BioSystems 153 (2017) 45–62

Contents lists available at ScienceDirect

BioSystems

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

Unitary 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

ARTICLE INFO

Article history:
Received 4 November 2016
Received in revised form 9 January 2017
Accepted 13 February 2017
Available online 24 February 2017

Keywords:
Simple repeats
Repeated dinucleotides
Repeated trinucleotides
Repeated tetranucleotides
Low complexity DNA

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. The origin of this circular code X in genes is an open problem since its discovery in 1996. Here, we first show that the unitary circular codes (*UCC*), i.e. sets of one word, allow to generate unitary circular code motifs (*UCC* motifs), i.e. a concatenation of the same motif (simple repeats) leading to low complexity DNA. Three classes of *UCC* motifs are studied here: repeated dinucleotides (D^+ motifs), repeated trinucleotides (T^+ motifs) and repeated tetranucleotides ($\mathcal{T}^+$ motifs). Thus, the D^+, T^+ and $\mathcal{T}^+$ motifs allow to retrieve, synchronize and maintain a frame modulo 2, modulo 3 and modulo 4, respectively, and their shifted frames (1 modulo 2; 1 and 2 modulo 3; 1, 2 and 3 modulo 4 according to the C^2, C^3 and C^4 properties, respectively) in the DNA sequences. The statistical distribution of the D^+, T^+ and $\mathcal{T}^+$ motifs is analyzed in the genomes of eukaryotes. A *UCC* motif and its complementary *UCC* motif have the same distribution in the eukaryotic genomes. Furthermore, a *UCC* motif and its complementary *UCC* motif have increasing occurrences contrary to their number of hydrogen bonds, very significant with the T^+ motifs. The longest D^+, T^+ and $\mathcal{T}^+$ motifs in the studied eukaryotic genomes are also given. Surprisingly, a scarcity of repeated trinucleotides (T^+ motifs) in the large eukaryotic genomes is observed compared to the D^+ and $\mathcal{T}^+$ motifs. This result has been investigated and may be explained by two outcomes. Repeated trinucleotides (T^+ motifs) are identified in the X motifs of low composition (cardinality less than 10) in the genomes of eukaryotes. Furthermore, identical trinucleotide pairs of the circular code X are preferentially used in the gene sequences of eukaryotes. These two results suggest that the unitary circular codes of trinucleotides may have been involved in the formation of the trinucleotide circular code X. Indeed, repeated trinucleotides in the X motifs in the genomes of eukaryotes may represent an intermediary evolution from repeated trinucleotides of cardinality 1 (T^+ motifs) in the genomes of eukaryotes up to the X motifs of cardinality 20 in the gene sequences of eukaryotes.

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 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 two shifted frames (Arquès and Michel, 1996). The two sets X_1 and X_2 can be deduced from

* 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.biosystems.2017.02.001
0303-2647/

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 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). 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.

The origin of this circular code X in genes is an open problem since its discovery in 1996. However, this circular code concept observed in genes is also found in eukaryotic genomes. Indeed, the unitary circular codes (*UCC*), i.e. sets of one word, allow to generate unitary circular code motifs (*UCC* motifs), i.e. a concatenation of the same motif (simple repeats). Simple repeats have an important role in the non-coding DNA regions. For instance they are involved in forming a wide variety of unusual DNA structures with simple and complex loop-folding patterns (Rich et al., 1984). They are considered hot spots for recombination as well (Jeffreys et al., 1998). They are located within gene deserts, dominant in the telomeric and centromeric regions of a chromosome (Canapa et al., 2002) but also often occur within coding and regulatory regions (Gemayel et al., 2010). Among repeats found in coding sequences, repeated trinucleotides (and hexanucleotides, i.e. multiple of three nucleotides) are by far the most common. Furthermore, they are extremely unstable; mutation rates are often 10–100,000 times higher than average mutation rates in other parts of the genome. The longer and purer the repeat is, the more unstable it is. Mutation in repeats increases its evolutionary stability. We study here the *UCC* motifs in relation to the simple repeats in the eukaryotic genomes and to the X motifs (of the trinucleotide circular code X) in genes. Indeed, low-complexity sequences are theorized to play a role in the formation of new protein coding sequences (Ohno and Epplen, 1983; Li et al., 2004; Toll-Riera et al., 2012) and recently, we have identified X motifs in non-coding regions of eukaryotic genomes (El Soufi and Michel, 2016). Three classes of *UCC* motifs with repeated dinucleotides (D^+ motifs), repeated trinucleotides (T^+ motifs) and repeated tetranucleotides (T^{+} motifs) are analyzed in 126 available complete eukaryotic genomes containing 91,350,244,263 bases. The statistical distribution of the D^+, T^+ and T^{+} motifs in the eukaryotic genomes identifies new properties that are being related to the distribution of trinucleotide pairs (two consecutive trinucleotides) of the circular code X in the eukaryotic gene sequences.

2. Method

2.1. Recall

The definitions of complementary map $\mathcal{C}$, permutation map $\mathcal{P}$, circular code, unitary circular code, comma-free code, strong comma-free code, self-complementary circular code and C^3 circular code of trinucleotides are extended here to an m-nucleotide code, i.e. a code containing words at m letters, in order to identify new properties with the circular code X observed in genes.

Let the 4-letter alphabet $B = \{A, C, G, T\}$ (nucleotides or bases) be the genetic alphabet. For $m \in \mathbb{N}$ with $m \geq 2$, an m-nucleotide code is a subset $X \subseteq B^m$. The statistical analysis applied here to the eukaryotic genomes is related to m-nucleotide codes in B^m with $m \in \{2, 3, 4\}$. Precisely, $B^2 = \{AA, \ldots, TT\}$ is the set of the 16 words of length 2 (dinucleotides or diletters), $B^3 = \{AAA, \ldots, TTT\}$ is the set of the 64 words of length 3 (trinucleotides or triletters) and $B^4 = \{AAAA, \ldots, TTTT\}$ is the set of the 256 words of length 4 (tetranucleotides or tetraletters).

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

Definition 1. The *nucleotide complementarity map* $\mathcal{C} : B \to B$ 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 *m-nucleotide complementarity map* $\mathcal{C} : B^m \to B^m$ is defined by $\mathcal{C}(l_1 \ldots l_m) = \mathcal{C}(l_m) \ldots \mathcal{C}(l_1)$ for all $l_1, \ldots, l_m \in B$, e.g. $\mathcal{C}(ACG) = CGT$ on B^3 $(m = 3)$. By extension to an m-nucleotide code X, the *m-nucleotide code complementarity map* $\mathcal{C} : \mathbb{P}(B^m) \to \mathbb{P}(B^m)$, $\mathbb{P}$ being the set of all subsets of B^m, is defined by $\mathcal{C}(X) = \{v : u, v \in B^m, u \in X, v = \mathcal{C}(u)\}$, i.e. a complementary m-nucleotide code $\mathcal{C}(X)$ is obtained by applying the complementarity map $\mathcal{C}$ to all its m-nucleotides, e.g. $\mathcal{C}(\{ACG, AGT\}) = \{ACT, CGT\}$ on B^3.

Definition 2. The *m-nucleotide circular permutation map* $\mathcal{P} : B^m \to B^m$ is defined by $\mathcal{P}(l_1 l_2 \ldots l_m) = l_2 \ldots l_m l_1$ for all $l_1, l_2, \ldots, l_m \in B$, e.g. $\mathcal{P}(ACG) = CGA$ on B^3. The mth iterate of $\mathcal{P}$ is denoted $\mathcal{P}^m$, e.g. $\mathcal{P}^2(ACG) = GAC$ on B^3. By extension to an m-nucleotide code X, the *m-nucleotide code circular permutation map* $\mathcal{P} : \mathbb{P}(B^m) \to \mathbb{P}(B^m)$ is defined by $\mathcal{P}(X) = \{v : u, v \in B^m, u \in X, v = \mathcal{P}(u)\}$, i.e. a permuted m-nucleotide code $\mathcal{P}(X)$ is obtained by applying the circular permutation map $\mathcal{P}$ to all its m-nucleotides, e.g. $\mathcal{P}(\{ACG, AGT\}) = \{CGA, GTA\}$ and $\mathcal{P}^2(\{ACG, AGT\}) = \{GAC, TAG\}$ on B^3.

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). We briefly present here the circular codes with the most recent and powerful approach which relates a directed graph to any m-nucleotide code. We refer the reader to the mentioned articles for details that are outside the scope of this paper.

We recall the definition which associates a directed graph to any m-nucleotide code.

Definition 3. (Fimmel et al., 2016). Let $X \subseteq B^m$, $m \in \mathbb{N}$ with $m \geq 2$, be an m-nucleotide code. The directed graph $\mathcal{G}(X) = (V(X), E(X))$ associated with X has a set of vertices $V(X)$ and a set of edges $E(X)$ defined as follows:

$$\begin{cases} V(X) = \{N_1 \ldots N_i, N_{i+1} \ldots N_m : N_1 \ldots N_m \in \mathrm{X}, 1 \leq \mathrm{i} \leq \mathrm{m} - 1\} \\ E(X) = \{[N_1 \ldots N_i, N_{i+1} \ldots N_m] : N_1 \ldots N_m \in \mathrm{X}, 1 \leq \mathrm{i} \leq \mathrm{m} - 1\} \end{cases}.$$

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

In other words, a trinucleotide code $X \subseteq B^3$ is circular if any word over the alphabet B^3 written on a circle (the next letter after the last letter of the word being the first letter) has at most one decomposition (factorization) into words of X.

The theorem below gives a relation between an m-nucleotide code which is circular and its associated graph.

Theorem 1. (Fimmel et al., 2016). *Given an m-nucleotide code* $X \subseteq B^m$, $m \in \mathbb{N}$ *with* $m \geq 2$, *the following statements are equivalent:*

(1) The code X is circular.
(2) The graph $\mathcal{G}(X)$ is acyclic.

There are several varieties of circular codes, in particular the comma-free codes (Golomb et al., 1958a,b; Michel et al., 2008a,b, 2012; Michel and Pirillo, 2011).

Definition 5. A trinucleotide code $X \subseteq B^3$ is *comma-free* if for each $x \in X$ and $u, v \in B^*$ such that $uxv = y_1 \ldots y_n$ with $y_1, \ldots, y_n \in X, n \geq 1$, it holds that $u, v \in X^*$.

In other words, a trinucleotide code $X \subseteq B^3$ is comma-free if given any two trinucleotides $x_1, x_2 \in X$, any trinucleotide from the concatenation x_1x_2 and different from x_1 and x_2, does not belong to X. Comma-free codes are obviously circular but the converse is not true.

The theorem below gives a relation between an m-nucleotide code which is comma-free and its associated graph.

Theorem 2. (Fimmel et al., 2016). *Given an m-nucleotide code $X \subseteq B^m$, $m \in \mathbb{N}$ with $m \geq 2$, the following statements are equivalent:*

(1) The code X is comma-free.
(2) The maximal length of a path in $\mathcal{G}(X)$ is 2.

Recently, a new subclass of comma-free codes was identified by using the graph approach.

Definition 6. (Fimmel et al., 2017). A trinucleotide code $X \subseteq B^3$ is strong comma-free if its associated graph $\mathcal{G}(X)$ has only paths of length 1, i.e. the maximal length of a path in $\mathcal{G}(X)$ is 1.

Strong comma-free codes are obviously comma-free but the converse is not true.

Remark 1. From Definition 6, an m-nucleotide unitary circular code $\{l_1 \ldots l_m\}$ with $l_i \in B$, such that $l_1 = l_m$ with $m \geq 3$, i.e. starting and ending by the same letter, cannot be strong comma-free.

Definition 7. An m-nucleotide circular code $X \subseteq B^m$ is *self-complementary* if, for each $u \in X, \mathcal{C}(u) \in X$, i.e. $X = \mathcal{C}(X)$.

Definition 8. An m-nucleotide circular code $X \subseteq B^m$ is C^m if the m permuted m-nucleotide codes $X_1 = \mathcal{P}(X), \ldots, X_m = \mathcal{P}^m(X)$ are circular. An m-nucleotide comma-free $X \subseteq B^m$ (strong comma-free, respectively) is CF^m (SCF^m, respectively) if the m permuted m-nucleotide codes $X_1 = \mathcal{P}(X), \ldots, X_m = \mathcal{P}^m(X)$ are comma-free (strong comma-free, respectively).

Definition 9. An m-nucleotide unitary circular code (UCC) $X \subseteq B^m$ contains an unique word w of m-nucleotides (letters).

Definition 10. An m-nucleotide unitary circular code motif (UCC motif) generated by an m-nucleotide unitary circular code (UCC), is a concatenation of n motifs w, i.e. n times the single motif w noted $w^n = \underbrace{ww \ldots w}_{n}$ of cardinality $|w| = m$ nucleotides (letters). The class of the motifs w^n for all n is noted w^+.

Put another way, n is the number of repeats of a word w of length m nucleotides.

Definition 11. Two UCC motifs w_1^+ and w_2^+ are said equivalent if w_1^+ and w_2^+ are related by the circular permutation map $\mathcal{P}$ (Definition 2). By convention, the UCC motif studied is the 1st motif in lexicographical order in the equivalence class, i.e. w_1^+ with $w_1 < w_2$ and $A < C < G < T$.

Remark 2. An m-nucleotide code X containing either one periodic permuted m-nucleotide $P^m = \{A^m, C^m, G^m, T^m\}$ or two non-periodic permuted m-nucleotides $\{u, \mathcal{P}(u)\}$ for an m-nucleotide $u \in B^m \backslash P^m$ cannot be circular.

For simplification and without loss of generality, we recall here the main property of circular codes with the trinucleotide circular codes (on B^3). 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.

Fig. 1. Retrieval of the reading frame of the word $w = \ldots AGGTAATTACCAG \ldots$ constructed with 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 $\ldots A{\cdot}GGT{\cdot}AAT{\cdot}TAC{\cdot}CAG{\cdot}\ldots$. Thus, the first letter A of w is the 3rd letter of a trinucleotide of X.

Example 1. Suppose that the word $w = \ldots AGGTAATTACCAG \ldots$ has been constructed with the trinucleotide circular code X of Eq. (1) (Fig. 1). 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. 1). Thus, the unique factorization of w is $w_1 = \ldots A{\cdot}GGT{\cdot}AAT{\cdot}TAC{\cdot}CAG{\cdot}\ldots$. 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. 1). 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 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 $X = \mathcal{C}(X)$, Definition 7) 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, Definition 8 with $m = 3$). The trinucleotide circular codes X_1 and X_2 are related by the permutation map, i.e. $X_2 = \mathcal{P}(X_1)$, and by the complementary map, i.e. $X_1 = \mathcal{C}(X_2)$ and $X_2 = \mathcal{C}(X_1)$ (Bussoli et al., 2012).

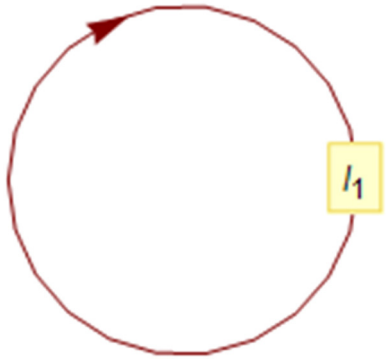

Fig. 2. A dinucleotide unitary code $D = \{l_1 l_1\}$ with $l_1 \in B$ is not circular as its associated graph $\mathcal{G}(D)$ is cyclic.

Fig. 3. A trinucleotide unitary code $T = \{l_1 l_1 l_1\}$ with $l_1 \in B$ is not circular as its associated graph $\mathcal{G}(T)$ is cyclic.

