	1.3	20.0
Anopheles gambiae	8.6	15.6	*Felis catus*	1.4	19.4	*Pan paniscus*	1.1	9.4
Apis mellifera	8.2	3.5	*Ficedula albicollis*	2.5	19.6	*Pan troglodytes*	1.1	8.8
Arabidopsis thaliana	38.6	5.4	*Fragaria vesca*	18.5	5.1	*Papio anubis*	1.3	7.5
Aspergillus fumigatus	93.7	8.7	*Gallus gallus*	2.9	16.6	*Phaeodactylum tricornutum*	114.7	3.0
Babesia bigemina	196.0	5.2	*Glycine max*	6.9	4.5	*Phaseolus vulgaris*	7.2	4.1
Babesia bovis	213.3	9.1	*Gorilla gorilla*	1.2	9.4	*Plasmodium cynomolgi*	69.9	5.6
Babesia microti	263.9	5.4	*Gossypium raimondii*	6.3	6.0	*Plasmodium falciparum*	111.1	79.3
Beta vulgaris	7.0	4.0	*Homo sapiens*	1.2	8.4	*Plasmodium knowlesi*	90.1	15.6
Bombus terrestris	8.1	3.8	*Kazachstania africana*	239.2	8.4	*Plasmodium vivax*	93.1	16.6
Bos taurus	1.3	21.9	*Kluyveromyces lactis*	223.9	9.8	*Poecilia reticulata*	5.9	16.5
Brachypodium distachyon	14.0	6.6	*Komagataella phaffii*	358.7	6.4	*Pongo abelii*	1.1	9.0
Brassica napus	13.9	4.8	*Lachancea thermotolerans*	260.5	16.4	*Populus trichocarpa*	13.2	2.8
Brassica oleracea	12.9	5.0	*Leishmania braziliensis*	94.8	9.3	*Prunus mume*	17.3	5.6
Brassica rapa	23.1	6.4	*Leishmania donovani*	82.2	6.8	*Rattus norvegicus*	1.4	10.2
Caenorhabditis briggsae	28.9	6.5	*Leishmania infantum*	95.0	8.8	*Saccharomyces cerevisiae*	257.2	15.2
Caenorhabditis elegans	36.1	6.4	*Leishmania major*	91.6	8.2	*Salmo salar*	3.3	11.7
Callithrix jacchus	1.2	8.8	*Leishmania mexicana*	96.5	7.8	*Scheffersomyces stipitis*	125.3	3.3
Camelina sativa	19.7	4.3	*Leishmania panamensis*	90.1	8.1	*Schizosaccharomyces pombe*	131.4	6.0
Candida dubliniensis	156.4	3.4	*Lepisosteus oculatus*	3.7	21.9	*Sesamum indicum*	15.1	5.0
Candida glabrata	179.8	9.3	*Macaca fascicularis*	1.2	7.6	*Setaria italica*	9.8	7.4
Candida orthopsilosis	202.1	6.9	*Macaca mulatta*	1.2	7.6	*Solanum lycopersicum*	4.4	4.1
Canis lupus	1.5	13.0	*Magnaporthe oryzae*	70.0	11.8	*Solanum pennellii*	3.9	3.3
Capra hircus	1.2	20.7	*Malus domestica*	7.4	7.1	*Sorghum bicolor*	6.0	6.7
Chlorocebus sabaeus	1.3	7.3	*Medicago truncatula*	14.2	4.6	*Sus scrofa*	1.2	26.4
Chrysemys picta	1.3	8.8	*Meleagris gallopavo*	2.7	14.6	*Taeniopygia guttata*	2.4	18.9
Cicer arietinum	9.0	4.5	*Micromonas sp.*	228.4	2.4	*Takifugu rubripes*	11.3	10.4
Ciona intestinalis	24.8	6.6	*Microtus ochrogaster*	1.5	15.8	*Tetrapisispora blattae*	165.4	7.7
Citrus sinensis	13.0	3.6	*Monodelphis domestica*	1.0	4.5	*Tetrapisispora phaffii*	197.6	9.5
Cryptococcus gattii	124.6	5.4	*Mus musculus*	1.3	12.5	*Thalassiosira pseudonana*	119.0	1.5
Cryptococcus neoformans	115.2	6.6	*Myceliophthora thermophila*	57.4	18.5	*Theileria annulata*	266.6	8.0
Cryptosporidium parvum	298.9	4.6	*Nasonia vitripennis*	13.6	11.4	*Theileria equi*	223.3	3.1
Cucumis sativus	15.2	6.5	*Naumovozyma castellii*	286.4	8.6	*Theileria orientalis*	216.1	2.7
Cyanidioschyzon merolae	81.5	3.7	*Naumovozyma dairenensis*	175.6	6.9	*Theileria parva*	215.3	5.1
Cynoglossus semilaevis	9.0	7.2	*Neospora caninum*	44.8	11.8	*Theobroma cacao*	11.6	6.7
Danio rerio	3.4	7.0	*Neurospora crassa*	58.1	11.5	*Thielavia terrestris*	58.4	14.6
Debaryomyces hansenii	288.2	7.7	*Nomascus leucogenys*	1.2	8.5	*Torulaspora delbrueckii*	367.6	6.6
Dictyostelium discoideum	161.8	13.3	*Ogataea parapolymorpha*	545.6	7.6	*Tribolium castaneum*	11.4	1.2
Drosophila melanogaster	17.4	11.1	*Oreochromis niloticus*	5.7	14.8	*Trypanosoma brucei*	150.1	2.1
Drosophila pseudoobscura	23.9	7.6	*Ornithorhynchus anatinus*	1.1	3.2	*Ustilago maydis*	156.3	8.6
Drosophila simulans	14.7	13.6	*Oryctolagus cuniculus*	1.1	19.2	*Vigna radiata*	9.8	4.9
Drosophila yakuba	20.3	10.5	*Oryza brachyantha*	12.8	12.4	*Vitis vinifera*	8.0	6.0
Elaeis guineensis	4.3	11.2	*Oryza sativa*	8.7	8.1	*Yarrowia lipolytica*	85.2	14.9
Equus caballus	1.4	18.1	*Oryzias latipes*	4.9	18.2	*Zea mays*	2.2	4.0
Eremothecium cymbalariae	202.6	3.1	*Ostreococcus lucimarinus*	231.1	1.3	*Zygosaccharomyces rouxii*	319.5	3.6
Eremothecium gossypii	335.6	15.8	*Ostreococcus tauri*	437.3	1.6	*Zymoseptoria tritici*	56.8	4.0
						Mean	79.2	9.3
						Median	15.2	7.6

