Mixed circular codes

Elena Fimmel, Christian Michel, François Pirot, Jean-Sébastien Sereni, Lutz Strüngmann

To cite this version:

Elena Fimmel, Christian Michel, François Pirot, Jean-Sébastien Sereni, Lutz Strüngmann. Mixed circular codes. Mathematical Biosciences, 2019, 317, pp.108231. 10.1016/j.mbs.2019.108231 . hal-02188407v2

HAL Id: hal-02188407
https://hal.science/hal-02188407v2

Submitted on 18 Nov 2019

MIXED CIRCULAR CODES

ELENA FIMMEL[1], CHRISTIAN J. MICHEL[2,*], FRANÇOIS PIROT[2,3], JEAN-SÉBASTIEN SERENI[2] AND LUTZ STRÜNGMANN[1]

[1]*Institute of Mathematical Biology*
Faculty for Computer Sciences
Mannheim University of Applied Sciences
68163 Mannheim, Germany

[2]*Theoretical Bioinformatics, ICube,*
C.N.R.S., University of Strasbourg,
300 Boulevard Sébastien Brant
67400 Illkirch, France
[*]*Corresponding author*

[3]*LORIA (Orpailleur) and Dept. of Mathematics*
University of Lorraine and Radboud University
Vandœuvre-lès-Nancy, France and Nijmegen, Netherlands

Abstract. By an extensive statistical analysis in genes of bacteria, archaea, eukaryotes, plasmids and viruses, a maximal C^3-self-complementary trinucleotide circular code has been found to have the highest average occurrence in the reading frame of the ribosome during translation. Circular codes may play an important role in maintaining the correct reading frame. On the other hand, as several evolutionary theories propose primeval codes based on dinucleotides, trinucleotides and tetranucleotides, mixed circular codes are investigated.

By using a graph-theoretical approach of circular codes recently developed, we study mixed circular codes, which are the union of a dinucleotide circular code, a trinucleotide circular code and a tetranucleotide circular code. Maximal mixed circular codes of (di,tri)-nucleotides, (tri,tetra)-nucleotides and (di,tri,tetra)-nucleotides are constructed, respectively. In particular, we show that any maximal dinucleotide circular code of size 6 can be embedded into a maximal mixed (di,tri)-nucleotide circular code such that its trinucleotide component is a maximal C^3-comma-free code. The growth function of self-complementary mixed circular codes of dinucleotides and trinucleotides is given. Self-complementary mixed circular codes could have been involved in primitive genetic processes.

1. Introduction

The genomes of all species, *i.e.* archaea, bacteria, eukaryota, viruses, plasmids, mitochondria and chloroplasts, contain regions for coding proteins, *i.e.* a series of amino acids. These genomic

E-mail address: e.fimmel@hs-mannheim.de, l.struengmann@hs-mannheim.de, c.michel@unistra.fr, sereni@kam.mff.cuni.cz, francois.pirot@loria.fr.

Date: October 28, 2019.

Key words and phrases. mixed circular codes, dinucleotide circular codes, trinucleotide circular codes, tetranucleotide circular codes, graph properties.

regions are called (protein coding) genes. A gene is a series of words of the same length equal to 3 nucleotides, *i.e.* a series of trinucleotides, also called *codons*. This genetic information can be easily revealed by signal processing, hence without any biological experimental method. For example, the correlation function and the power spectrum identify a nucleotide periodicity modulo 3 in genes (Shepherd, 1981; Fickett, 1982; Michel, 1986, Figure 1). Ten years later, this periodicity modulo 3 has been explained by a maximal C^3-self-complementary trinucleotide circular code which has been found to have the highest average occurrence in the reading frame of the ribosome during translation, compared to the two shifted frames, of genes of bacteria, archaea, eukaryotes, plasmids and viruses (Michel, 2017, 2015; Arquès and Michel, 1996). It contains the following 20 trinucleotides

$$\begin{aligned} X = \{&AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC,\\ &GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\} \end{aligned} \tag{1}$$

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

$$\{\text{Ala,Asn,Asp,Gln,Glu,Gly,Ile,Leu,Phe,Thr,Tyr,Val}\},$$

$$\{A, N, D, Q, E, G, I, L, F, T, Y, V\}.$$

Two reviews on circular codes, separated by ten years, present the scientific progress in this research field (Michel, 2008; Fimmel and Strüngmann, 2018). In particular, these reviews describe the new mathematical approaches using group theory and graph theory, the hierarchies of circular codes: strong comma-free, comma-free and general circular, and the properties of the different classes of circular codes: maximality, self-complementarity, growth functions and number of encoded amino acids, for example.