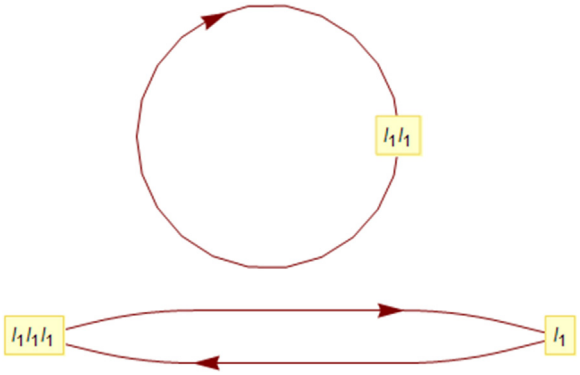

Fig. 4. A tetranucleotide unitary code $\mathcal{T} = \{l_1 l_1 l_1 l_1\}$ with $l_1 \in B$ is not circular as its associated graph $\mathcal{G}(\mathcal{T})$ is cyclic.

The statistical distribution of three classes of unitary circular code motifs (*UCC* motifs) with repeated dinucleotides (D^+ motifs), repeated trinucleotides (T^+ motifs) and repeated tetranucleotides ($\mathcal{T}^+$ motifs) is analyzed in the genomes of eukaryotes. A fundamental property of the *UCC* motifs which has been largely ignored from a theoretical point of view, is the fact that the *UCC* motifs are generated by *m*-nucleotide unitary circular codes (*UCC*, Definition 9). Thus, the D^+, T^+ and $\mathcal{T}^+$ motifs allow to retrieve, synchronize and maintain a frame modulo 2, modulo 3 and modulo 4, respectively, in the DNA sequences. In the section below, we will show that the studied *m*-nucleotide unitary circular codes ($m \in \{2, 3, 4\}$) are either comma-free (Definition 5 and Theorem 2) or strong comma-free (Definition 6).

2.2. Unitary circular codes of dinucleotides, trinucleotides and tetranucleotides

The four dinucleotide (2-nucleotide) unitary codes $D = \{l_1 l_1\}$ with $l_1 \in B$ are obviously not circular (Fig. 2).

The four trinucleotide (3-nucleotide) unitary codes $T = \{l_1 l_1 l_1\}$ with $l_1 \in B$ are obviously not circular (Fig. 3).

The four tetranucleotide (4-nucleotide) unitary codes $\mathcal{T} = \{l_1 l_1 l_1 l_1\}$ with $l_1 \in B$ are obviously not circular (Fig. 4).

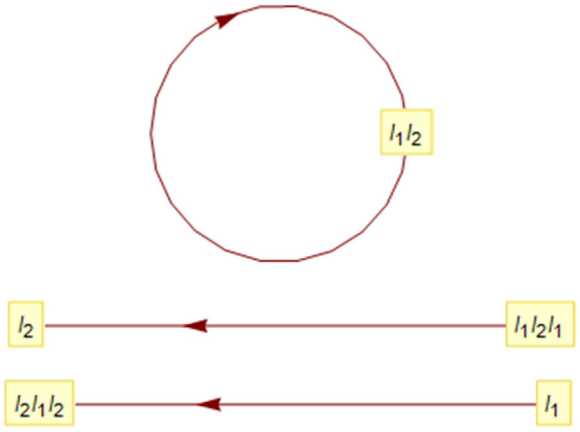

Fig. 5. A tetranucleotide unitary code $\mathcal{T} = \{l_1 l_2 l_1 l_2\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ is not circular as its associated graph $\mathcal{G}(\mathcal{T})$ is cyclic.

Fig. 6. A dinucleotide unitary code $D = \{l_1 l_2\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ is circular as its associated graph $\mathcal{G}(D)$ is acyclic and in addition strong comma-free as the unique path length in $\mathcal{G}(D)$ is 1.

The 12 tetranucleotide (4-nucleotide) unitary codes $\mathcal{T} = \{l_1 l_2 l_1 l_2\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ are also not circular (Fig. 5).

We describe some additional combinatorial properties for the unitary codes *D*, *T* and $\mathcal{T}$ which are circular.

2.2.1. Unitary circular codes of dinucleotides

All the dinucleotide unitary codes which are circular are also C^2 (Definition 8).

The 12 dinucleotide unitary codes $\{l_1 l_2\}$ and $\{\mathcal{P}(l_1 l_2)\} = \{l_2 l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$, i.e. $\{AC\}$, $\{AG\}$, $\{AT\}$, $\{CA\}$, $\{CG\}$, $\{CT\}$, $\{GA\}$, $\{GC\}$, $\{GT\}$, $\{TA\}$, $\{TC\}$ and $\{TG\}$, are circular and in addition strong comma-free (*SCF*) by Definition 6 (Fig. 6).

Thus, a dinucleotide unitary strong comma-free code $\{l_1 l_2\}$ has a permuted code $\{l_2 l_1\}$ which is also strong comma-free (SCF^2 property, see Definition 8).

Furthermore, the four dinucleotide unitary strong comma-free codes $\{AT\}$, $\{CG\}$, $\{GC\}$ and $\{TA\}$ are self-complementary.

2.2.2. Unitary circular codes of trinucleotides

All the trinucleotide unitary codes which are circular are also C^3 (Definition 8).

The 24 trinucleotide unitary codes $\{l_1 l_2 l_3\}$, $\{\mathcal{P}(l_1 l_2 l_3)\} = \{l_2 l_3 l_1\}$ and $\{\mathcal{P}^2(l_1 l_2 l_3)\} = \{l_3 l_1 l_2\}$ with $l_1, l_2, l_3 \in B$ and $l_1 \neq l_2 \neq l_3$, i.e. $\{ACG\}$, $\{ACT\}$, $\{AGC\}$, $\{AGT\}$, $\{ATC\}$, $\{ATG\}$, $\{CAG\}$, $\{CAT\}$, $\{CGA\}$, $\{CGT\}$, $\{CTA\}$, $\{CTG\}$, $\{GAC\}$, $\{GAT\}$, $\{GCA\}$, $\{GCT\}$, $\{GTA\}$, $\{GTC\}$, $\{TAC\}$, $\{TAG\}$, $\{TCA\}$, $\{TCG\}$, $\{TGA\}$ and $\{TGC\}$, are circular and in addition strong comma-free (*SCF*) by Definition 6 (Fig. 7).

Thus, a trinucleotide unitary strong comma-free code $\{l_1 l_2 l_3\}$ has two permuted codes $\{l_2 l_3 l_1\}$ and $\{l_3 l_1 l_2\}$ which are also strong comma-free (SCF^3 property, see Definition 8).

The 24 trinucleotide unitary codes $\{l_1 l_1 l_2\}$ and $\{\mathcal{P}^2(l_1 l_1 l_2)\} = \{l_2 l_1 l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$, i.e. $\{AAC\}$, $\{AAG\}$, $\{AAT\}$, $\{ACC\}$, $\{AGG\}$, $\{ATT\}$, $\{CAA\}$, $\{CCA\}$, $\{CCG\}$, $\{CCT\}$, $\{CGG\}$, $\{CTT\}$, $\{GAA\}$, $\{GCC\}$, $\{GGA\}$, $\{GGC\}$, $\{GGT\}$, $\{GTT\}$, $\{TAA\}$,

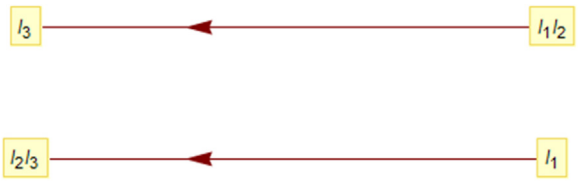

Fig. 7. A trinucleotide unitary code $T = \{l_1l_2l_3\}$ with $l_1, l_2, l_3 \in B$ and $l_1 \neq l_2 \neq l_3$ is circular as its associated graph $\mathcal{G}(T)$ is acyclic and in addition strong comma-free as the two path lengths in $\mathcal{G}(T)$ are 1.

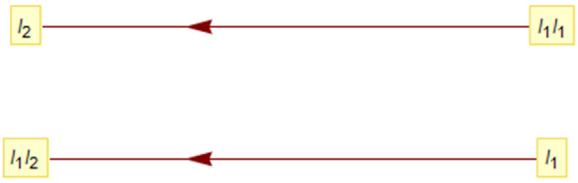

Fig. 8. A trinucleotide unitary code $T = \{l_1l_1l_2\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ is circular as its associated graph $\mathcal{G}(T)$ is acyclic and in addition strong comma-free as the two path lengths in $\mathcal{G}(T)$ are 1.

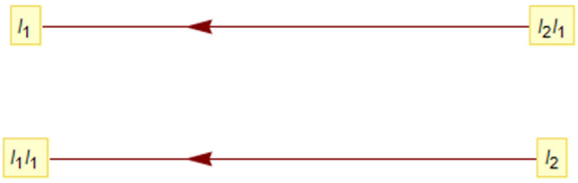

Fig. 9. A trinucleotide unitary code $T = \{l_2l_1l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ is circular as its associated graph $\mathcal{G}(T)$ is acyclic and in addition strong comma-free as the two path lengths in $\mathcal{G}(T)$ are 1.

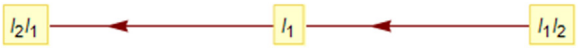

Fig. 10. A trinucleotide unitary code $T = \{l_1l_2l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ is circular as its associated graph $\mathcal{G}(T)$ is acyclic and in addition comma-free as the unique path length in $\mathcal{G}(T)$ is 2.

$\{TCC\}$, $\{TGG\}$, $\{TTA\}$, $\{TTC\}$ and $\{TTG\}$, are circular and in addition strong comma-free (SCF) by Definition 6 (Figs. 8 and 9).

The 12 trinucleotide unitary codes $\{\mathcal{P}(l_1l_1l_2)\} = \{l_1l_2l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$, i.e. $\{ACA\}$, $\{AGA\}$, $\{ATA\}$, $\{CAC\}$, $\{CGC\}$, $\{CTC\}$, $\{GAG\}$, $\{GCG\}$, $\{GTG\}$, $\{TAT\}$, $\{TCT\}$ and $\{TGT\}$, are circular and in addition comma-free (CF) by Definition 5 and Theorem 2 (Fig. 10).

Thus, a trinucleotide unitary strong comma-free code $\{l_1l_1l_2\}$ has one permuted code $\{l_2l_1l_1\}$ which is also strong comma-free and one permuted code $\{l_1l_2l_1\}$ which is comma-free code. Corollary, a trinucleotide unitary comma-free code $\{l_1l_2l_1\}$ has two permuted codes $\{l_1l_1l_2\}$ and $\{l_2l_1l_1\}$ which are strong comma-free. Obviously, there is no self-complementary trinucleotide unitary circular code.

2.2.3. Unitary circular codes of tetranucleotides

All the tetranucleotide unitary codes which are circular are also C^4 (Definition 8).

The 24 tetranucleotide unitary codes $\{l_1l_2l_3l_4\}$, $\{\mathcal{P}(l_1l_2l_3l_4)\} = \{l_2l_3l_4l_1\}$, $\{\mathcal{P}^2(l_1l_2l_3l_4)\} = \{l_3l_4l_1l_2\}$ and $\{\mathcal{P}^3(l_1l_2l_3l_4)\} = \{l_4l_1l_2l_3\}$ with $l_1, l_2, l_3, l_4 \in B$ and $l_1 \neq l_2 \neq l_3 \neq l_4$ are circular and in addition strong comma-free (SCF) by Definition 6 (Fig. 11).

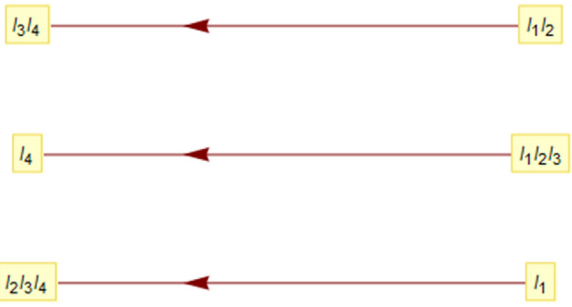

Fig. 11. A tetranucleotide unitary code $\mathcal{T} = \{l_1l_2l_3l_4\}$ with $l_1, l_2, l_3, l_4 \in B$ and $l_1 \neq l_2 \neq l_3 \neq l_4$ is circular as its associated graph $\mathcal{G}(\mathcal{T})$ is acyclic and in addition strong comma-free as the three path lengths in $\mathcal{G}(\mathcal{T})$ are 1.

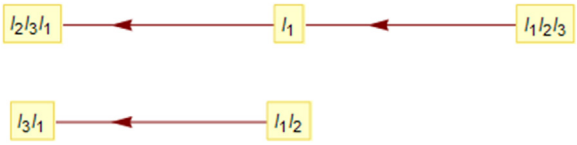

Fig. 12. A tetranucleotide unitary code $\mathcal{T} = \{l_1l_2l_3l_1\}$ with $l_1, l_2, l_3 \in B$ and $l_1 \neq l_2 \neq l_3$ is circular as its associated graph $\mathcal{G}(\mathcal{T})$ is acyclic and in addition comma-free as one path length in $\mathcal{G}(\mathcal{T})$ is 2.

Thus, a tetranucleotide unitary strong comma-free code $\{l_1l_2l_3l_4\}$ has three permuted codes $\{l_2l_3l_4l_1\}$, $\{l_3l_4l_1l_2\}$ and $\{l_4l_1l_2l_3\}$ which are also strong comma-free (SCF^4 property, see Definition 8).

The 48 tetranucleotide unitary codes $\{l_1l_2l_1l_3\}$, $\{\mathcal{P}(l_1l_2l_1l_3)\} = \{l_2l_1l_3l_1\}$, $\{\mathcal{P}^2(l_1l_2l_1l_3)\} = \{l_1l_3l_1l_2\}$ and $\{\mathcal{P}^3(l_1l_2l_1l_3)\} = \{l_3l_1l_2l_1\}$ with $l_1, l_2, l_3 \in B$ and $l_1 \neq l_2 \neq l_3$ are circular and in addition strong comma-free (SCF) by Definition 6 (figure not shown). Thus, a tetranucleotide unitary strong comma-free code $\{l_1l_2l_1l_3\}$ has three permuted codes $\{l_2l_1l_3l_1\}$, $\{l_1l_3l_1l_2\}$ and $\{l_3l_1l_2l_1\}$ which are also strong comma-free (SCF^4 property).

The 72 tetranucleotide unitary codes $\{l_1l_1l_2l_3\}$, $\{\mathcal{P}^2(l_1l_1l_2l_3)\} = \{l_2l_3l_1l_1\}$ and $\{\mathcal{P}^3(l_1l_1l_2l_3)\} = \{l_3l_1l_1l_2\}$ with $l_1, l_2, l_3 \in B$ and $l_1 \neq l_2 \neq l_3$ are circular and in addition strong comma-free (SCF) by Definition 6 (figure not shown). The 24 tetranucleotide unitary codes $\{\mathcal{P}(l_1l_1l_2l_3)\} = \{l_1l_2l_3l_1\}$ with $l_1, l_2, l_3 \in B$ and $l_1 \neq l_2 \neq l_3$ are circular and in addition comma-free (CF) by Definition 5 and Theorem 2 (Fig. 12).

Thus, a tetranucleotide unitary strong comma-free code $\{l_1l_1l_2l_3\}$ has two permuted codes $\{l_2l_3l_1l_1\}$ and $\{l_3l_1l_1l_2\}$ which are also strong comma-free and one permuted code $\{l_1l_2l_3l_1\}$ which is comma-free. Corollary, a tetranucleotide unitary comma-free code $\{l_1l_2l_3l_1\}$ has three permuted codes $\{l_1l_1l_2l_3\}$, $\{l_2l_3l_1l_1\}$ and $\{l_3l_1l_1l_2\}$ which are strong comma-free.