15 to 21 trinucleotides. Fig. 6 shows that the large X motifs $m(X)$ with lengths $l \geq 15$ trinucleotides (Eq. (5)) have the highest occurrence compared to all the 25 other classes of large motifs $m(\Pi(X))$, $m(X_1)$ and $m(X_2)$ (with lengths $l \geq 15$ trinucleotides) for all cardinalities Card from 10 to 15 trinucleotides.

3.3. Largest X motifs $m(X)$ in the genomes of eukaryotes

Table 2 gives the top 20 largest X motifs $m_{\mathcal{G}_{Chr}}(X)$ with cardinality (composition) Card ≥ 10 trinucleotides (Eq. (5)) in the chromosomes $\mathcal{G}_{Chr}$ of the 138 complete eukaryotic genomes $\mathcal{G}$ in decreasing order of their trinucleotide lengths $l \geq 45$. The 1st largest X motif $m_{Solanum_3}(X)$ is observed in a non-gene region of the chromosome $Chr = 3$ in the genome $\mathcal{G} = Solanum\ pennellii$. It has a length of $l = 155$ trinucleotides (465 nucleotides) and an expectation $\mathbb{E}\left[N\left(m_{Solanum_3}(X)\right)\right] = 10^{-71}$ (Eq. (6)). The 2nd and 3rd largest X motifs $m_{Salmo_{15}}(X)$ are observed in non-gene regions of the chromosome $Chr = 15$ in the genome $\mathcal{G} = Salmo\ salar$. They have a different composition but the same length $l = 118$ trinucleotides (354 nucleotides) and an expectation $\mathbb{E}\left[N\left(m_{Salmo_{15}}(X)\right)\right] = 10^{-52}$. The biological function and evolution of these unexpected large X motifs in the eukaryotic genomes are unknown.

3.4. Largest X motifs $m(X)$ in the chromosomes of the human genome

Table 3 shows the largest X motifs $m_{\mathcal{H}_{Chr}}(X)$ with cardinality (composition) Card ≥ 10 trinucleotides (Eq. (5)) and expectation $\mathbb{E} < 1$ (Eq. (6)) in the chromosomes $\mathcal{H}_{Chr}$ of the human genome $\mathcal{G} = \mathcal{H} = Homo\ sapiens$. The largest X motif $m_{\mathcal{H}_{13}}(X)$ is found in a non-gene region of the human chromosome $Chr = 13$. It has a length of $l = 36$ trinucleotides and an expectation $\mathbb{E}\left[N\left(m_{\mathcal{H}_{13}}(X)\right)\right] = 7.5 \times 10^{-11}$ (Eq. (6)).