Bacterial, viral and organelle genomes all possess a compact architecture where genes represent about 90% of a genome, in contrast to eukaryotes where genes only constitute 10±5% of a genome (Bobay and Ochman, 2017). Non-coding regions of eukaryotic genomes contain different DNA structures along the chromosome: pseudogenes, RNA-coding genes, introns, tandem repeats: minisatellites and microsatellites, retrotransposons: long terminal repeats (LTR), non-long terminal repeats (Non-LTR), long interspersed elements (LINE) and short interspersed elements (SINE), DNA transposons, *etc.* Introns of eukaryotes have no nucleotide periodicity modulo 3 (Figure 2 in Michel, 1986) but a nucleotide periodicity modulo 2 generated by dinucleotides (Konopka and Smythers, 1987; Arquès and Michel, 1987). Furthermore, in intergenic regions of eukaryotes, pure and mixed, as well as short and long, repeated dinucleotides, trinucleotides and tetranucleotides are very common (consult, among others, Canapa *et al.*, 2002; Ellegren, 2004; Gemayel *et al.*, 2010; El Soufi and Michel, 2017). Some dinucleotide repeats are highly enriched in enhancers which are genomic elements involved in gene expression (Yáñez-Cuna *et al.*, 2014). However, most of these genetic structures have an unknown biological activity and their functions remain an important open question in today's biology. This genetic information could in part be related to traces of ancestral codes (El Soufi and Michel, 2017). Indeed, in recent years, various authors have formulated hypotheses about the origin of the modern genetic code in which it is assumed that single amino acids were first encoded by dinucleotides or tetranucleotides, and then later by trinucleotides (see, among others, Baranov *et al.*, 2009; Gonzalez *et al.*, 2012; Patel, 2005; Seligmann, 2014; Wilhelm and Nikolaeva, 2004; Wu *et al.*, 2005). This coding process could have been necessary because the number of amino acids to be encoded increased during the evolution while at the same time nature needed robustness against mutations and translational errors. As several biochemical mechanisms would have been used for reading oligonucleotides of different lengths ("primitive codons" of different lengths), we are interested in mixed circular codes, which allow to generalize the reading frame retrieval to words of different sizes.

We indeed prove that mixed codes can be circular, and we identify several of their mathematical properties. Moreover, we formulate criteria for mixed codes of di-, tri- and tetranucleotides to be circular and show how such codes, even of maximal possible size, can be constructed. We also mention some biological implications of mixed circular codes, which could have been involved in primitive gene coding processes.

2. Definitions and examples

Let us first state some definitions and results that will be used in the sequel. The *genetic alphabet* is

$$\mathcal{B} := \{A, C, G, T\}$$

where A stands for *adenine*, C for *cytosine*, G for *guanine* and T for *thymine*. As commonly used in word theory, $\mathcal{B}^* = \{N_1 \cdots N_n \mid N_i \in \mathcal{B}, n \in \mathbf{N}\}$ is the set of all *words* over $\mathcal{B}$ of finite length including the *empty word* ε while $\mathcal{B}^+ = \mathcal{B}^* \backslash \{\varepsilon\}$.

Definition 1.

(1) A set $X \subseteq \mathcal{B}^*$ is a *code* if every word $w \in X^*$ has a single decomposition into words from X, *i.e.* whenever $w = w_1 \cdots w_n = w'_1 \cdots w'_m$ for words $w_i, w'_j \in X$, then $m = n$ and $w_i = w'_i$ follow.

(2) For $\ell \in \mathbf{N}$ with $\ell \geq 2$, an *ℓ-letter code* is a subset of $\mathcal{B}^\ell$.

(3) Elements of $\mathcal{B}^\ell$ are called *ℓ-nucleotides.*

(4) 2-nucleotides, 3-nucleotides and 4-nucleotides are also called *dinucleotides*, *trinucleotides* and *tetranucleotides*, respectively.

(5) Given two finite words w_1 and w_2 in B^*, we define $w_1 \sqcap w_2$ to be the largest tail segment (*suffix*) of w_1 that is an initial segment (*prefix*) of w_2. (Note that $w_1 \sqcap w_2$ might be the empty word ε.) If $w_1 \neq w_2$ and $w_1 \sqcap w_2 \neq \varepsilon$, then w_1 *overlaps* w_2.

Remark 1. Any set of ℓ-nucleotides is always a code, for example the genetic code $\mathcal{B}^3$ and the sets $\mathcal{B}^2$ of dinucleotides and $\mathcal{B}^4$ of tetranucleotides. However, this property is not always true when allowing *mixed* sets containing ℓ-nucleotides for several values of ℓ.