The 24 tetranucleotide unitary codes $\{l_1l_1l_1l_2\}$ and $\{\mathcal{P}^3(l_1l_1l_1l_2)\} = \{l_2l_1l_1l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ are circular and in addition strong comma-free (SCF) by Definition 6 (figure not shown). The 24 tetranucleotide unitary codes $\{\mathcal{P}(l_1l_1l_1l_2)\} = \{l_1l_1l_2l_1\}$ and $\{\mathcal{P}^2(l_1l_1l_1l_2)\} = \{l_1l_2l_1l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ are circular and in addition comma-free (CF) by Definition 5 and Theorem 2 (figure not shown). Thus, a tetranucleotide unitary strong comma-free code $\{l_1l_1l_1l_2\}$ has one permuted code $\{l_2l_1l_1l_1\}$ which is also strong comma-free and two permuted codes $\{l_1l_1l_2l_1\}$ and $\{l_1l_2l_1l_1\}$ which are comma-free. Corollary, a tetranucleotide unitary comma-free code $\{l_1l_1l_2l_1\}$ has one permuted code $\{l_1l_2l_1l_1\}$ which is also comma-free and

two permuted codes $\{l_1l_1l_1l_2\}$ and $\{l_2l_1l_1l_1\}$ which are strong comma-free.

The 12 tetranucleotide unitary codes $\{l_1l_1l_2l_2\}$ and $\{\mathcal{P}^2(l_1l_1l_2l_2)\} = \{l_2l_2l_1l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ are circular and in addition strong comma-free (*SCF*) by Definition 6 (figure not shown). The 12 tetranucleotide unitary codes $\{\mathcal{P}(l_1l_1l_2l_2)\} = \{l_1l_2l_2l_1\}$ and $\{\mathcal{P}^3(l_1l_1l_2l_2)\} = \{l_2l_1l_1l_2\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ are circular and in addition comma-free (*CF*) by Definition 5 and Theorem 2 (figure not shown). Thus, a tetranucleotide unitary strong comma-free code $\{l_1l_1l_2l_2\}$ has one permuted code $\{l_2l_2l_1l_1\}$ which is also strong comma-free and two permuted codes $\{l_1l_2l_2l_1\}$ and $\{l_2l_1l_1l_2\}$ which are comma-free. Corollary, a tetranucleotide unitary comma-free code $\{l_1l_2l_2l_1\}$ has one permuted code $\{l_2l_1l_1l_2\}$ which is also comma-free and two permuted codes $\{l_1l_1l_2l_2\}$ and $\{l_2l_2l_1l_1\}$ which are strong comma-free.

Furthermore, the 12 tetranucleotide unitary strong comma-free codes $\{AATT\}$, $\{ACGT\}$, $\{AGCT\}$, $\{CATG\}$, $\{CCGG\}$, $\{CTAG\}$, $\{GATC\}$, $\{GGCC\}$, $\{GTAC\}$, $\{TCGA\}$, $\{TGCA\}$ and $\{TTAA\}$ are self-complementary. There is no tetranucleotide unitary comma-free code which is self-complementary.

Remark 3. We describe here the combinatorial properties of the 64 tetranucleotides which were associated to the mitochondrial genetic code (Table 10 in Gonzalez et al., 2012). The four codes $\{AAAA\}$, $\{CCCC\}$, $\{GGGG\}$ and $\{TTTT\}$ belonging to the class $\{l_1l_1l_1l_1\}$ are not circular (Fig. 4); the 12 codes $\{ACAC\}$, $\{AGAG\}$, $\{ATAT\}$, $\{CACA\}$, $\{CGCG\}$, $\{CTCT\}$, $\{GAGA\}$, $\{GCGC\}$, $\{GTGT\}$, $\{TATA\}$, $\{TCTC\}$ and $\{TGTG\}$ belonging to the class $\{l_1l_2l_1l_2\}$ are not circular (Fig. 5); the 24 codes $\{ACGT\}$, $\{ACTG\}$, $\{AGCT\}$, $\{AGTC\}$, $\{ATCG\}$, $\{ATGC\}$, $\{CAGT\}$, $\{CATG\}$, $\{CGAT\}$, $\{CGTA\}$, $\{CTAG\}$, $\{CTGA\}$, $\{GACT\}$, $\{GATC\}$, $\{GCAT\}$, $\{GCTA\}$, $\{GTAC\}$, $\{GTCA\}$, $\{TACG\}$, $\{TAGC\}$, $\{TCAG\}$, $\{TCGA\}$, $\{TGAC\}$ and $\{TGCA\}$ belonging to the class $\{l_1l_2l_3l_4\}$ are strong comma-free (Fig. 11); the 12 codes $\{AACC\}$, $\{AAGG\}$, $\{AATT\}$, $\{CCAA\}$, $\{CCGG\}$, $\{CCTT\}$, $\{GGAA\}$, $\{GGCC\}$, $\{GGTT\}$, $\{TTAA\}$, $\{TTCC\}$ and $\{TTGG\}$ belonging to the class $\{l_1l_1l_2l_2\}$ are strong comma-free (figure not shown); and the 12 codes $\{ACCA\}$, $\{AGGA\}$, $\{ATTA\}$, $\{CAAC\}$, $\{CGGC\}$, $\{CTTC\}$, $\{GAAG\}$, $\{GCCG\}$, $\{GTTG\}$, $\{TAAT\}$, $\{TCCT\}$ and $\{TGGT\}$ belonging to the class $\{l_1l_2l_2l_1\}$ are comma-free (figure not shown).

2.3. Definition of unitary circular code motifs

The unitary circular code motifs (*UCC* motifs) are generated from the unitary circular codes (*UCC*). They are defined by two parameters: their equivalence class and their length in nucleotides.

2.3.1. Dinucleotide unitary circular code motifs

A repeated dinucleotide $d^+ = (l_1l_2)^+$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ belongs to one of the $(16-4)/2 = 6$ equivalence classes $\{(l_1l_2)^+, (l_2l_1)^+\}$ (Definition 11): $\{(AC)^+, (CA)^+\}$, $\{(AG)^+, (GA)^+\}$, $\{(AT)^+, (TA)^+\}$, $\{(CG)^+, (GC)^+\}$, $\{(CT)^+, (TC)^+\}$ and $\{(GT)^+, (TG)^+\}$. By convention, the six dinucleotide *UCC* motifs D^+ are defined by the repeated dinucleotides d^+ which are the 1st repeated motifs in lexicographical order in each equivalence class:

$$D^+ = \{(AC)^+, (AG)^+, (AT)^+, (CG)^+, (CT)^+, (GT)^+\}. \quad (4)$$

The repeated dinucleotides d^n studied have lengths $l = n \times |d| \geq 30$ nucleotides ($|d|$ being the number of letters of d), i.e. $n \geq 15$.

2.3.2. Trinucleotide unitary circular code motifs

A repeated trinucleotide $t^+ = (l_1l_2l_3)^+$ with $l_1, l_2, l_3 \in B$ and $l_1l_2 \neq l_2l_3$ belongs to one of the $(64-4)/3 = 20$ equivalence classes $\{(l_1l_2l_3)^+, (l_2l_3l_1)^+, (l_3l_1l_2)^+\}$ (Definition 11): $\{(AAC)^+, (ACA)^+, (CAA)^+\}, \ldots, \{(GTT)^+, (TTG)^+, (TGT)^+\}$. Similarly by convention, the 20 trinucleotide *UCC* motifs T^+ are defined by:

$$\begin{aligned} T^+ = \{&(AAC)^+, (AAG)^+, (AAT)^+, (ACC)^+, (ACG)^+, (ACT)^+, \\ &(AGC)^+, (AGG)^+, (AGT)^+, (ATC)^+, (ATG)^+, (ATT)^+, (CCG)^+, \\ &(CCT)^+, (CGG)^+, (CGT)^+, (CTG)^+, (CTT)^+, (GGT)^+, (GTT)^+\}. \end{aligned} \quad (5)$$

The repeated trinucleotides t^n studied have lengths $l = n \times |t| \geq 30$ nucleotides, i.e. $n \geq 10$.

2.3.3. Tetranucleotide unitary circular code motifs

A repeated tetranucleotide $t^+ = (l_1l_2l_3l_4)^+$ with $l_1, l_2, l_3, l_4 \in B$ and $l_1l_2 \neq l_3l_4$ belongs to one of the $(256-4-12)/4 = 60$ equivalence classes $\{(l_1l_2l_3l_4)^+, (l_2l_3l_4l_1)^+, (l_3l_4l_1l_2)^+, (l_4l_1l_2l_3)^+\}$ (Definition 11). Similarly by convention, the 60 tetranucleotide *UCC* motifs $\mathcal{T}^+$ are defined by:

$$\mathcal{T}^+ = \{(AAAC)^+, \ldots, (GTTT)^+\}. \quad (6)$$

The repeated tetranucleotides t^n studied have lengths $l = n \times |t| \geq 28$ nucleotides, i.e. $n \geq 7$.

Remark 4. Among the 64 tetranucleotides which were associated to the mitochondrial genetic code (Gonzalez et al., 2012), only 48 tetranucleotides are *UCC* codes (Remark 3). These 48 *UCC* codes are associated to the 12 following repeated tetranucleotides $(ACGT)^+$, $(ACTG)^+$, $(AGCT)^+$, $(AGTC)^+$, $(ATCG)^+$, $(ATGC)^+$, $(AACC)^+$, $(AAGG)^+$, $(AATT)^+$, $(CCGG)^+$, $(CCTT)^+$ and $(GGTT)^+$.

2.4. Occurrence number of an unitary circular code motif in the eukaryotic genomes

Let $r^n \in \{d^n, t^n, t^n\}$ be a repeated motif r of nucleotide length $n \times |r|$ ($|r| \in \{2, 3, 4\}$ being the number of letters of r) where $r = d$ for a repeated dinucleotide d^n, $r = t$ for a repeated trinucleotide t^n and $r = t$ for a repeated tetranucleotide t^n. The number $N(r^n, \mathcal{G})$ counts the occurrences of a repeated motif r^n for a given number n in a eukaryotic genome $\mathcal{G}$. Then, the occurrence number $N(r^+)$ of a repeated motif $r^+ \in \{d^+, t^+, t^+\}$ for all the genomes of eukaryotes $\mathbb{E}$ is obtained by summing for all $\mathcal{G}$ in $\mathbb{E}$ and for all n

$$N(r^+) = \sum_{\mathcal{G} \in \mathbb{E}} \sum_{n} N(r^n, \mathcal{G}) \quad (7)$$

with $n \geq 15$ for computing $N(d^+)$ of a repeated dinucleotide $d^+ \in D^+$ (Eq. (4)), $n \geq 10$ for computing $N(t^+)$ of a repeated trinucleotide $t^+ \in T^+$ (Eq. (5)) and $n \geq 7$ for computing $N(t^+)$ of a repeated tetranucleotide $t^+ \in \mathcal{T}^+$ (Eq. (6)). These occurrence numbers $N(d^+)$, $N(t^+)$ and $N(t^+)$ are computed in the eukaryotic genomes according to the following algorithm.

The algorithm searches for repeated motifs in a DNA sequence such that their lengths are greater than or equal to the parameter minsize and returns a frequency map for the association of a word and how many times it was repeated. The algorithm is adaptive in regards to the input set as a parameter. It determines the number of frames required with respect to word length (two frames for dinucleotides, three frames for trinucleotides and four frames for tetranucleotides). This approach allows us to retrieve all the repeated motifs without the issue of overlaps between different frames.

```
1.  Read sequence
2.  INIT Y AS a set of words
3.  INIT minsize AS the minimum number of words in a repeats motif
4.  INIT wordsize from Y
5.  INIT mapFreq AS a map using the association of a word and number
      of repeats as key with their frequency as value
6.  FOR EACH frame in wordsize
7.    INIT wordCurrent AS empty
8.    INIT streak AS 0, number of successive wordCurrent
9.    FOR EACH word in sequence starting from frame AS wordSeq
10.     IF Y contains wordSeq THEN
11.       IF wordCurrent equals wordSeq THEN
12.               increment streak by 1
13.       ELSE
14.         IF streak is greater than or equal to minsize THEN
15.           increment the frequency of wordCurrent with streak in
                mapFreq by 1
16.         ENDIF
17.         INIT wordCurrent As wordSeq
18.         INIT streak AS 1
19.       ENDIF
20.     ELSE
21.       IF streak is greater than or equal to minsize THEN
22.         increment the frequency of wordCurrent with streak in
              mapFreq by 1
23.       ENDIF
24.       INIT wordCurrent As empty
25.       INIT streak AS 0
26.     ENDIF
27.   ENDFOR
28. ENDFOR
```

Example 2. If the trinucleotide $t = AAC$ occurs with two repeats t^{n_1} with $n_1 = 10$ in the genome $\mathcal{G}_1$, i.e. $N\left(t^{10}, \mathcal{G}_1\right) = 2$, and three repeats t^{n_2} with $n_2 = 20$ in the genome $\mathcal{G}_2$, i.e. $N\left(t^{20}, \mathcal{G}_2\right) = 3$, then the occurrence number $N(t^+)$ of the repeated trinucleotide $(AAC)^+$ in the genomes of eukaryotes $\mathbb{E}$ is equal to $N(t^+) = N\left(t^{10}, \mathcal{G}_1\right) + N\left(t^{20}, \mathcal{G}_2\right) = 2 + 3 = 5$ (Eq. (7)).

2.5. Base number of an unitary circular code motif in the eukaryotic genomes

The base number $B(r^+)$ of a repeated motif $r^+ \in \left\{d^+, t^+, t^+\right\}$ for all the genomes of eukaryotes $\mathbb{E}$ is

$$B\left(r^+\right) = |r| \sum_{\mathcal{G} \in \mathbb{E}} \sum_{n} N\left(r^n, \mathcal{G}\right) \times n \tag{8}$$

where $N(r^n, \mathcal{G})$ is defined in Section 2.4 for Eq. (7) and with $n \geq 15$ for computing $B(d^+)$ of a repeated dinucleotide $d^+ \in D^+$ (Eq. (4)), $n \geq 10$ for computing $B\left(t^+\right)$ of a repeated trinucleotide $t^+ \in T^+$ (Eq. (5)) and $n \geq 7$ for computing $B(t^+)$ of a repeated tetranucleotide $t^+ \in \mathcal{T}^+$ (Eq. (6)), $|r| \in \left\{2, 3, 4\right\}$ being the number of letters of r.

Example 3. If the trinucleotide $t = AAC$ occurs with two repeats t^{n_1} with $n_1 = 10$ in the genome $\mathcal{G}_1$, i.e. $N(t^{n_1}, \mathcal{G}_1) = 2$, and three repeats t^{n_2} with $n_2 = 20$ in the genome $\mathcal{G}_2$, i.e. $N(t^{n_2}, \mathcal{G}_2) = 3$, then the base number $B(t^+)$ of the repeated trinucleotide $(AAC)^+$ in the genomes of eukaryotes $\mathbb{E}$ is equal to $B(t^+) = |t|(N(t^{n_1}, \mathcal{G}_1) \times n_1 + N(t^{n_2}, \mathcal{G}_2) \times n_2) = 3(2 \times 10 + 3 \times 20) = 240$ (Eq. (8)).

2.6. Total base number of unitary circular code motifs in the eukaryotic genomes

The total base number $B(R^+, \mathcal{G})$ of all the repeated motifs r^+ in $R^+ \in \left\{D^+, T^+, \mathcal{T}^+\right\}$ (Eqs. (4), (5) and (6)) for a eukaryotic genome $\mathcal{G}$ is

$$B\left(R^+, \mathcal{G}\right) = |r| \sum_{r^+ \in R^+} \sum_{n} N\left(r^n, \mathcal{G}\right) \times n \tag{9}$$

where $N(r^n, \mathcal{G})$ is defined in Section 2.4 for Eq. (7) and with $n \geq 15$ for computing $B(D^+, \mathcal{G})$ of the six repeated dinucleotides d^+ in D^+ (Eq. (4)), $n \geq 10$ for computing $B(T^+, \mathcal{G})$ of the 20 repeated trinucleotides t^+ in T^+ (Eq. (5)) and $n \geq 7$ for computing $B(\mathcal{T}^+, \mathcal{G})$ of the 60 repeated tetranucleotides t^+ in $\mathcal{T}^+$ (Eq. (6)), $|r| \in \left\{2, 3, 4\right\}$ being the number of letters of r.