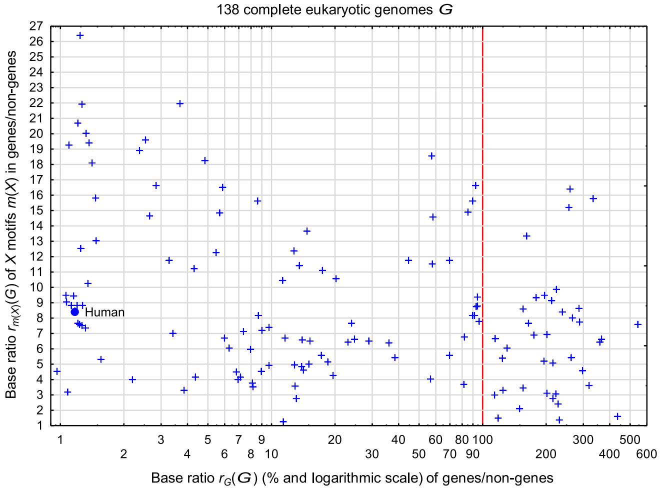


Fig. 7. Base ratio $r_G(\mathcal{G})$ (Eq. (7) in %) of genes/non-genes and base ratio $r_{m(X)}(\mathcal{G})$ (Eq. (9)) of X motifs $m(X)$ of length $l \geq 10$ trinucleotides and cardinality (composition) Card ≥ 5 trinucleotides (Eq. (8)) in genes/non-genes of the 138 complete eukaryotic genomes $\mathcal{G}$ (see Appendix A). The vertical red line $r_G(\mathcal{G}) = 100\%$ makes a partition of genomes $\mathcal{G}$ according to their base content in genes. When $r_G(\mathcal{G}) < 100\%$, the total base number $N(\mathcal{G}_G)$ of all genes $\mathcal{G}_G$ in the genome $\mathcal{G}$ is less than the total base number $N(\mathcal{G}_{\bar{G}})$ of all non-gene regions $\mathcal{G}_{\bar{G}}$ in $\mathcal{G}$, and conversely when $r_G(\mathcal{G}) > 100\%$ (see Remark 3). The genome $\mathcal{G}$ = *Plasmodium falciparum* with $r_{m(X)}(\mathcal{G}) = 79.3$ is not represented in the figure (see Table 4). There is no correlation between $r_G(\mathcal{G})$ and $r_{m(X)}(\mathcal{G})$ $(r = -0.12)$.

Table 5

Base ratio $r_G(\mathcal{H}_{Chr})$ (Eq. (7) in %) of genes/non-genes and base ratio $r_{m(X)}(\mathcal{H}_{Chr})$ (Eq. (9)) of X motifs $m(X)$ of length $l \geq 10$ trinucleotides and cardinality (composition) Card ≥ 5 trinucleotides (Eq. (8)) in genes/non-genes of the 24 chromosomes $\mathcal{H}_{Chr}$ in the human genome $\mathcal{G} = \mathcal{H}$ = *Homo sapiens* (see Appendix B).

$\mathcal{H}_{Chr}$	$r_G(\mathcal{H}_{Chr})$ (%)	$r_{m(X)}(\mathcal{H}_{Chr})$
1	1.5	8.6
2	1.1	8.0
3	1.0	8.7
4	0.8	5.2
5	0.9	7.8
6	1.1	8.3
7	1.1	6.9
8	0.8	7.3
9	1.1	7.9
10	1.1	5.8
11	1.6	8.2
12	1.4	6.5
13	0.6	7.7
14	1.1	9.1
15	1.2	7.4
16	1.7	6.6
17	2.5	7.2
18	0.7	7.1
19	4.1	6.5
20	1.3	9.2
21	0.8	10.4
22	1.6	12.0
X	0.9	9.7
Y	0.2	11.9
Mean	1.3	8.1
Median	1.1	7.8

3.5. *X motifs $m(X)$ in genes and non-gene regions of eukaryotic genomes*

The maximal C^3 self-complementary trinucleotide circular code X is a well-known coding property of genes. Indeed, it is observed in genes of bacteria, eukaryotes, plasmids and viruses (Michel, 2015; Arquès and Michel, 1996).

Table 4 gives the base ratio $r_G(\mathcal{G})$ (Eq. (7) in %) of genes/non-genes and the base ratio $r_{m(X)}(\mathcal{G})$ (Eq. (9)) of X motifs $m(X)$ of length $l \geq 10$ trinucleotides and cardinality (composition) Card ≥ 5 trinucleotides (Eq. (8)) in genes/non-genes of the 138 complete eukaryotic genomes $\mathcal{G}$.

The lowest value $r_G(\mathcal{G})$ of genes/non-genes is observed with the genome $\mathcal{G}$ = *Monodelphis domestica* with $r_G(\mathcal{G}) = 1.0\%$ ($r_{m(X)}(\mathcal{G}) = 4.5$). The highest value $r_G(\mathcal{G})$ of genes/non-genes is observed with the genome $\mathcal{G}$ = *Ogataea parapolymorpha* with $r_G(\mathcal{G}) = 545.6\%$ ($r_{m(X)}(\mathcal{G}) = 7.6$). The mean value is $\bar{r}_G(\mathcal{G}) = 79.2\%$ and the median value $\tilde{r}_G(\mathcal{G}) = 15.2\%$.