Example 4. If the trinucleotide $t_1 = AAC$ occurs with two repeats $t_1^{n_1}$ with $n_1 = 10$ and three repeats $t_1^{n_2}$ with $n_2 = 20$ in a genome $\mathcal{G}$, i.e. $N((AAC)^{n_1}, \mathcal{G}) = 2$ and $N((AAC)^{n_2}, \mathcal{G}) = 3$, and if the trinucleotide $t_2 = AAG$ occurs with four repeats $t_2^{n_3}$ with $n_3 = 30$ in the same genome $\mathcal{G}$, i.e. $N((AAG)^{n_3}, \mathcal{G}) = 3$, then the total base number $B(T^+, \mathcal{G})$ of all the repeated motifs T^+ in the genome $\mathcal{G}$ is equal to $B(T^+, \mathcal{G}) = |t|(N((AAC)^{n_1}, \mathcal{G}) \times n_1 + N((AAC)^{n_2}, \mathcal{G}) \times n_2 + N((AAG)^{n_3}, \mathcal{G}) \times n_3) = 3(2 \times 10 + 3 \times 20 + 4 \times 30) = 600$ (Eq. (9)).

In order to normalize the total base number $B(R^+, \mathcal{G})$ (Eq. (9)) for eukaryotic genomes of different sizes, the ratio $r(R^+, \mathcal{G})$ giving the proportion of the total base number $B(R^+, \mathcal{G})$ of all the repeated motifs r^+ in R^+ in a eukaryotic genome $\mathcal{G}$ of size $N(\mathcal{G})$ in bases (given in Appendix A) is defined by

$$r\left(R^+, \mathcal{G}\right) = \frac{B(R^+, \mathcal{G})}{N(\mathcal{G})}. \tag{10}$$

Finally, $\bar{r}(R^+)$ is the mean of the ratios $r(R^+, \mathcal{G})$ for all the genomes of eukaryotes $\mathbb{E}$

$$\bar{r}\left(R^+\right) = \frac{1}{|\mathbb{E}|} \sum_{\mathcal{G} \in \mathbb{E}} r\left(R^+, \mathcal{G}\right) \tag{11}$$

where $|\mathbb{E}|$ is the number of genomes $\mathcal{G}$ in $\mathbb{E}$ ($|\mathbb{E}| = 126$ here) and

$\tilde{r}\left(R^+\right)$ is the median of the ratios $r\left(R^+, \mathcal{G}\right)$ for all the genomes of eukaryotes $\mathbb{E}$. (12)

2.7. Occurrence number of trinucleotide pairs in the eukaryotic gene sequences

In order to identify a new property of the circular code X, we study the occurrence of trinucleotide pairs $tt' \in B^6$ where $t, t' \in B^3$ ($|B^6| = 4096$ motifs tt' of two consecutive trinucleotides) in the eukaryotic gene sequences. Among these 4096 motifs, there are 400 trinucleotide pairs $tt' \in X^2$ associated to the circular code X (cardinality of 20 trinucleotides).

The number $N(tt', \mathcal{G}_{GS})$ counts the occurrences of a trinucleotide pair $tt' \in B^6$ in the gene sequences $\mathcal{G}_{GS}$ of a eukaryotic genome $\mathcal{G}$. Note that when $t = t'$, the trinucleotide pair tt is a repeated trinucleotide t^n with $n = 2$ (see Section 2.4) leading to $N(tt, \mathcal{G}_{GS}) = N\left(t^2, \mathcal{G}_{GS}\right)$. Then, the occurrence number $N(tt')$ of a trinucleotide pair $tt' \in B^6$ in all the gene sequences of eukaryotes $\mathbb{E}$ is

$$N\left(tt'\right) = \sum_{\mathcal{G}_{GS} \in \mathbb{E}} N\left(tt', \mathcal{G}_{GS}\right). \tag{13}$$

The observed probability $P(tt')$ of a trinucleotide pair $tt' \in B^6$ in all the gene sequences of $\mathbb{E}$ is

$$P\left(tt'\right) = \frac{N(tt')}{\sum_{tt' \in B^6} N(tt')}. \tag{14}$$

Due to the codon usage, in particular, this probability $P(tt')$ must be normalized. The observed probability $P(t)$ of a trinucleotide $t \in B^3$ in all the gene sequences of $\mathbb{E}$ is

$$P(t) = \frac{N(t)}{\sum_{t \in B^3} N(t)} \tag{15}$$

with $N(t) = \sum_{\mathcal{G}_{GS} \in \mathbb{E}} N(t, \mathcal{G}_{GS})$ where $N(t, \mathcal{G}_{GS})$ (a repeated trinucleotide t^n with $n = 1$) is the occurrence number of t in the gene sequences $\mathcal{G}_{GS}$ of a eukaryotic genome $\mathcal{G}$. By taking the hypothesis of independent events then the estimated theoretical probability $\hat{P}(tt')$ of a trinucleotide pair $tt' \in B^6$ in all the gene sequences of $\mathbb{E}$ is

$$\hat{P}\left(tt'\right) = P(t) \times P\left(t'\right). \tag{16}$$

Finally, the observed/theoretical ratio $r(tt')$ of a trinucleotide pair $tt' \in B^6$ in all the gene sequences of eukaryotes $\mathbb{E}$ is equal to

$$r\left(tt'\right) = \frac{P(tt')}{\hat{P}(tt')}. \tag{17}$$

Two other ratios also analyze the occurrence of trinucleotide pairs in the eukaryotic gene sequences.

The observed probability $P(tt', \mathcal{G}_{GS})$ of a trinucleotide pair $tt' \in B^6$ in the gene sequences $\mathcal{G}_{GS}$ of a eukaryotic genome $\mathcal{G}$ is

$$P\left(tt', \mathcal{G}_{GS}\right) = \frac{N(tt', \mathcal{G}_{GS})}{\sum_{tt' \in B^6} N(tt', \mathcal{G}_{GS})}. \tag{18}$$

Eq. (18) in the gene sequences $\mathcal{G}_{GS}$ of a eukaryotic genome $\mathcal{G}$ is similar to Eq. (14) in all the gene sequences of eukaryotes $\mathbb{E}$. Similarly as previously, the observed probability $P(t, \mathcal{G}_{GS})$ of a trinucleotide $t \in B^3$ in the gene sequences $\mathcal{G}_{GS}$ of $\mathcal{G}$ is

$$P(t, \mathcal{G}_{GS}) = \frac{N(t, \mathcal{G}_{GS})}{N(\mathcal{G}_{GS})} \tag{19}$$

where $N(t, \mathcal{G}_{GS})$ defined for Eq. (15) is the occurrence number of t in the gene sequences $\mathcal{G}_{GS}$ of $\mathcal{G}$ and $N(\mathcal{G}_{GS}) = \sum_{t \in B^3} N(t, \mathcal{G}_{GS})$ is the size in trinucleotides (given in Appendix A) of gene sequences $\mathcal{G}_{GS}$ of $\mathcal{G}$. By taking the hypothesis of independent events then the theoretical probability $\hat{P}(tt', \mathcal{G}_{GS})$ of a trinucleotide pair $tt' \in B^6$ in the gene sequences $\mathcal{G}_{GS}$ of $\mathcal{G}$ is

$$\hat{P}\left(tt', \mathcal{G}_{GS}\right) = P(t, \mathcal{G}_{GS}) \times P\left(t', \mathcal{G}_{GS}\right). \tag{20}$$

Then, the observed/theoretical ratio $r(tt', \mathcal{G}_{GS})$ of a trinucleotide pair $tt' \in B^6$ in the gene sequences $\mathcal{G}_{GS}$ of $\mathcal{G}$ is equal to

$$r\left(tt', \mathcal{G}_{GS}\right) = \frac{P(tt', \mathcal{G}_{GS})}{\hat{P}(tt', \mathcal{G}_{GS})}. \tag{21}$$

Finally, $\bar{r}(tt')$ is the mean of the observed/theoretical ratios $r(tt', \mathcal{G}_{GS})$ of a trinucleotide pair $tt' \in B^6$ in all the gene sequences of eukaryotes $\mathbb{E}$

$$\bar{r}\left(tt'\right) = \frac{1}{|\mathbb{E}|} \sum_{\mathcal{G}_{GS} \in \mathbb{E}} r\left(tt', \mathcal{G}_{GS}\right) \tag{22}$$

where $|\mathbb{E}|$ is the number of genomes $\mathcal{G}$ in $\mathbb{E}$ ($|\mathbb{E}| = 126$ here) and

$$\tilde{r}\left(tt'\right) \text{ is the median of the observed/theoretical ratios } r\left(tt', \mathcal{G}_{GS}\right) \text{ of a trinucleotide pair } tt' \in B^6 \text{ in all the gene sequences of eukaryotes } \mathbb{E}. \tag{23}$$

Remark 5. The three observed/theoretical ratios $r(tt')$ (Eq. (17)), $\bar{r}(tt')$ (Eq. (22)) and $\tilde{r}(tt')$ (Eq. (23)) of a trinucleotide pair $tt' \in B^6$ have the same statistical property. When $r(tt') > 1$, $\bar{r}(tt') > 1$ or $\tilde{r}(tt') > 1$, the trinucleotide pair tt' is overrepresented in the eukaryotic gene sequences, and conversely when $r(tt') < 1$, $\bar{r}(tt') < 1$ or $\tilde{r}(tt') < 1$.

The three observed/theoretical ratios $r(tt')$, $\bar{r}(tt')$ and $\tilde{r}(tt')$ of trinucleotide pairs will lead to the same statistical results with the circular code X in the eukaryotic gene sequences. In Section 3.7, the results with the circular code X in the eukaryotic gene sequences are presented with the median $\tilde{r}(tt')$ (Eq. (23)) where the trinucleotide pairs $tt' \in X^2$.

2.8. 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. Then, the Genbank file is retrieved for each chromosome in order to extract the coordinates of its gene sequences (GenBank keyword CDS).

Thus, 126 complete genomes of eukaryotes are extracted from GenBank (http://www.ncbi.nlm.nih.gov/genbank/, June 2016). The list of 126 eukaryotic genome $\mathcal{G}$, the size $N(\mathcal{G}_{GS})$ in trinucleotides for the gene sequences $\mathcal{G}_{GS}$ of $\mathcal{G}$ and the size $N(\mathcal{G})$ in bases of $\mathcal{G}$ are given in Appendix A. This genome information represents a total of 91,350,244,263 bases. Genomes of prokaryotes which contain very low proportion of non-coding DNA compared to coding DNA (about 10% for an order of magnitude) are not studied.

3. Results

3.1. Occurrence of repeated dinucleotides in the genomes of eukaryotes

The repeated dinucleotides (Section 2.3.1) are generated from the unitary circular codes (UCC) of dinucleotides (Section 2.2.1). Fig. 13a,b give the occurrence number $N(d^+)$ (Eq. (7)) and the base

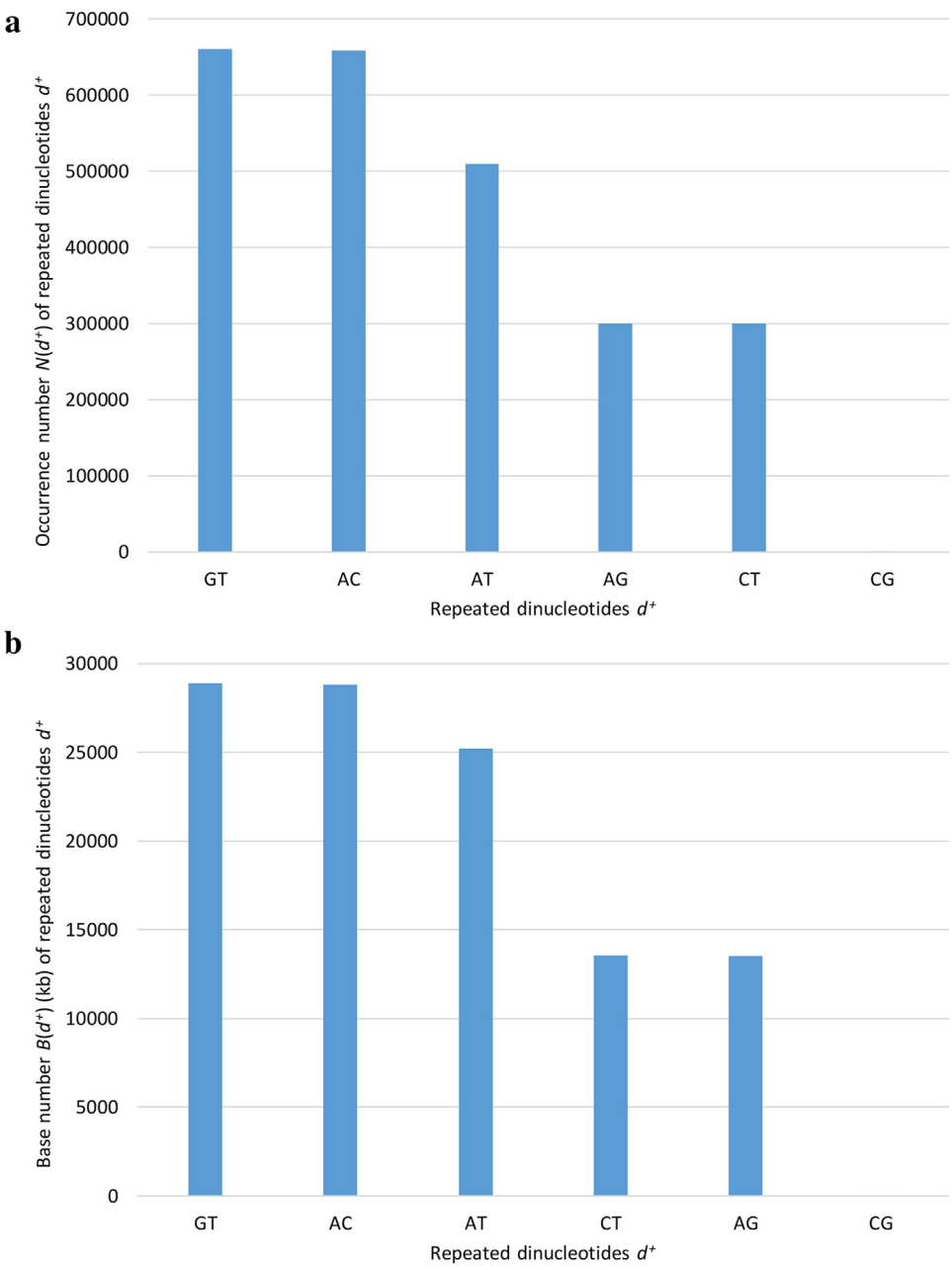

Fig. 13. a. Occurrence number $N(d^+)$ (Eq. (7)) (decreasing order) of the six repeated dinucleotides d^n (Eq. (4)) of length $l = 2n \geq 30$ nucleotides ($n \geq 15$) in the eukaryotic genomes (see Appendix A). The repeated dinucleotides are generated from the unitary circular codes (UCC) of dinucleotides. b. Base number $B(d^+)$ (Eq. (8)) (kb for kilobases; decreasing order) of the six repeated dinucleotides d^n (Eq. (4)) of length $l = 2n \geq 30$ nucleotides ($n \geq 15$) in the eukaryotic genomes (see Appendix A). The repeated dinucleotides are generated from the unitary circular codes (UCC) of dinucleotides.

number $B(d^+)$ (Eq. (8)) of the six repeated dinucleotides d^n (Eq. (4)) of length $l = 2n \geq 30$ nucleotides ($n \geq 15$) in the genomes of eukaryotes. The results in the two Fig. 13a,b are altogether consistent. The repeats $(AC)^+$ and $(GT)^+ = (\mathcal{C}(AC))^+$ have the highest occurrences in the eukaryotic genomes. Then, the repeat $(AT)^+$ (note that $\mathcal{C}(AT) = AT$) has a lower occurrence. The repeats $(AG)^+$ and $(CT)^+ = (\mathcal{C}(AG))^+$ have occurrences lower than $(AT)^+$. The repeat $(CG)^+$ (note that $\mathcal{C}(CG) = CG$) is almost absent.