The lowest value $r_{m(X)}(\mathcal{G})$ of X motifs in genes/non-genes is observed with the genome $\mathcal{G}$ = *Tribolium castaneum* with $r_{m(X)}(\mathcal{G}) = 1.2$ ($r_G(\mathcal{G}) = 11.4\%$). The highest value $r_{m(X)}(\mathcal{G})$ of X motifs in genes/non-genes is observed with the genome $\mathcal{G}$ = *Plasmodium falciparum* with $r_{m(X)}(\mathcal{G}) = 79.3$ ($r_G(\mathcal{G}) = 111.1\%$). The mean value is $\bar{r}_{m(X)}(\mathcal{G}) = 9.3$ and the median value $\tilde{r}_{m(X)}(\mathcal{G}) = 7.6$.

Fig. 7 gives a graphical representation of Table 4. There is no correlation between $r_G(\mathcal{G})$ and $r_{m(X)}(\mathcal{G})$ $(r = -0.12)$.

Thus, as expected according to previous works, the X motifs $m(X)$ occur preferentially in genes of genomes with a factor of about 8 ($\tilde{r}_{m(X)}(\mathcal{G}) = 7.6 < 8 < \bar{r}_{m(X)}(\mathcal{G}) = 9.3$). Furthermore, this circular code property is verified whatever the base content of genes in the genomes $(r = -0.12)$.

3.6. *X motifs $m(X)$ in genes and non-gene regions of the 24 chromosomes in the human genome*

Table 5 gives the base ratio $r_G(\mathcal{H}_{Chr})$ (Eq. (7) in %) of genes/non-genes and the base ratio $r_{m(X)}(\mathcal{H}_{Chr})$ (Eq. (9)) of X motifs $m(X)$ of length $l \geq 10$ trinucleotides and cardinality (composition) Card ≥ 5 trinucleotides (Eq. (8)) in genes/non-genes of the 24 chromosomes $\mathcal{H}_{Chr}$ in the human genome $\mathcal{G} = \mathcal{H}$ = *Homo sapiens*.

The lowest value $r_G(\mathcal{H}_{Chr})$ of genes/non-genes is observed with the chromosome $Chr = Y$ with $r_G(\mathcal{H}_Y) = 0.2\%$ ($r_{m(X)}(\mathcal{H}_Y) = 11.9$). The highest value $r_G(\mathcal{H}_{Chr})$ of genes/non-genes is observed with the chromosome $Chr = 19$ with $r_G(\mathcal{H}_{19}) = 4.1\%$ ($r_{m(X)}(\mathcal{H}_{19}) = 6.5$). The mean value is $\bar{r}_G(\mathcal{H}_{Chr}) = 1.3\%$ and the median value $\tilde{r}_G(\mathcal{H}_{Chr}) = 1.1\%$.

Remark 5. These two values $\bar{r}_G(\mathcal{H}_{Chr}) = 1.3\%$ and $\tilde{r}_G(\mathcal{H}_{Chr}) = 1.1\%$ are very close from $r_G(\mathcal{H}) = 1.2\%$ (Table 4).

The lowest value $r_{m(X)}(\mathcal{H}_{Chr})$ of X motifs in genes/non-genes is observed with the chromosome $Chr = 4$ with $r_{m(X)}(\mathcal{H}_4) = 5.2$ ($r_G(\mathcal{H}_4) = 0.8\%$). The highest value $r_{m(X)}(\mathcal{H}_{Chr})$ of X motifs in genes/non-genes is observed with the chromosome $Chr = 22$ with $r_{m(X)}(\mathcal{H}_{22}) = 12.0$ $\left(r_G(\mathcal{H}_{22}) = 1.6\%\right)$. The mean value is $\bar{r}_{m(X)}(\mathcal{H}_{Chr}) = 8.1$ and the median value $\tilde{r}_{m(X)}(\mathcal{H}_{Chr}) = 7.8$.

Remark 6. These two values $\bar{r}_{m(X)}(\mathcal{H}_{Chr}) = 8.1$ and $\tilde{r}_{m(X)}(\mathcal{H}_{Chr}) = 7.8$ are also very close from $r_{m(X)}(\mathcal{H}) = 8.4$ (Table 4).

Fig. 8 gives a graphical representation of Table 5. There is no correlation between $r_G(\mathcal{H}_{Chr})$ and $r_{m(X)}(\mathcal{H}_{Chr})$ $(r = -0.26)$.

As in the general case, the X motifs $m(X)$ occur preferentially in genes of human chromosomes with a factor of about 8 ($\tilde{r}_{m(X)}(\mathcal{H}_{Chr}) = 7.8 < 8 < \bar{r}_{m(X)}(\mathcal{H}_{Chr}) = 8.1$). Furthermore, this circular code property is also verified whatever the base content of genes in human chromosomes $(r = -0.26)$.

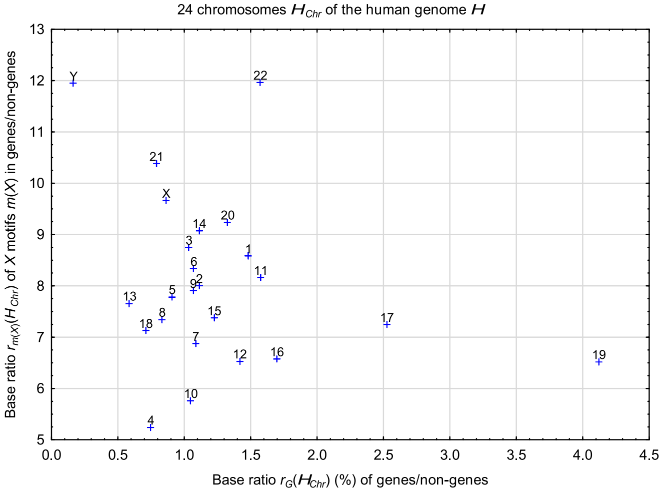


Fig. 8. Base ratio $r_G(\mathcal{H}_{Chr})$ (Eq. (7) in %) of genes/non-genes and base ratio $r_{m(X)}(\mathcal{H}_{Chr})$ (Eq. (9)) of X motifs $m(X)$ of length $l \geq 10$ trinucleotides and cardinality (composition) Card ≥ 5 trinucleotides (Eq. (8)) in genes/non-genes of the 24 chromosomes $\mathcal{H}_{Chr}$ in the human genome $\mathcal{G} = \mathcal{H} = $ *Homo sapiens* (see Appendix B). There is no correlation between $r_G(\mathcal{H}_{Chr})$ and $r_{m(X)}(\mathcal{H}_{Chr})$ ($r = -0.26$).

4. Discussion

X circular code motifs are found in genes of bacteria, eukaryotes, plasmids and viruses (Michel, 2015; Arquès and Michel, 1996), tRNAs of prokaryotes and eukaryotes, and rRNAs of prokaryotes (16S) and eukaryotes (18S), in particular in the ribosome decoding center (Michel, 2012, 2013; El Soufi and Michel, 2014, 2015). The universally conserved nucleotides G530, A1492 and A1493 are included in X motifs (Michel, 2012; El Soufi and Michel, 2014, 2015). These short X motifs in tRNAs and rRNAs (see Introduction and Tables 2, 3, 4a,t in El Soufi and Michel, 2015) have the circular code property for retrieving, synchronizing and maintaining the reading frame in genes, the C^3 property for retrieving the two shifted frames in genes and the complementary property for pairing, in particular between DNAs-DNAs, DNAs-mRNAs, mRNAs-rRNAs, mRNAs-tRNAs and rRNAs-tRNAs, as shown with a 3D visualization of X motifs in the ribosome (Michel, 2012; El Soufi and Michel, 2014, 2015). All these properties suggest a possible translation (framing) code in genes based on the circular code (Michel, 2012).

New properties of this circular code theory are identified here with robust statistical studies of X motifs $m(X)$ in genomes of eukaryotes. This study shines light on non-gene regions, that were not examined previously, as well as gene regions. It has also been proposed that the circular code X, which is associated with the regular RNA transcription, may use its bijective transformation codes $\Pi(X)$ for coding nucleotide exchanging RNA transcription (Michel and Seligmann, 2014). The large X motifs $m(X)$ (having lengths $l \geq 15$ trinucleotides and cardinalities (composition) Card ≥ 10 trinucleotides, Eq. (5)) have the highest occurrence in genomes of eukaryotes compared to (i) its 23 large bijective motifs $m(\Pi(X))$ from the bijective transformation circular codes $\Pi(X)$, (ii) its two large permuted motifs $m(X_1)$ and $m(X_2)$ from the permuted circular codes $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$, and (iii) large random motifs $m(R)$ from random codes R (Section 3.2 and Figs. 1–3). The largest X motifs identified in genomes are presented (Section 3.3 and Table 2), e.g. an X motif in a non-gene region of the genome *Solanum pennellii* with a length of 155 trinucleotides (465 nucleotides) and an expectation $\mathbb{E} = 10^{-71}$ (Eq. (6)), two X motifs in non-gene regions of the genome *Salmo salar* with lengths of 118 trinucleotides (354 nucleotides) and an expectation $\mathbb{E} = 10^{-52}$, etc. Large X motifs are also found in the human genome (Section 3.4 and Table 3). The largest X motif occurs in a non-gene region of the human chromosome 13 with a length of 36 trinucleotides and an expectation $\mathbb{E} = 10^{-11}$. X motifs in non-gene regions of genomes are possibly evolutionary relics of primitive genes using the circular code for translation. However, the mean value $\bar{r}_{m(X)}(\mathcal{G})$ and the median value $\tilde{r}_{m(X)}(\mathcal{G})$ giving the proportion of X motifs (having lengths $l \geq 10$ trinucleotides and cardinalities Card ≥ 5 trinucleotides, Eq. (8)) in genes/non-genes of the 138 complete eukaryotic genomes $\mathcal{G}$ are close to 8 ($\bar{r}_{m(X)}(\mathcal{G}) = 9.3 \approx \tilde{r}_{m(X)}(\mathcal{G}) = 7.6 \approx 8$, Section 3.5 and Table 4). This factor of 8 is retrieved for the X motifs in genes/non-genes of the 24 human chromosomes $\mathcal{H}_{Chr}$ ($\bar{r}_{m(X)}(\mathcal{H}_{Chr}) = 8.1 \approx \tilde{r}_{m(X)}(\mathcal{H}_{Chr})$ $=7.8 \approx 8$, Section 3.6 and Table 5). Thus, the X motifs occur preferentially in genes. This property is true whatever the base content of genes in the genomes as there is no correlation between the base proportion of genes/non-genes in genomes and the base proportion of X motifs in genes/non-genes of genomes (Figs. 4, 5). From a biological point of view, this property may be explained by the fact that mutations (substitution, insertion and deletion of nucleotides) are more frequent in non-gene regions compared to genes. Finally, the statistical analysis developed here is based on the search of exact X motifs. X motifs with a few mutations in genomes of eukaryotes should also be investigated in future.