A repeated dinucleotide d^+ and its complementary repeated dinucleotide $(\mathcal{C}(d))^+$ have the same occurrences in the eukaryotic genomes: $N\left((AC)^+\right) \approx N\left((GT)^+\right) \approx 659400$, $B\left((AC)^+\right) \approx B\left((GT)^+\right) \approx 28800$ kb, $N\left((AG)^+\right) \approx N\left((CT)^+\right) \approx 299900$ and $B\left((AG)^+\right) \approx B\left((CT)^+\right) \approx 13500$ kb (Fig. 13a,b). This property is related to the complementary property of the DNA double helix.

3.2. Occurrence of repeated trinucleotides in the genomes of eukaryotes

The repeated trinucleotides (Section 2.3.2) are generated from the unitary circular codes (UCC) of trinucleotides (Section 2.2.2). Fig. 14a,b give the occurrence number $N(t^+)$ (Eq. (7)) and the base number $B(t^+)$ (Eq. (8)) of the 20 repeated trinucleotides t^n (Eq. (5)) of length $l = 3n \geq 30$ nucleotides ($n \geq 10$) in the genomes of eukaryotes. Again, the results in the two Fig. 14a,b are altogether consistent. The repeats $(AAT)^+$ and $(ATT)^+ = (\mathcal{C}(AAT))^+$ have the highest occurrences in the eukaryotic genomes. Then, the following repeats are observed by decreasing order of occurrence: $(AAG)^+$ and $(CTT)^+ = (\mathcal{C}(AAG))^+$, $(AAC)^+$ and $(GTT)^+ = (\mathcal{C}(AAC))^+$, $(ATC)^+$ and $(ATG)^+ = (\mathcal{C}(\mathcal{P}^2(ATC)))^+$ (i.e. $(ATG)^+$ and $(GAT)^+ = (\mathcal{C}(ATC))^+$

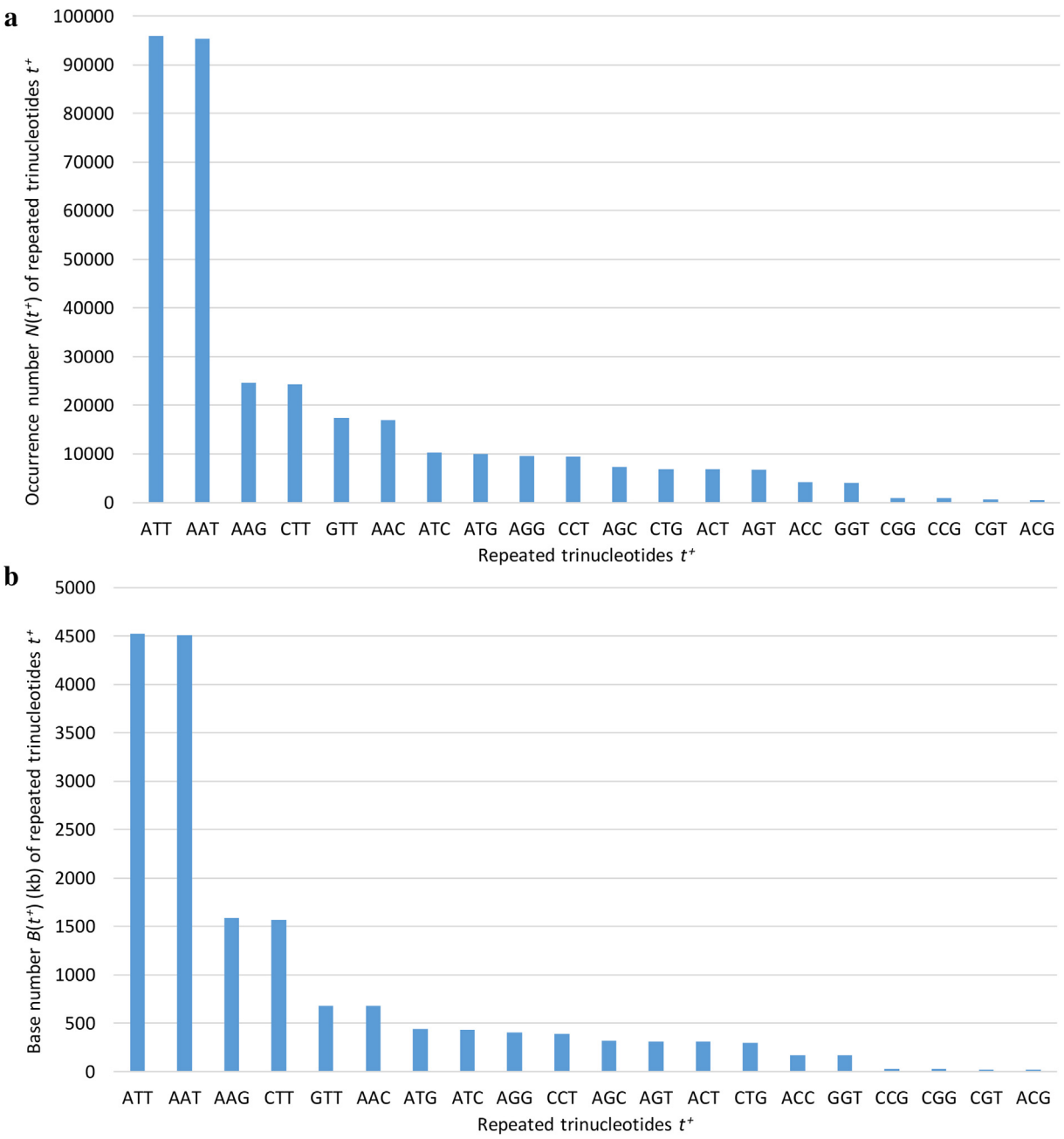

Fig. 14. a. Occurrence number $N(t^+)$ (Eq. (7)) (decreasing order) of the 20 repeated trinucleotides t^n (Eq. (5)) of length $l = 3n \geq 30$ nucleotides ($n \geq 10$) in the eukaryotic genomes (see Appendix A). The repeated trinucleotides are generated from the unitary circular codes (UCC) of trinucleotides. b. Base number $B(t^+)$ (Eq. (8)) (kb for kilobases; decreasing order) of the 20 repeated trinucleotides t^n (Eq. (5)) of length $l = 3n \geq 30$ nucleotides ($n \geq 10$) in the eukaryotic genomes (see Appendix A). The repeated trinucleotides are generated from the unitary circular codes (UCC) of trinucleotides.

belong to the same equivalence class by the circular permutation map $\mathcal{P}$), $(AGG)^+$ and $(CCT)^+ = (\mathcal{C}(AGG))^+$, $(AGC)^+$ and $(CTG)^+ = (\mathcal{C}(\mathcal{P}^2(AGC)))^+$ (i.e. $(CTG)^+$ and $(GCT)^+ = (\mathcal{C}(AGC))^+$ belong to the same equivalence class), $(ACT)^+$ and $(AGT)^+ = (\mathcal{C}(ACT))^+$, and $(ACC)^+$ and $(GGT)^+ = (\mathcal{C}(ACC))^+$. The repeats $(ACG)^+$ and $(CGT)^+ = (\mathcal{C}(ACG))^+$, and $(CCG)^+$ and $(CGG)^+ = (\mathcal{C}(CCG))^+$ are almost absent.

A repeated trinucleotide t^+ and its complementary repeated trinucleotide $(\mathcal{C}(t))^+$ have the same occurrences in the eukaryotic genomes: $N\left((AAT)^+\right) \approx N\left((ATT)^+\right) \approx 95700$, $B\left((AAT)^+\right) \approx B\left((ATT)^+\right) \approx 4500$ kb, $N\left((AAG)^+\right) \approx N\left((CTT)^+\right) \approx 24500$ and $B\left((AAG)^+\right) \approx B\left((CTT)^+\right) \approx 1600$ kb, etc. (Fig. 14a,b). This property is again related to the complementary property of the DNA double helix.

This result is also confirmed by the correlation matrix of the base number $B(t^+)$ of the 20 repeated trinucleotides t^+ (Table 1).

Remark 6. The correlation matrix of n random variables $X_1, \ldots, X_n$ is the $n \times n$ matrix whose i, j entry is the Pearson product-moment correlation coefficient commonly called simply "the correlation coefficient" $corr\left(X_i, X_j\right)$. The correlation matrix is symmetric because $corr\left(X_i, X_j\right) = corr\left(X_j, X_i\right)$.

The highest correlation is always observed between the repeated trinucleotides t^+ and $(\mathcal{C}(t))^+$ in the eukaryotic genomes. There is no significant correlation between a repeated trinucleotide t^+ and the size $N(\mathcal{G})$ of genomes (given in Appendix A) as well as the A, C, G, T and GC content (data not shown) of genomes.

Table 1
Correlation matrix of the base number $B(t^+)$ (Eq. (8) and Fig. 14b) of the 20 repeated trinucleotides t^n (Eq. (5)) of length $l = 3n \geq 30$ nucleotides ($n \geq 10$) in the eukaryotic genomes (see Appendix A). The highest correlation (in bold) is always between a repeated trinucleotide t^+ and its complementary repeated trinucleotide $(\mathcal{C}(t))^+$ (note that $(ATC)^+$ and $(ATG)^+ = (\mathcal{C}(\mathcal{P}^2(ATC)))^+$, and $(AGC)^+$ and $(CTG)^+ = (\mathcal{C}(\mathcal{P}^2(AGC)))^+$, details in Section 3.2). There is no significant correlation between a repeated trinucleotide t^+ and the size $N(\mathcal{G})$ of genomes (given in Appendix A) as well as the *A*, *C*, *G*, *T* and *GC* content (data not shown) of genomes.

	Size	*A* content	*C* content	*G* content	*T* content	*GC* content	*AAC*	*AAG*	*AAT*	*ACC*	*ACG*	*ACT*	*AGC*	*AGG*	*AGT*	*ATC*	*ATG*	*ATT*	*CCG*	*CCT*	*CGG*	*CGT*	*CTG*	*CTT*	*GGT*	*GTT*
Size	1.00	0.11	−0.11	−0.11	0.11	−0.11	0.45	0.43	0.17	0.30	0.08	0.38	0.41	0.35	0.38	0.37	0.38	0.17	0.56	0.35	0.50	0.07	0.40	0.43	0.32	0.46
A content	0.11	1.00	−1.00	−1.00	1.00	−1.00	0.14	0.07	0.17	0.00	−0.19	0.05	−0.01	0.02	0.05	0.08	0.10	0.16	−0.03	0.02	−0.03	−0.14	0.00	0.06	0.01	0.12
C content	−0.11	−1.00	1.00	1.00	−1.00	1.00	−0.14	−0.07	−0.17	0.00	0.19	−0.05	0.01	−0.02	−0.05	−0.08	−0.10	−0.16	0.02	−0.02	0.03	0.14	0.00	−0.06	−0.01	−0.12
G content	−0.11	−1.00	1.00	1.00	−1.00	1.00	−0.14	−0.07	−0.17	0.00	0.19	−0.05	0.01	−0.02	−0.05	−0.08	−0.10	−0.16	0.03	−0.02	0.03	0.14	0.00	−0.06	0.00	−0.12
T content	0.11	1.00	−1.00	−1.00	1.00	−1.00	0.14	0.07	0.17	0.00	−0.19	0.05	−0.01	0.02	0.05	0.08	0.10	0.16	−0.02	0.02	−0.03	−0.14	0.00	0.06	0.00	0.12
GC content	−0.11	−1.00	1.00	1.00	−1.00	1.00	−0.14	−0.07	−0.17	0.00	0.19	−0.05	0.01	−0.02	−0.05	−0.08	−0.10	−0.16	0.03	−0.02	0.03	0.14	0.00	−0.06	0.00	−0.12
AAC	0.45	0.14	−0.14	−0.14	0.14	−0.14	1.00	0.57	0.75	0.55	0.19	0.38	0.29	0.58	0.38	0.85	0.86	0.75	0.30	0.57	0.23	0.21	0.29	0.56	0.56	**0.98**
AAG	0.43	0.07	−0.07	−0.07	0.07	−0.07	0.57	1.00	0.18	0.75	0.29	0.56	0.36	0.92	0.58	0.57	0.59	0.19	0.50	0.92	0.43	0.25	0.37	**1.00**	0.77	0.57
AAT	0.17	0.17	−0.17	−0.17	0.17	−0.17	0.75	0.18	1.00	0.11	0.00	0.27	0.04	0.20	0.27	0.77	0.76	**1.00**	0.04	0.20	0.02	0.01	0.04	0.18	0.10	0.77
ACC	0.30	0.00	0.00	0.00	0.00	0.00	0.55	0.75	0.11	1.00	0.44	0.43	0.49	0.72	0.43	0.53	0.56	0.11	0.38	0.70	0.26	0.38	0.50	0.74	**1.00**	0.55
ACG	0.08	−0.19	0.19	0.19	−0.19	0.19	0.19	0.29	0.00	0.44	1.00	0.28	0.18	0.28	0.27	0.21	0.21	0.00	0.19	0.28	0.14	**0.93**	0.18	0.28	0.44	0.18
ACT	0.38	0.05	−0.05	−0.05	0.05	−0.05	0.38	0.56	0.27	0.43	0.28	1.00	0.25	0.42	**0.99**	0.51	0.52	0.27	0.25	0.42	0.19	0.21	0.25	0.57	0.44	0.40
AGC	0.41	−0.01	0.01	0.01	−0.01	0.01	0.29	0.36	0.04	0.49	0.18	0.25	1.00	0.37	0.25	0.30	0.32	0.04	0.21	0.36	0.15	0.15	**1.00**	0.37	0.49	0.29
AGG	0.35	0.02	−0.02	−0.02	0.02	−0.02	0.58	0.92	0.20	0.72	0.28	0.42	0.37	1.00	0.43	0.50	0.51	0.20	0.39	**1.00**	0.32	0.22	0.38	0.93	0.73	0.58
AGT	0.38	0.05	−0.05	−0.05	0.05	−0.05	0.38	0.58	0.27	0.43	0.27	**0.99**	0.25	0.43	1.00	0.52	0.53	0.27	0.25	0.43	0.18	0.21	0.25	0.58	0.44	0.39
ATC	0.37	0.08	−0.08	−0.08	0.08	−0.08	0.85	0.57	0.77	0.53	0.21	0.51	0.30	0.50	0.52	1.00	**0.99**	0.77	0.28	0.49	0.21	0.19	0.31	0.56	0.54	0.87
ATG	0.38	0.10	−0.10	−0.10	0.10	−0.10	0.86	0.59	0.76	0.56	0.21	0.52	0.32	0.51	0.53	**0.99**	1.00	0.76	0.28	0.51	0.21	0.22	0.32	0.58	0.56	0.88
ATT	0.17	0.16	−0.16	−0.16	0.16	−0.16	0.75	0.19	**1.00**	0.11	0.00	0.27	0.04	0.20	0.27	0.77	0.76	1.00	0.04	0.20	0.02	0.01	0.04	0.18	0.11	0.78
CCG	0.56	−0.03	0.02	0.03	−0.02	0.03	0.30	0.50	0.04	0.38	0.19	0.25	0.21	0.39	0.25	0.28	0.28	0.04	1.00	0.39	**0.95**	0.17	0.22	0.50	0.39	0.31
CCT	0.35	0.02	−0.02	−0.02	0.02	−0.02	0.57	0.92	0.20	0.70	0.28	0.42	0.36	**1.00**	0.43	0.49	0.51	0.20	0.39	1.00	0.32	0.22	0.37	0.93	0.72	0.58
CGG	0.50	−0.03	0.03	0.03	−0.03	0.03	0.23	0.43	0.02	0.26	0.14	0.19	0.15	0.32	0.18	0.21	0.21	0.02	**0.95**	0.32	1.00	0.13	0.16	0.42	0.27	0.24
CGT	0.07	−0.14	0.14	0.14	−0.14	0.14	0.21	0.25	0.01	0.38	**0.93**	0.21	0.15	0.22	0.21	0.19	0.22	0.01	0.17	0.22	0.13	1.00	0.16	0.24	0.38	0.16
CTG	0.40	0.00	0.00	0.00	0.00	0.00	0.29	0.37	0.04	0.50	0.18	0.25	**1.00**	0.38	0.25	0.31	0.32	0.04	0.22	0.37	0.16	0.16	1.00	0.37	0.50	0.30
CTT	0.43	0.06	−0.06	−0.06	0.06	−0.06	0.56	**1.00**	0.18	0.74	0.28	0.57	0.37	0.93	0.58	0.56	0.58	0.18	0.50	0.93	0.42	0.24	0.37	1.00	0.76	0.56
GGT	0.32	0.01	−0.01	0.00	0.00	0.00	0.56	0.77	0.10	**1.00**	0.44	0.44	0.49	0.73	0.44	0.54	0.56	0.11	0.39	0.72	0.27	0.38	0.50	0.76	1.00	0.55
GTT	0.46	0.12	−0.12	−0.12	0.12	−0.12	**0.98**	0.57	0.77	0.55	0.18	0.40	0.29	0.58	0.39	0.87	0.88	0.78	0.31	0.58	0.24	0.16	0.30	0.56	0.55	1.00

A second property is identified with the repeated trinucleotides t^+ and $(\mathcal{C}(t))^+$. Indeed, the repeated trinucleotides t^+ and $(\mathcal{C}(t))^+$ have increasing occurrences in the eukaryotic genomes contrary to their number of hydrogen bonds (two hydrogen bonds between A and $T = \mathcal{C}(A)$ and three hydrogen bonds between C and $G = \mathcal{C}(C)$), from the highest occurrences for the two repeats $(AAT)^+$ and $(ATT)^+$ with a total of six hydrogen bonds to the lowest occurrences for the two repeats $(CCG)^+$ and $(CGG)^+$ with a total of nine hydrogen bonds.