Appendix A. Data of eukaryotic genomes

List and total base numbers $N(\mathcal{G}_G)$ of genes $\mathcal{G}_G$ and $N(\mathcal{G}_{\bar{G}})$ of non-gene regions $\mathcal{G}_{\bar{G}}$, and their sum $N(\mathcal{G}) = N(\mathcal{G}_G)$+$N(\mathcal{G}_{\bar{G}})$ for the 138 complete eukaryotic genomes $\mathcal{G}$ extracted from the GenBank (http://www.ncbi.nlm.nih.gov/genbank/, April 2016):

Genomes $\mathcal{G}$	Gene bases $N(\mathcal{G}_G)$	Non-gene bases $N(\mathcal{G}_{\bar{G}})$	Total bases $N(\mathcal{G})$	Genomes $\mathcal{G}$	Gene bases $N(\mathcal{G}_G)$	Non-gene bases $N(\mathcal{G}_{\bar{G}})$	Total bases $N(\mathcal{G})$
Anolis carolinensis	16670366	1064974225	1081644591	*Malus domestica*	36138315	490059574	526197889
Anopheles gambiae	1935976	22457132	24393108	*Medicago truncatula*	47725325	336741668	384466993
Apis mellifera	16592730	203036882	219629612	*Meleagris gallopavo*	25172179	947030988	972203167
Arabidopsis thaliana	33175579	85970769	119146348	*Micromonas sp.*	14597320	6392006	20989326
Aspergillus fumigatus	14214225	15170733	29384958	*Microtus ochrogaster*	23979075	1631404432	1655383507
Babesia bigemina	6801553	3469771	10271324	*Monodelphis*	26405867	2727912010	2754317877