3.3. Occurrence of repeated tetranucleotides in the genomes of eukaryotes

The repeated tetranucleotides (Section 2.3.3) are generated from the unitary circular codes (*UCC*) of tetranucleotides (Section 2.2.3). Fig. 15a,b give the occurrence number $N(t^+)$ (Eq. (7)) and the base number $B(t^+)$ (Eq. (8)) of the repeated tetranucleotides t^n (Eq. (6)) of length $l = 4n \geq 28$ nucleotides ($n \geq 7$) occurring in the genomes of eukaryotes significantly. The results in the two Fig. 15a,b are consistent and identify two classes of repeated tetranucleotides with high occurrences. The 1st class with the highest occurrences in the eukaryotic genomes contains eight repeated tetranucleotides, by decreasing order: $(AAAT)^+$ and $(ATTT)^+ = (\mathcal{C}(AAAT))^+$, $(AAAG)^+$ and $(CTTT)^+ = (\mathcal{C}(AAAG))^+$, $(AGAT)^+$ and $(ATCT)^+ = (\mathcal{C}(AGAT))^+$, and $(AAGG)^+$ and $(CCTT)^+ = (\mathcal{C}(AAGG))^+$ (Fig. 15a). Note that this repeated tetranucleotide order is different in Fig. 15b. The 2nd class with high occurrences in the eukaryotic genomes contains 12 repeated tetranucleotides, by decreasing order: $(ATCC)^+$ and $(ATGG)^+ = (\mathcal{C}(\mathcal{P}^2(ATCC)))^+$, $(AAAC)^+$ and $(GTTT)^+ = (\mathcal{C}(AAAC))^+$, $(ACAG)^+$ and $(CTGT)^+ = (\mathcal{C}(ACAG))^+$, $(ACAT)^+$ and $(ATGT)^+ = (\mathcal{C}(ACAT))^+$, $(AATG)^+$ and $(ATTC)^+ = (\mathcal{C}(\mathcal{P}^3(AATG)))^+$, and $(AGGG)^+$ and $(CCCT)^+ = (\mathcal{C}(AGGG))^+$ (Fig. 15a). Note that this repeated tetranucleotide order is also different in Fig. 15b. The repeats $(CCGG)^+$ (note that $\mathcal{C}(CCGG) = CCGG$), and $(CCCG)^+$ and $(CGGG)^+ = (\mathcal{C}(CCCG))^+$ are almost absent (results not shown). A repeated tetranucleotide t^+ and its complementary repeated tetranucleotide $(\mathcal{C}(t))^+$ also have the same occurrences in the eukaryotic genomes (Fig. 15a,b). Overall, the repeated tetranucleotides t^+ and $(\mathcal{C}(t))^+$ have increasing occurrences contrary to their number of hydrogen bonds, from the highest occurrences for the two repeats $(AAAT)^+$ and $(ATTT)^+$ (Fig. 15a, less significant in Fig. 15b) with a total of eight hydrogen bonds to the lowest occurrences for the three repeats $(CCCG)^+$, $(CCGG)^+$ and $(CGGG)^+$ (results not shown) with a total of 12 hydrogen bonds.

Remark 7. Among the 12 repeated tetranucleotides (Remark 4) associated to the 48 *UCC* codes of the mitochondrial genetic code (Gonzalez et al., 2012), only the two repeated tetranucleotides $(AAGG)^+$ and $(CCTT)^+$ belong to the 1st class (see above) and thus, they have high occurrences in the eukaryotic genomes. The 10 other repeated tetranucleotides $(ACGT)^+$, $(ACTG)^+$, $(AGCT)^+$, $(AGTC)^+$, $(ATCG)^+$, $(ATGC)^+$, $(AACC)^+$, $(AATT)^+$, $(CCGG)^+$ and $(GGTT)^+$ which do not belong to any class (see above), have no significant occurrence in the eukaryotic genomes.

3.4. Largest nucleotide lengths of unitary circular code motifs in the genomes of eukaryotes

Table 2 shows the largest nucleotide lengths $l = 2n$ for the six repeated dinucleotides d^n, $l = 3n$ for the 20 repeated trinucleotides t^n and $l = 4n$ for the 10 largest repeated tetranucleotides t^n in the eukaryotic genomes. The largest repeated dinucleotide $(AT)^n$ of length $l = 11254$ nucleotides is observed in the chromosome 1 of *Medicago truncatula*. The largest repeated trinucleotide $(ATT)^n$ of length $l = 19275$ nucleotides is found in the chromosome 9 of *Citrus sinensis*. The largest repeated tetranucleotide $(ATCC)^n$ of length $l = 6952$ nucleotides is present in the chromosome 11 of *Solanum pennellii*.

Table 2
Largest nucleotide lengths $l = 2n$ for the six repeated dinucleotides d^n, $l = 3n$ for the 20 repeated trinucleotides t^n and $l = 4n$ for the 10 largest repeated tetranucleotides t^n in the eukaryotic genomes (see Appendix A). The 1st column indicates the unitary circular code (*UCC*) motif, the 2nd and 3rd columns mention the genome $\mathcal{G}$ and its chromosome number $\mathcal{G}_{Chr}$, respectively, and the 4th column gives the nucleotide length l of the *UCC* motif.

UCC motif	Genome $\mathcal{G}$	$\mathcal{G}_{Chr}$	Length l of *UCC* motif (in bases)
$(AC)^n$	*Citrus sinensis*	7	4500
$(AG)^n$	*Citrus sinensis*	3	7648
$(AT)^n$	*Medicago truncatula*	1	11,254
$(CG)^n$	*Cucumis sativus*	7	432
$(CT)^n$	*Citrus sinensis*	5	7232
$(GT)^n$	*Beta vulgaris*	5	2680
$(AAC)^n$	*Solanum pennellii*	9	8655
$(AAG)^n$	*Solanum pennellii*	12	10,536
$(AAT)^n$	*Citrus sinensis*	4	12,951
$(ACC)^n$	*Oryza brachyantha*	11	363
$(ACG)^n$	*Bombus terrestris*	B04	144
$(ACT)^n$	*Solanum pennellii*	12	1728
$(AGC)^n$	*Ficedula albicollis*	21	3555
$(AGG)^n$	*Ficedula albicollis*	6	822
$(AGT)^n$	*Zea mays*	10	1926
$(ATC)^n$	*Camelina sativa*	5	2145
$(ATG)^n$	*Citrus sinensis*	9	2076
$(ATT)^n$	*Citrus sinensis*	9	19,275
$(CCG)^n$	*Oryza brachyantha*	9	555
$(CCT)^n$	*Ficedula albicollis*	14	1065
$(CGG)^n$	*Oryza brachyantha*	7	210
$(CGT)^n$	*Solanum pennellii*	8	1815
$(CTG)^n$	*Ficedula albicollis*	12	723
$(CTT)^n$	*Cicer arietinum*	Ca6	4263
$(GGT)^n$	*Homo sapiens*	2	630
$(GTT)^n$	*Cicer arietinum*	Ca8	4239
$(ATCC)^n$	*Solanum pennellii*	11	6952
$(ATCT)^n$	*Solanum pennellii*	6	5492
$(ATGG)^n$	*Solanum pennellii*	9	5076
$(AGAT)^n$	*Solanum pennellii*	10	4904
$(AAAG)^n$	*Ficedula albicollis*	1	4780
$(ATTT)^n$	*Cynoglossus semilaevis*	5	4268
$(CTTT)^n$	*Ficedula albicollis*	1	4200
$(AAGG)^n$	*Ficedula albicollis*	15	4048
$(CCTT)^n$	*Ficedula albicollis*	1	3668
$(AGTG)^n$	*Cicer arietinum*	Ca2	3036

3.5. Scarcity of repeated trinucleotides in the large genomes of eukaryotes

Table 3 shows the ratios $r(D^+, \mathcal{G})$, $r(T^+, \mathcal{G})$ and $r(\mathcal{T}^+, \mathcal{G})$ (Eq. (10)) giving the proportion of the total base numbers $B(D^+, \mathcal{G})$, $B(T^+, \mathcal{G})$ and $B(\mathcal{T}^+, \mathcal{G})$ (Eq. (9)) of the six repeated dinucleotides d^+ in D^+ (Eq. (4)), the 20 repeated trinucleotides t^+ in T^+ (Eq. (5)) and the 60 repeated tetranucleotides t^+ in $\mathcal{T}^+$ (Eq. (6)) in the 59 large eukaryotic genomes $\mathcal{G}$ (sizes $N(\mathcal{G}) > 300000$ kb). Interestingly, the means $\bar{r}(D^+)$, $\bar{r}(T^+)$ and $\bar{r}(\mathcal{T}^+)$ (Eq. (11)) and the medians $\tilde{r}(D^+)$, $\tilde{r}(T^+)$ and $\tilde{r}(\mathcal{T}^+)$ (Eq. (12)) of the ratios $r(D^+, \mathcal{G})$, $r(T^+, \mathcal{G})$ and $r(\mathcal{T}^+, \mathcal{G})$, respectively, in the eukaryotic genomes leads both to the same following result

$$\begin{cases} \bar{r}(D^+) > \bar{r}(\mathcal{T}^+) > \bar{r}(T^+) \\ \tilde{r}(D^+) > \tilde{r}(\mathcal{T}^+) > \tilde{r}(T^+) \end{cases}. \tag{24}$$

These inequalities (24) are evaluated by two statistical tests: a paired sample Student's t-test (parametric statistical hypothesis test assuming a normal distribution of the population) and a Wilcoxon signed-rank W-test (non-parametric statistical hypothesis test). The comparisons of the means $\bar{r}(D^+)$ and $\bar{r}(\mathcal{T}^+)$, and the means $\bar{r}(\mathcal{T}^+)$ and $\bar{r}(T^+)$ with the t-test have significant p-values equal to 3×10^{-5} and 7×10^{-3}, respectively. The comparisons of

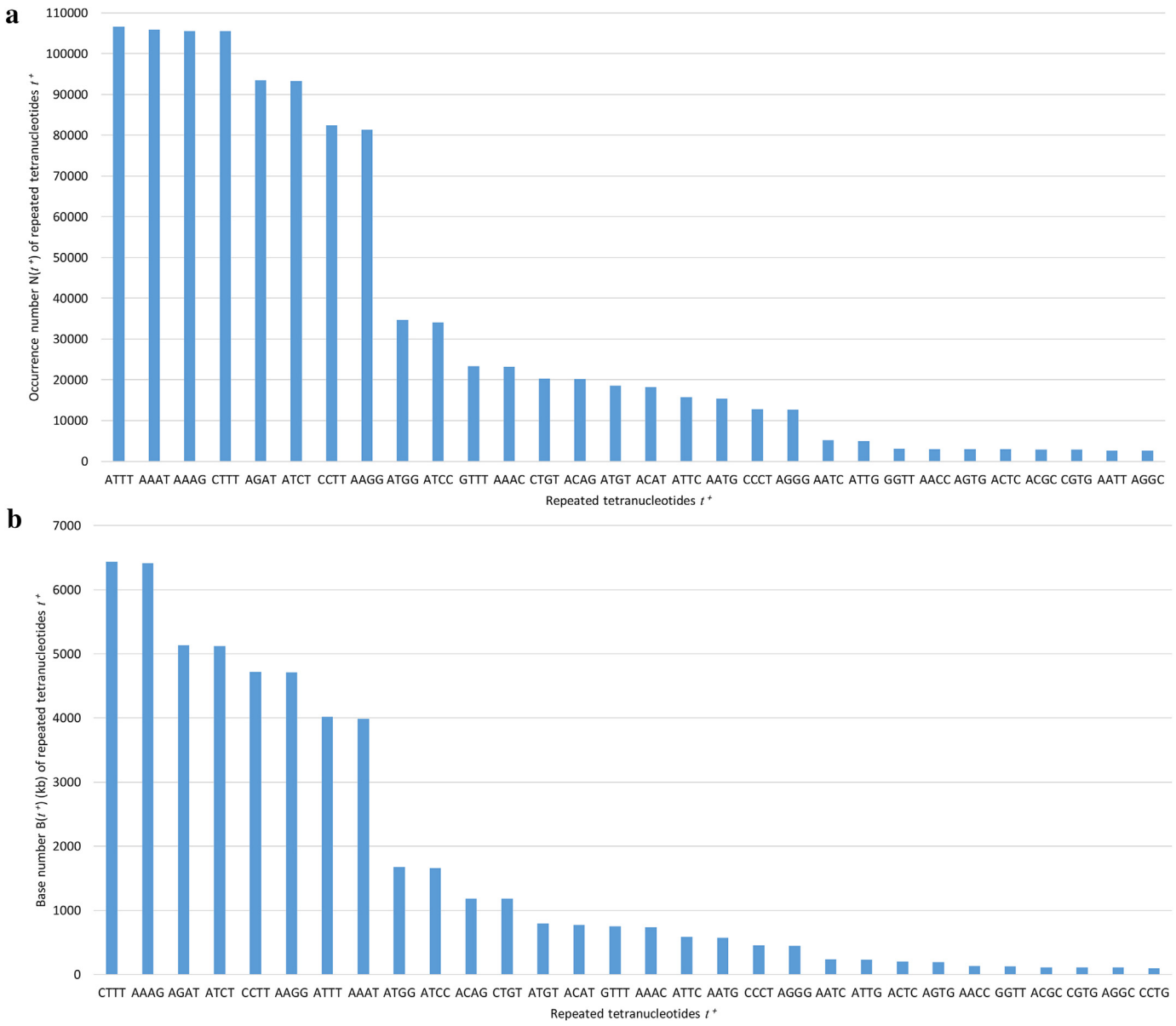

Fig. 15. a. Occurrence number $N(t^+)$ (Eq. (7)) (decreasing order) of the repeated tetranucleotides t^n (Eq. (6)) of lengths $l = 4n \geq 28$ nucleotides ($n \geq 7$) in the eukaryotic genomes (see Appendix A). Only the repeated tetranucleotides t^n with an occurrence greater than 2500 (the first 30 repeated tetranucleotides) are represented. The repeated tetranucleotides are generated from the unitary circular codes (UCC) of tetranucleotides. b. Base number $B(t^+)$ (Eq. (8)) (kb for kilobases; decreasing order) of the repeated tetranucleotides t^n (Eq. (6)) of lengths $l = 4n \geq 28$ nucleotides ($n \geq 7$) in the eukaryotic genomes (see Appendix A). Only the repeated tetranucleotides t^n greater than 100 kb (the first 30 repeated tetranucleotides) are represented. The repeated tetranucleotides are generated from the unitary circular codes (UCC) of tetranucleotides.

the distribution of D^+ and $\mathcal{T}^+$, and the distribution of $\mathcal{T}^+$ and T^+ with the Wilcoxon test also have significant p-values equal to 10^{-6} and 9×10^{-3}, respectively. Thus, the total base proportion of D^+ is greater than the total base proportion of $\mathcal{T}^+$ which is greater than the total base proportion of T^+. In other words, there is a scarcity of repeated trinucleotides in the large eukaryotic genomes compared to the repeated dinucleotides and the repeated tetranucleotides. For the eukaryotic genomes $\mathcal{G}$ of small sizes $N(\mathcal{G}) < 300000$ kb, the analysis has the same statistical trend. However, it is not conclusive and should be investigated in the future with the increase of genome data.

The scarcity of repeated trinucleotides in the large genomes of eukaryotes may be explained by the two following results: (i) the observation of repeated trinucleotides in the X motifs in the genomes of eukaryotes and (ii) the preferential use of identical trinucleotide pairs of the maximal C^3 self-complementary trinucleotide circular code X in the gene sequences of eukaryotes.

3.6. Repeated trinucleotides in the **X** motifs in the genomes of eukaryotes

Mutation, in particular substitution, in unitary circular code motifs (D^+, T^+ and $\mathcal{T}^+$ motifs and more general repeated motifs, e.g. repeated pentanucleotides, repeated hexanucleotides, etc.) associated with unitary circular codes of cardinality 1 (composition with one unique word) may generate complex circular code motifs associated with circular codes of cardinality greater than 1 up to circular codes of maximal cardinality. Low composition X motifs are abundant in the eukaryotic genomes (El Soufi and Michel, 2016). Table 4 shows 10 examples of X motifs with repeated trinucleotides. Further mutation would increase the composition of X motifs while decreasing the amount of repeats it holds.

Table 3
Scarcity of repeated trinucleotides (T^+ motifs) in the large eukaryotic genomes. The 1st column mentions the 59 eukaryotic genomes $\mathcal{G}$ of large sizes $N(\mathcal{G}) > 300000$ kb (see Appendix A), the 2nd, 3rd and 4th provide the ratios $r(D^+, \mathcal{G})$ (‰), $r(T^+, \mathcal{G})$ (‰) and $r(\mathcal{T}^+, \mathcal{G})$ (‰), respectively, (Eq. (10)) giving the proportion of the total base numbers $B(D^+, \mathcal{G})$, $B(T^+, \mathcal{G})$ and $B(\mathcal{T}^+, \mathcal{G})$, respectively, (Eq. (9)) of the six repeated dinucleotides d^+ in D^+ (Eq. (4)), the 20 repeated trinucleotides t^+ in T^+ (Eq. (5)) and the 60 repeated tetranucleotides t^+ in $\mathcal{T}^+$ (Eq. (6)), respectively, in the large eukaryotic genomes. The means $\bar{r}(D^+)$, $\bar{r}(T^+)$ and $\bar{r}(\mathcal{T}^+)$ (Eq. (11)) and the medians $\tilde{r}(D^+)$, $\tilde{r}(T^+)$ and $\tilde{r}(\mathcal{T}^+)$ (Eq. (12)) of the ratios $r(D^+, \mathcal{G})$, $r(T^+, \mathcal{G})$ and $r(\mathcal{T}^+, \mathcal{G})$ in the large eukaryotic genomes lead to Eq. (24).

Genome $\mathcal{G}$	$r(D^+, \mathcal{G})$ (‰)	$r(T^+, \mathcal{G})$ (‰)	$r(\mathcal{T}^+, \mathcal{G})$ (‰)
Anolis carolinensis	0.814	2.602	0.560
Beta vulgaris	0.827	0.668	0.050
Bos taurus	0.450	0.022	0.008
Brassica napus	0.332	0.072	0.013
Brassica oleracea	0.459	0.093	0.011
Callithrix jacchus	0.791	0.033	0.391
Camelina sativa	0.707	0.202	0.009
Canis lupus familiaris	1.119	0.207	1.734
Capra hircus	0.458	0.041	0.027
Chlorocebus sabaeus	0.524	0.106	0.870
Chrysemys picta bellii	0.565	0.007	0.091
Cicer arietinum	0.948	1.285	0.170
Cynoglossus semilaevis	1.297	0.613	0.520
Danio rerio	8.289	0.987	3.884
Elaeis guineensis	0.880	0.096	0.054
Equus caballus	0.180	0.012	0.057
Esox lucius	1.028	0.022	0.056
Felis catus	2.624	0.152	1.027
Ficedula albicollis	0.201	0.389	1.783
Gallus gallus	0.070	0.029	0.323
Glycine max	2.034	0.260	0.013
Gorilla gorilla gorilla	0.448	0.055	0.315
Gossypium raimondii	0.240	0.199	0.044
Homo sapiens	0.713	0.086	0.502
Lepisosteus oculatus	0.051	0.325	0.028
Macaca fascicularis	0.612	0.109	0.979
Macaca mulatta	0.595	0.093	0.778
Malus domestica	0.900	0.056	0.023
Medicago truncatula	1.206	0.156	0.014
Meleagris gallopavo	0.088	0.035	0.162
Microtus ochrogaster	3.528	0.319	1.483
Monodelphis domestica	2.447	0.256	2.066
Mus musculus	5.061	0.812	2.510
Nomascus leucogenys	0.541	0.066	0.494
Oreochromis niloticus	1.684	0.125	0.292
Ornithorhynchus anatinus	0.223	0.090	0.262
Oryctolagus cuniculus	1.086	0.044	0.378
Oryza sativa Japonica	0.859	0.133	0.069
Oryzias latipes	0.188	0.061	0.827
Ovis aries	0.506	0.060	0.033
Pan paniscus	0.503	0.070	0.355
Pan troglodytes	0.501	0.072	0.371
Papio anubis	0.462	0.082	0.774
Phaseolus vulgaris	0.571	0.134	0.005
Poecilia reticulata	1.296	0.413	1.182
Pongo abelii	0.432	0.068	0.320
Populus trichocarpa	1.289	0.210	0.018
Rattus norvegicus	5.972	0.570	1.423
Salmo salar	6.069	0.089	1.268
Setaria italica	0.225	0.033	0.022
Solanum lycopersicum	0.878	0.176	0.034
Solanum pennellii	0.634	0.331	0.135
Sorghum bicolor	0.577	0.204	0.058
Sus scrofa	0.779	0.048	0.455
Taeniopygia guttata	0.124	0.079	0.214
Theobroma cacao	0.562	0.048	0.011
Vigna radiata	3.389	0.251	0.023
Vitis vinifera	0.987	0.284	0.035
Zea mays	0.170	0.038	0.008
Means $\bar{r}(D^+), \bar{r}(T^+), \bar{r}(\mathcal{T}^+)$	1.203	0.240	0.502
Medians $\tilde{r}(D^+), \tilde{r}(T^+), \tilde{r}(\mathcal{T}^+)$	0.634	0.096	0.214

Table 4
Repeated trinucleotides in the X motifs in the genomes of eukaryotes. The 1st, 2nd and 3rd columns give the genome $\mathcal{G}$, its chromosome number $\mathcal{G}_{Chr}$ and its base size $N(\mathcal{G}_{Chr})$, respectively, the 4th column shows the X motif containing repeated trinucleotides, the 5th and 6th columns indicate the start and end positions of the X motif in the chromosome $\mathcal{G}_{Chr}$ and the 7th column gives the nucleotide length l of the X motif.

Genome $\mathcal{G}$	$\mathcal{G}_{Chr}$	Size $N(\mathcal{G}_{Chr})$ (in bases)	Repeated trinucleotides in the X motifs	Start position	End position	Length l of X motifs (in bases)
Homo sapiens	1	248956422	CTG,GCC,GTT,GTC, $(ACC)^{30}$	161051956	161052057	102
Homo sapiens	2	242193529	$(GAA)^{11}$, $(GAC)^3$,AAC, $(GGT)^2$,GAG	20102731	20102784	54
Homo sapiens	3	198295559	$(GGT)^7$,GAC,AAT,GAT, $(GAA)^2$	182632405	182632440	36
Homo sapiens	5	181538259	$(GGC)^{16}$,GTA,GCC,GTA,GAG,GGT,GAG	443230	443295	66
Saccharomyces cerevisiae	I	230218	ACC,GCC, $(GTT)^9$,ATT, $(GTT)^2$,ATT, $(GTT)^2$,ATT, $(GTT)^2$,ATC	113044	113097	54
Saccharomyces cerevisiae	XV	1091291	GTC, $(ATC)^9$,ACC, $(ATC)^2$, $(ATT)^3$,GGT	63055	63105	51
Mus musculus	1	195471971	CAG,GTC, $(TTC)^{21}$, $(CTC)^{11}$,CTG	15964453	15964557	105
Mus musculus	1	195471971	TTC,CAG,GGC, $(ATC)^5$, $(ATT)^{14}$	23812669	23812734	66
Zea mays	1	301433382	GCC,GTC,ACC,GTC,ACC,GTC,GCC,ACC, $(GTC)^8$,CTC,ATC,CTC,GTC,GCC,$(GTC)^2$	11375278	11375346	69
Zea mays	1	301433382	GAC,GGC,AAC,GAG,GAC,GAG, $(GAC)^5$,GGC, $(GAC)^4$,GGT,GGC,GAC,GGT,GAC,GGC	14012017	14012082	66

3.7. Identical trinucleotide pairs of the maximal C^3 self-complementary trinucleotide circular code X preferentially used in the gene sequences of eukaryotes

Unitary circular codes (UCC) of dinucleotides, trinucleotides and tetranucleotides are associated with the repeated dinucleotides (D^+ motifs, Eq. (4)), the repeated trinucleotides (T^+ motifs, Eq. (5)) and the repeated tetranucleotides ($\mathcal{T}^+$ motifs, Eq. (6)) which are identified in the genomes of eukaryotes. Furthermore, there is a scarcity of T^+ motifs in the large eukaryotic genomes compared to the D^+ and $\mathcal{T}^+$ motifs (Section 3.5). Otherwise, a circular code X is observed in genes of bacteria, eukaryotes, plasmids and viruses (Michel, 2015; Arquès and Michel, 1996) which is a set of 20 trinucleotides (Eq. (1)). The problem investigated here is whether the unitary circular codes of trinucleotides in genomes may have some traces in the trinucleotide circular code X in genes.

Fig. 16 identifies a new property of the circular code X. Indeed, by varying the 20 trinucleotides $t' \in X$ for a given trinucleotide $t \in X$, the medians $\tilde{r}(tt')$ (Eq. (23)) of the observed/theoretical ratios $r(tt', \mathcal{G}_{GS})$ (Eq. (21)) of the 400 trinucleotide pairs $tt' \in X^2$ in all the gene sequences $\mathcal{G}_{GS}$ of eukaryotic genomes identify 14 trinucleotide pairs such that the values of $\tilde{r}(tt')$ are maximal when the trinucleotide $t' = t$. These 14 identical trinucleotide pairs tt are described according to t as follows:

$$t \in X' = \{AAC, ACC, ATC, CAG, CTC, CTG, GAA, GAG, GAT, GGT, GTA, GTT, TAC, TTC\} \quad (25)$$

where X' is a subset of X.

The eight trinucleotide pairs $ACCACC$, $ATCATC$, $CTCCTC$, $GAAGAA$, $GAGGAG$, $GATGAT$, $GGTGGT$ and $TTCTTC$ have the eight highest values $\tilde{r}(tt) \geq 1.90$ (Table 5) among all the 400 trinucleotide pairs $tt' \in X^2$ (data not shown). The six other trinucleotide pairs $AACAAC$, $CAGCAG$, $CTGCTG$, $GTAGTA$, $GTTGTT$ and $TACTAC$ have values $\tilde{r}(tt) \geq 1.57$ (Table 5) and belong to the rank interval [9..25] among the 400 trinucleotide pairs $tt' \in X^2$ (data not shown). The six trinucleotide pairs $AATAAT$ (186th rank), $ATTATT$ (167th rank), $GACGAC$ (70th rank), $GCCGCC$ (65th rank), $GGCGGC$ (64th rank) and $GTCGTC$ (85th rank) with values $\tilde{r}(tt)$ close to 1 (see Remark 5) and t-values less than 10 (Table 5) do not have a particular statistical distribution.

Surprisingly, as with the circular code X, the trinucleotide set X' is also self-complementary, i.e. $X' = \mathcal{C}(X')$. All these results are retrieved with the two other ratios $r(tt')$ (Eq. (17)) and $\bar{r}(tt')$ (Eq. (22)) (results not shown). Thus, with a few exceptions, identical trinucleotide pairs of the circular code X are preferentially used in the eukaryotic gene sequences.

Remark 8. 199 trinucleotide pairs have values $\tilde{r}(tt') < 1.00$ in the gene sequences of eukaryotes. The 10 trinucleotide pairs having the 10 lowest values $\tilde{r}(tt') \leq 0.52$ among the 400 trinucleotide pairs $tt' \in X^2$ in the gene sequences of eukaryotes are: $TTCGAA$ with $\tilde{r}(tt') = 0.38$, $TACGTA$, $CTCGAA$, $TTCGAG$, $GTTAAC$, $TTCGGT$, $ACCGAA$, $ATCGAA$, $TTCGAT$ and $TACGTT$ with $\tilde{r}(tt') = 0.52$.

Very interestingly, the 14 trinucleotides $t \in X'$ (Eq. (25)) identified in the gene sequences of eukaryotes are associated to the repeated trinucleotides of high occurrences in the eukaryotic genomes (Fig. 14a,b) by excluding the two repeats $(AAT)^+$ and $(ATT)^+$ of highest occurrences and with a particular statistical distribution compared to the other repeats (Fig. 14a,b) (note that $CAG \in X'$ and $AGC = \mathcal{P}(CAG)$ belong to the same equivalence class by the circular permutation map $\mathcal{P}$, and similarly for $CTC \in X'$ and $CCT = \mathcal{P}^2(CTC)$, $GAA \in X'$ and $AAG = \mathcal{P}(GAA)$, $GAG \in X'$ and $AGG = \mathcal{P}(GAG)$, $GAT \in X'$ and $ATG = \mathcal{P}(GAT)$, $GTA \in X'$ and $AGT = \mathcal{P}^2(GTA)$, $TAC \in X'$ and $ACT = \mathcal{P}(TAC)$ and $TTC \in X'$ and $CTT = \mathcal{P}^2(TTC)$).

4. Discussion

A maximal C^3 self-complementary trinucleotide circular code is identified in genes of bacteria, eukaryotes, plasmids and viruses (Michel, 2015; Arquès and Michel, 1996). X motifs, i.e. motifs from this circular code X, are found in (i) genes of several kingdoms; (ii) tRNAs of prokaryotes and eukaryotes; (iii) 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; and (iii) genomes (non-coding regions) of eukaryotes (Arquès and Michel, 1996; Michel, 2012, 2013, 2015; El Soufi and Michel, 2014, 2015, 2016). These X motifs have the circular code property for retrieving, maintaining and synchronizing 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), and retrieving the two reading frames and the four shifted frames. All these properties suggest a possible translation (framing) code in genes based on the circular code (Michel, 2012).