				domestica			
Babesia bovis	2942868	1379871	4322739	*Mus musculus*	14915903	1190656585	1205572488
Babesia microti	4627831	1753458	6381289	*Myceliophthora thermophila*	5976840	10408460	16385300
Beta vulgaris	24577029	352006668	376583697	*Nasonia vitripennis*	13849880	102179764	116029644
Bombus terrestris	16328176	200521166	216849342	*Naumovozyma castellii*	8316040	2903499	11219539
Bos taurus	34037257	2681728647	2715765904	*Naumovozyma dairenensis*	8618630	4908950	13527580
Brachypodium distachyon	33348761	238427717	271776478	*Neospora caninum*	17793122	39754298	57547420
Brassica napus	94788198	680325795	775113993	*Neurospora crassa*	14868399	25594673	40463072
Brassica oleracea	51062958	395822924	446885882	*Nomascus leucogenys*	32421857	2762838188	2795260045
Brassica rapa	48059517	208363946	256423463	*Ogataea parapolymorpha*	7499949	1374640	8874589
Caenorhabditis briggsae	20457533	70777254	91234787	*Oreochromis niloticus*	35450942	621900030	657350972
Caenorhabditis elegans	26613936	73658671	100272607	*Ornithorhynchus anatinus*	4691381	432388643	437080024
Callithrix jacchus	32941100	2737278115	2770219215	*Oryctolagus cuniculus*	24420043	2223332061	2247752104
Camelina sativa	95232658	483211609	578444267	*Oryza brachyantha*	28480058	222443280	250923338
Candida dubliniensis	8917936	5700486	14618422	*Oryza sativa*	30547069	351603876	382150945
Candida glabrata	7914961	4403284	12318245	*Oryzias latipes*	33534282	689907207	723441489
Candida orthopsilosis	8468943	4190458	12659401	*Ostreococcus lucimarinus*	9216998	3987890	13204888
Canis lupus	34021609	2293612375	2327633984	*Ostreococcus tauri*	10138133	2318218	12456351
Capra hircus	30265609	2494397111	2524662720	*Ovis aries*	33794145	2551021749	2584815894
Chlorocebus sabaeus	35692308	2708423003	2744115311	*Pan paniscus*	33289497	3118617730	3151907227
Chrysemys picta	5803852	455943505	461747357	*Pan troglodytes*	34403316	3056708897	3091112213
Cicer arietinum	28623811	318623566	347247377	*Papio anubis*	34126862	2690200812	2724327674
Ciona intestinalis	15550846	62745309	78296155	*Phaeodactylum tricornutum*	13966979	12171777	26138756
Citrus sinensis	27421823	211577885	238999708	*Phaseolus vulgaris*	34393133	480427395	514820528
Cryptococcus gattii	10193549	8181211	18374760	*Plasmodium cynomolgi*	9350600	13377735	22728335
Cryptococcus neoformans	10546316	9153466	19699782	*Plasmodium falciparum*	12245290	11019048	23264338
Cryptosporidium parvum	6820333	2281991	9102324	*Plasmodium knowlesi*	11118740	12343447	23462187
Cucumis sativus	25366500	166492524	191859024	*Plasmodium vivax*	10906305	11714766	22621071
Cyanidioschyzon merolae	7429255	9117492	16546747	*Poecilia reticulata*	38655401	658045552	696700953
Cynoglossus semilaevis	36786472	408352885	445139357	*Pongo abelii*	32110815	2997380214	3029491029
Danio rerio	44231259	1296199332	1340430591	*Populus trichocarpa*	44068914	334476651	378545565
Debaryomyces hansenii	9022180	3130306	12152486	*Prunus mume*	29298313	169554093	198852406
Dictyostelium discoideum	20979100	12963972	33943072	*Rattus norvegicus*	37109116	2744903486	2782012602
Drosophila melanogaster	4239527	24318227	28557754	*Saccharomyces cerevisiae*	8691722	3379604	12071326
Drosophila pseudoobscura	9775205	40832070	50607275	*Salmo salar*	71170520	2169034471	2240204991
Drosophila simulans	2308887	15683400	17992287	*Scheffersomyces stipitis*	8587907	6853272	15441179
Drosophila yakuba	3900892	19244445	23145337	*Schizosaccharomyces pombe*	7138394	5433426	12571820
Elaeis guineensis	27281976	630686860	657968836	*Sesamum indicum*	30558151	202664230	233222381
Equus caballus	32994722	2334058725	2367053447	*Setaria italica*	35711158	365585260	401296418
Eremothecium cymbalariae	6473618	3195806	9669424	*Solanum lycopersicum*	33641774	768496446	802138220
Eremothecium gossypii	7007631	2088117	9095748	*Solanum pennellii*	34535865	891890599	926426464
Esox lucius	36362423	664661728	701024151	*Sorghum bicolor*	37431478	621797889	659229367

Felis catus	32750934	2386461976	2419212910	*Sus scrofa*	32005814	2564633642	2596639456
Ficedula albicollis	25797019	1018268272	1044065291	*Taeniopygia guttata*	23660215	997802725	1021462940
Fragaria vesca	30904815	167212294	198117109	*Takifugu rubripes*	28676957	252895405	281572362
Gallus gallus	28351491	993087537	1021439028	*Tetrapisispora blattae*	8755400	5293193	14048593
Glycine max	60862926	888313116	949176042	*Tetrapisispora phaffii*	8034104	4066086	12100190
Gorilla gorilla	33405815	2884281198	2917687013	*Thalassiosira pseudonana*	15615332	13118203	28733535
Gossypium raimondii	44683789	704544301	749228090	*Theileria annulata*	6074113	2278407	8352520
Homo sapiens	35915410	3052354422	3088269832	*Theileria equi*	4155315	1860488	6015803
Kazachstania africana	7848851	3281289	11130140	*Theileria orientalis*	6141721	2841875	8983596
Kluyveromyces lactis	7388969	3300187	10689156	*Theileria parva*	3080757	1431157	4511914
Komagataella phaffii	7207175	2009203	9216378	*Theobroma cacao*	34445261	296010936	330456197
Lachancea thermotolerans	7509690	2883172	10392862	*Thielavia terrestris*	13614268	23297988	36912256
Leishmania braziliensis	15200552	16037552	31238104	*Torulaspora delbrueckii*	7248844	1971834	9220678
Leishmania donovani	14635818	17809150	32444968	*Tribolium castaneum*	19166507	168328462	187494969
Leishmania infantum	15556104	16368749	31924853	*Trypanosoma brucei*	13292454	8855634	22148088
Leishmania major	15710352	17144737	32855089	*Ustilago maydis*	11979357	7664534	19643891
Leishmania mexicana	15190689	15747000	30937689	*Vigna radiata*	29662798	303645666	333308464
Leishmania panamensis	14542185	16146609	30688794	*Vitis vinifera*	31432213	394743796	426176009
Lepisosteus oculatus	31820297	859323780	891144077	*Yarrowia lipolytica*	9430576	11072405	20502981
Macaca fascicularis	34465017	2837360992	2871826009	*Zea mays*	44366789	2015334939	2059701728
Macaca mulatta	34674220	2801289170	2835963390	*Zygosaccharomyces rouxii*	7436797	2327838	9764635
Magnaporthe oryzae	16673311	23818662	40491973	*Zymoseptoria tritici*	14379863	25306388	39686251
				Total	3133622680	88287559350	91421182030

Appendix B. Data of human genome

List and total base numbers $N\left(\mathcal{H}_{Chr_G}\right)$ of genes Chr_G and $N\left(\mathcal{H}_{Chr_{\bar{G}}}\right)$ of non-gene regions $Chr_{\bar{G}}$, and their sum $N\left(\mathcal{H}_{Chr}\right) = N\left(\mathcal{H}_{Chr_G}\right) + N\left(\mathcal{H}_{Chr_{\bar{G}}}\right)$ for the 24 chromosomes $\mathcal{H}_{Chr}$ of the complete human genome $\mathcal{G} = \mathcal{H} = Homo\ sapiens$ extracted from the GenBank (http://www.ncbi.nlm.nih.gov/genbank/, April 2016):

Human chromosome $\mathcal{H}_{Chr}$	Gene bases $N\left(\mathcal{H}_{Chr_G}\right)$	Non-gene bases $N\left(\mathcal{H}_{Chr_{\bar{G}}}\right)$	Total bases $N\left(\mathcal{H}_{Chr}\right)$
1	3640059	245316363	248956422
2	2669855	239523674	242193529
3	2035080	196260479	198295559
4	1416419	188798136	190214555
5	1637897	179900362	181538259
6	1814234	168991745	170805979
7	1715132	157630841	159345973
8	1202323	143936313	145138636
9	1464139	136930578	138394717
10	1390981	132406441	133797422
11	2097637	132988985	135086622
12	1864798	131410511	133275309
13	669656	113694672	114364328
14	1179821	105863897	107043718

15	1236353	100754836	101991189
16	1508869	88829476	90338345
17	2051840	81205601	83257441
18	567889	79805396	80373285
19	2320757	56296859	58617616
20	844705	63599462	64444167
21	366968	46343015	46709983
22	787490	50030978	50818468
X	1337251	154703644	156040895
Y	95257	57132158	57227415
Total	35915410	3052354422	3088269832

References

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

El Soufi, K., Michel, C.J., 2014. Circular code motifs in the ribosome decoding center. Comput. Biol. Chem. 52, 9–17.

El Soufi, K., Michel, C.J., 2015. Circular code motifs near the ribosome decoding center. Comput. Biol. Chem. 59, 158–176.

Fimmel, E., Danielli, A., Strüngmann, L., 2013. On dichotomic classes and bijections of the genetic code. J. Theor. Biol. 336, 221–230.

Fimmel, E., Michel, C.J., Strüngmann, L., 2016. *n*-Nucleotide circular codes in graph theory. Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 374, 20150058.

Fimmel, E., Giannerini, S., Gonzalez, D.L., Strüngmann, L., 2014. Circular codes, symmetries and transformations. J. Math. Biol. 70, 1623–1644.

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

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

Michel, C.J., 2014. A genetic scale of reading frame coding. J. Theor. Biol. 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. J. Theor. Biol. 380, 156–177.

Michel, C.J., Pirillo, G., 2010. Identification of all trinucleotide circular codes. Comput. Biol. Chem. 34, 122–125.

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

Pirillo, G., 2003. A characterization for a set of trinucleotides to be a circular code. In: Pellegrini, C., Cerrai, P., Freguglia, P., Benci, V., Israel, G. (Eds.), Determinism, Holism, and Complexity. Kluwer Academic Publisher, New York, NY, USA.

Seligmann, H., 2013a. Systematic asymmetric nucleotide exchanges produce human mitochondrial RNAs cryptically encoding for overlapping protein coding genes. J. Theor. Biol. 324, 1–20.

Seligmann, H., 2013b. Polymerization of non-complementary RNA: systematic symmetric nucleotide exchanges mainly involving uracil produce mitochondrial RNA transcripts coding for cryptic overlapping genes. BioSystems 111, 156–174.