The origin of this trinucleotide circular code X in genes is an open problem since its discovery in 1996. We show here that the circular code concept, originally found in genes, exists in the eukaryotic genomes with the unitary circular codes (UCC) of dinucleotides, trinucleotides and tetranucleotides generating UCC motifs of repeated dinucleotides (D^+ motifs, Eq. (4)), repeated trinucleotides (T^+ motifs, Eq. (5)) and repeated tetranucleotides ($\mathcal{T}^+$ motifs, Eq. (6)). More precisely, the 12 unitary circular codes of dinucleotides are, also, strong comma-free, where four of them, $\{AT\}$, $\{CG\}$, $\{GC\}$ and $\{TA\}$, are self-complementary (Section 2.2.1). 48 unitary circular codes of trinucleotides are, also, strong comma-free and 12 unitary circular codes of trinucleotides are also comma-free (Section 2.2.2). 180 unitary circular codes of tetranucleotides are, also, strong comma-free, 12 of them $\{AATT\}$, $\{ACGT\}$, $\{AGCT\}$, $\{CATG\}$, $\{CCGG\}$, $\{CTAG\}$, $\{GATC\}$, $\{GGCC\}$, $\{GTAC\}$, $\{TCGA\}$, $\{TGCA\}$ and $\{TTAA\}$ being self-complementary and 60 unitary circular codes of tetranucleotides are, also, comma-free (Section 2.2.3). Thus, the D^+, T^+ and $\mathcal{T}^+$ motifs and their C^2, C^3 and C^4 properties (Definition 8) allow to retrieve, maintain and synchronize a frame modulo 2, modulo 3 and modulo 4, respectively, and their shifted frames (1 modulo 2; 1 and 2 modulo 3; 1, 2 and 3 modulo 4) in the DNA sequences of eukaryotic genomes. The D^+ and $\mathcal{T}^+$ motifs that are self-complementary (Definition 7) allows DNA-DNA and DNA-RNA pairing in the DNA sequences of eukaryotic genomes. A UCC motif and its complementary UCC motif have the same distribution in eukaryotic genomes, both from their occurrence number (Eq. (7)) and their total base number (Eq. (8)). This property is observed with the D^+, T^+ and $\mathcal{T}^+$ motifs (Sections 3.1, 3.2 and 3.3; Figs. 13a,b, 14a,b and 15a,b; Table 1). In addition for the T^+ and $\mathcal{T}^+$ motifs, a UCC motif and its complementary UCC motif have increasing occurrences contrary to their number of hydrogen bonds (Sections 3.2 and 3.3; Figs. 14a,b and 15a,b). For the D^+ motifs, the repeat $(CG)^+$ has the lowest occurrence but the repeat $(AT)^+$ has not the highest occurrence (3rd occurrence in Fig. 13a,b). The largest nucleotide lengths of D^+, T^+ and $\mathcal{T}^+$ motifs in the studied eukaryotic genomes are given in Table 2. Surprisingly, a scarcity of repeated trinucleotides (T^+ motifs) in the large eukaryotic genomes is observed compared to the D^+ and $\mathcal{T}^+$ motifs (Section 3.5; Table 3). This statistical result is found with the mean and the median (Eqs. (11) and (12)) and confirmed by two statistical tests (a paired sample Student's t-test and a Wilcoxon signed-rank W-test). The scarcity of repeated trinucleotides in the large genomes of eukaryotes may be explained by the two following

Fig. 16. 14 identical trinucleotide pairs (Eq. (25)) of the maximal C^3 self-complementary trinucleotide circular code X preferentially used in the eukaryotic gene sequences (see Appendix A). Median $\tilde{r}(tt')$ (Eq. (23)) of the observed/theoretical ratios $r(tt', \mathcal{G}_{GS})$ (Eq. (21)) of the 400 trinucleotide pairs $tt' \in X^2$ in all the gene sequences $\mathcal{G}_{GS}$ of eukaryotic genomes. Each figure gives in ordinate the median $\tilde{r}(tt')$ of a trinucleotide $t \in X$ (in label) by varying the 20 trinucleotides $t' \in X$ in abscissa.

Table 5
Statistical significance of the mean $\bar{r}(tt)$ (Eq. (22)) of the observed/theoretical ratios $r(tt, \mathcal{G}_{GS})$ (Eq. (21)) of the 20 trinucleotide pairs $tt \in X^2$ in all the gene sequences $\mathcal{G}_{GS}$ of eukaryotic genomes evaluated by a single sample Student t-test which determines whether the sample mean is statistically different from 1 (see Remark 5). The six trinucleotide pairs $tt \in \{AATAAT, ATTATT, GACGAC, GCCGCC, GGCGGC, GTCGTC\}$ have t-values less than 10 (in italics).

tt	Median $\tilde{r}(tt')$ (Eq. (23))	Mean $\bar{r}(tt')$ (Eq. (22))	Standard deviation	t-value
AACAAC	1.70	1.77	0.43	21.7
AATAAT	1.03	1.03	0.28	*1.2* (p-value = 0.25)
ACCACC	2.26	2.36	0.52	31.6
ATCATC	1.93	1.87	0.34	30.8
ATTATT	1.06	1.03	0.28	*1.4* (p-value = 0.16)
CAGCAG	1.79	1.95	0.76	15.1
CTCCTC	1.90	1.97	0.53	21.7
CTGCTG	1.71	1.89	0.82	13.1
GAAGAA	2.03	1.97	0.32	36.4
GACGAC	1.31	1.35	0.57	*7.4*
GAGGAG	1.98	2.00	0.54	22.1
GATGAT	1.93	1.88	0.35	30.2
GCCGCC	1.32	1.51	0.72	*8.5*
GGCGGC	1.33	1.52	0.68	*9.1*
GGTGGT	2.25	2.38	0.53	31.4
GTAGTA	1.57	1.61	0.32	22.8
GTCGTC	1.29	1.34	0.59	*6.9*
GTTGTT	1.69	1.76	0.41	22.0
TACTAC	1.65	1.67	0.36	22.1
TTCTTC	1.99	1.98	0.32	36.2

results. X motifs of low composition in particular, which are abundant in the eukaryotic genomes (El Soufi and Michel, 2016), contain repeated trinucleotides (Section 3.6). Identical trinucleotide pairs of the circular code X are preferentially used in the eukaryotic gene sequences (Section 3.7). Indeed, 14 trinucleotides (Eq. (25)) among 20 of the circular code X are preferentially followed by itself in the eukaryotic gene sequences. This statistical result is observed with three ratios (Eqs. (17), (22) and (23)). Thus, some statistical properties of repeated trinucleotides are persistent in the circular code X.

In conclusion, the unitary circular codes of trinucleotides in eukaryotic genomes may have been involved in the formation of the trinucleotide circular code X in genes. Indeed, repeated trinucleotides in the X motifs in the genomes of eukaryotes may represent an intermediary evolution from repeated trinucleotides of cardinality 1 (T^+ motifs) in the genomes of eukaryotes up to the X motifs of maximal cardinality 20 in the gene sequences of eukaryotes. For the first time since 20 years, the circular code theory in genes is extended here to genomes. Circular code could be a mathematical structure of genes as well as genomes.

Appendix A. Data of eukaryotic genomes

List of 126 complete eukaryotic genome $\mathcal{G}$, size $N(\mathcal{G}_{GS})$ in trinucleotides (for Eq. (19)) of the gene sequences $\mathcal{G}_{GS}$ of $\mathcal{G}$ and size $N(\mathcal{G})$ in bases (for Eq. (10)) of $\mathcal{G}$ extracted from the GenBank (http://www.ncbi.nlm.nih.gov/genbank/, June 2016):

Genomes $\mathcal{G}$	Size $N(\mathcal{G}_{GS})$ of gene sequences $\mathcal{G}_{GS}$ in trinucleotides	Size $N(\mathcal{G})$ of genome $\mathcal{G}$ in bases
Anolis carolinensis	11706361	1081644591
Anopheles gambiae	751389	24393108
Apis mellifera	13620545	219629612
Arabidopsis thaliana	14320038	119146348
Aspergillus fumigatus	4727465	29384958
Babesia bigemina	2263900	10271324
Beta vulgaris	11150232	376583697
Bombus terrestris	11910477	216849342
Bos taurus	17879564	2715765904
Brachypodium distachyon	15926201	271776478
Brassica napus	39633312	775113993
Brassica oleracea	22543289	446885882
Brassica rapa	20606555	256423463
Caenorhabditis briggsae	6776159	91234787
Caenorhabditis elegans	13157150	100272607
Callithrix jacchus	24520274	2770219215
Camelina sativa	42568721	608544003
Candida dubliniensis	2969509	14618422
Candida orthopsilosis	2818046	12659401
Canis lupus familiaris	31183423	2327633984
Capra hircus	16968387	2524662720
Chlorocebus sabaeus	44137205	2744115311
Chrysemys picta bellii	5137627	461747357
Cicer arietinum	13439237	347247377
Ciona intestinalis	5660346	78296155
Citrus sinensis	14131072	238999708
Cryptococcus gattii	3383994	18374760
Cryptococcus neoformans	3542591	19699782
Cryptosporidium parvum	2270778	9102324
Cucumis sativus	11974205	191859024
Cyanidioschyzon merolae	2475860	16546747
Cynoglossus semilaevis	18995753	445139357
Danio rerio	33008958	1340430591
Debaryomyces hansenii	3003875	12152486
Dictyostelium discoideum	6991644	33943072
Drosophila melanogaster	4135924	28557754
Drosophila pseudoobscura	7072517	50607275
Drosophila simulans	752899	17992287
Drosophila yakuba	2943314	23145337
Elaeis guineensis	14100095	657968836
Equus caballus	20967575	2367053447
Eremothecium cymbalariae	2154960	9669424
Esox lucius	24553197	701024151
Felis catus	19356838	2419212910
Ficedula albicollis	15259561	1044065291
Fragaria vesca	13793361	198117109
Gallus gallus	30345321	1021439028
Glycine max	32356518	949176042
Gorilla gorilla gorilla	16110141	2917687013
Gossypium raimondii	26666857	749228090
Homo sapiens	61732808	3088269832
Kazachstania africana	2612751	11130140
Kluyveromyces lactis	2459833	10689156
Leishmania braziliensis	5061373	31238104
Leishmania donovani	4924699	32444968
Leishmania infantum	5179812	31924853
Leishmania major	5231237	32855089
Leishmania mexicana	5058215	30937689
Leishmania panamensis	4842270	30688794
Lepisosteus oculatus	28333784	891144077
Macaca fascicularis	44026530	2871826009
Macaca mulatta	36978068	2835963390
Magnaporthe oryzae	5629242	40491973
Malus domestica	14990689	526197889
Medicago truncatula	18965012	384466993
Meleagris gallopavo	13804201	972203167
Micromonas	4869779	20989326
Microtus ochrogaster	13580395	1655383507
Monodelphis domestica	21250159	2754317877
Mus musculus	20048765	1205572488
Myceliophthora thermophila	1987996	16385300
Nasonia vitripennis	10169009	116029644
Naumovozyma castellii	2768311	11219539
Naumovozyma dairenensis	2869045	13527580
Neospora caninum	5915585	57547420
Neurospora crassa	5612485	40463072
Nomascus leucogenys	18369180	2795260045
Oreochromis niloticus	25308300	657350972
Ornithorhynchus anatinus	2332329	437080024
Oryctolagus cuniculus	17399825	2247752104
Oryza brachyantha	11811292	250923338
Oryza sativa Japonica Group	10201970	382150945
Oryzias latipes	19264271	723441489
Ostreococcus lucimarinus	3066943	13204888
Ostreococcus tauri	3372403	12456351
Ovis aries	27727712	2584815894
Pan paniscus	27628433	3151907227
Pan troglodytes	33498806	3091112213
Papio anubis	30756505	2724327674
Phaeodactylum tricornutum	4657840	26138756
Phaseolus vulgaris	13558399	514820528
Plasmodium cynomolgi strain B	3112005	22728335
Plasmodium falciparum	4078192	23264338
Plasmodium knowlesi strain H	3700350	23462187
Plasmodium vivax	3630043	22621071
Poecilia reticulata	29389640	696700953
Pongo abelii	17696459	3029491029
Populus trichocarpa	16350699	378545565
Prunus mume	11628870	198852406
Rattus norvegicus	30376473	2782012602
Saccharomyces cerevisiae	2914027	12071326
Salmo salar	64043007	2240204991
Scheffersomyces stipitis	2858379	15441179
Sesamum indicum	14672059	233222381
Setaria italica	15111311	401296418
Solanum lycopersicum	16407523	802138220
Solanum pennellii	15949831	926426464
Sorghum bicolor	12421163	659229367
Sus scrofa	24119139	2596639456
Taeniopygia guttata	9764512	1021462940
Takifugu rubripes	15387600	281572362
Tetrapisispora blattae	2914869	14048593
Tetrapisispora phaffii	2674494	12100190
Thalassiosira pseudonana	5194718	28733535
Theobroma cacao	19148424	330456197
Thielavia terrestris	4527921	36912256
Torulaspora delbrueckii	2413026	9220678
Tribolium castaneum	10388047	187494969
Trypanosoma brucei gambiense	4424477	22148088
Ustilago maydis	3995688	19643891
Vigna radiata	13618511	333308464

Vitis vinifera	17096642	426176009
Yarrowia lipolytica	3139371	20502981
Zea mays	22929943	2059701728
Zygosaccharomyces rouxii	2475704	9764635
Zymoseptoria tritici	4782485	39686251
Total	1757915083	91350244263

References

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

Bussoli, L., Michel, C.J., Pirillo, G., 2012. On conjugation partitions of sets of trinucleotides. Appl. Math. 3, 107–112.

Canapa, A., Cerioni, P.N., Barucca, M., Olmo, E., Caputo, V., 2002. A centromeric satellite DNA may be involved in heterochromatin compactness in gobiid fishes. Chromosome Res. 10, 297–304.

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.

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

Fimmel, E., Michel, C.J., Strüngmann, L., 2016. *n*-Nucleotide circular codes in graph theory. Philosophical transactions of the royal society A: mathematical. Phys. Eng. Sci. 374, 20150058.

Fimmel E., Michel C.J., Strüngmann L., 2017. Diletter circular codes over finite alphabets, submitted.

Gemayel, R., Vinces, M.D., Legendre, M., Verstrepen, K.J., 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu. Rev. Genet. 44, 445–477.

Golomb, S.W., Delbruck, M., Welch, L.R., 1958a. Construction and properties of comma-free codes. Biologiske Meddelelser Kongelige Danske Videnskabernes Selskab 23, 1–34.

Golomb, S.W., Gordon, B., Welch, L.R., 1958b. Comma-free codes. Can. J. Math. 10, 202–209.

Gonzalez, D.L., Giannerini, S., Rosa, R., 2012. On the origin of the mitochondrial genetic code: towards a unified mathematical framework for the management of genetic information. Nature precedings 1, http://dx.doi.org/10.1038/npre.2012.7136.

Jeffreys, A.J., Neil, D.L., Neumann, R., 1998. Repeat instability at human minisatellites arising from meiotic recombination. EMBO J. 17, 4147–4157.

Li, Y.-C., Korol, A.B., Fahima, T., Nevo, E., 2004. Microsatellites within genes: structure, function, and evolution. Mol. Biol. Evol. 21, 991–1007.

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

Michel, C.J., Pirillo, G., 2011. Strong trinucleotide circular codes. Int. J. Combinatorics 2011, 1–14 (Article ID 659567).

Michel, C.J., Pirillo, G., Pirillo, M.A., 2008a. Varieties of comma free codes. Comput. Math. Appl. 55, 989–996.

Michel, C.J., Pirillo, G., Pirillo, M.A., 2008b. A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theor. Comput. Sci. 401, 17–26.

Michel, C.J., Pirillo, G., Pirillo, M.A., 2012. A classification of 20-trinucleotide circular codes. Inf. Comput. 212, 55–63.

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., 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.

Ohno, S., Epplen, J.T., 1983. The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. Proc. Natl. Acad. Sci. U. S. A. 80, 3391–3395.

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.

Rich, A., Nordheim, A., Wang, A.H.J., 1984. The chemistry and biology of left-handed Z-DNA. Annu. Rev. Biochem. 53, 791–846.

Toll-Riera, M., Radó-Trilla, N., Martys, F., Albà, M.M., 2012. Role of low-complexity sequences in the formation of novel protein coding sequences. Mol. Biol. Evol. 29, 883–886.