Example 1. The mixed set $X = \{AC, GA, GT, ACG, TGA\}$ is not a code since the word $ACGTGA$ has two different decompositions into words from X, namely

$$ACG \mid TGA = AC \mid GT \mid GA.$$

Some codes may have additional properties.

Definition 2. A code $X \subseteq \mathcal{B}^+$ is

(1) *comma-free* if every concatenation $w_1 w_2$ of two words from X does not contain as a substring any word from X but w_1 as a prefix and w_2 as a suffix, that is, if

$$X^2 \cap \mathcal{B}^+ X \mathcal{B}^+ = \varnothing;$$

(2) *circular* if for any finite concatenation $w_1 \cdots w_m$ of elements from X ($m \in \mathbf{N}$), there is only one partition into elements from X when read on a circle. Any such partition is a *circular decomposition* of $w_1 \cdots w_m$.

Any comma-free code is also circular and the most important property of such codes is that they allow the detection of frameshifts. A major difference between these two classes of circular codes is the nucleotide window length for retrieving the reading frame in genes (Figures 1 and 2).

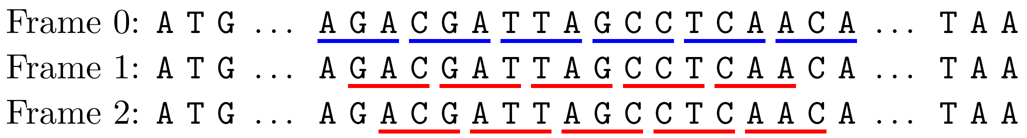
Frame 0: A T G ... A G A C G A T T A G C C T C A A C A ... T A A
Frame 1: A T G ... A G A C G A T T A G C C T C A A C A ... T A A
Frame 2: A T G ... A G A C G A T T A G C C T C A A C A ... T A A

FIGURE 1. Reading frame retrieval in genes with the comma-free code $X = \{ACA, AGA, CGA, GCC, TCA, TTA\}$. The trinucleotides (words of length 3) underlined in blue belong to X, while those underlined in red do not. A frameshift is detected immediately, that is within at most 3 nucleotides.

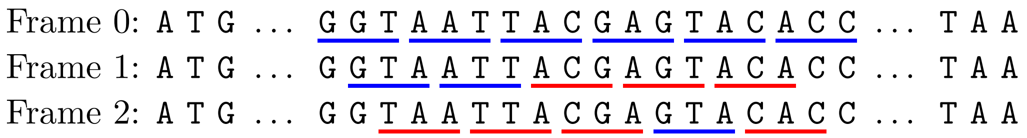
Frame 0: A T G ... G G T A A T T A C G A G T A C A C C ... T A A
Frame 1: A T G ... G G T A A T T A C G A G T A C A C C ... T A A
Frame 2: A T G ... G G T A A T T A C G A G T A C A C C ... T A A

FIGURE 2. Reading frame retrieval in genes with the maximal C^3 self-complementary trinucleotide circular code $X = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$ identified in genes (1). The trinucleotides underlined in blue belong to X, while those underlined in red do not. A frameshift is detected after a few nucleotides, specifically within at most 13 nucleotides.

It directly follows from the definition that a trinucleotide circular code over $\mathcal{B}$ cannot contain the trinucleotides AAA, CCC, GGG and TTT, hence we make the following remark.

Remark 2. The genetic code $\mathcal{B}^3$ as well as the codes $\mathcal{B}^2$ of dinucleotides and $\mathcal{B}^4$ of trinucleotides are (obviously) not circular.

In the following sections, we will need the notions of *self-complementary code* and C^3*-code*. To define them, we introduce a special transformation, the so called *Strong/Weak (SW) or complementing* (c) transformation

$$SW\ (or\ c)\colon (A, T, C, G) \to (T, A, G, C)$$

that exchanges A and T as well as C and G. Regarding the complementary structure of the DNA double helix, this transformation plays an important biological role. The second important biological transformation related to the antiparallel structure of the DNA double helix is the *reversing* permutation, which we indicate by $\overleftarrow{\cdot}$: a given trinucleotide $x = (b_1, b_2, b_3) \in \mathcal{B}^3$ leads to the trinucleotide $\overleftarrow{x} := (b_3, b_2, b_1)$. These two biological maps, the complementary transformation and the reversing permutation, are involved in gene coding.

In general, the *reversing permutation* inverts the order of bases in any ℓ-nucleotide, *i.e.* if $x = N_1N_2 \cdots N_{\ell-1}N_\ell \in \mathcal{B}^\ell$ then $\overleftarrow{x} = N_\ell N_{\ell-1} \cdots N_2N_1 \in \mathcal{B}^\ell$. If X is a code of ℓ-nucleotides, then $\overleftarrow{X} = \{\overleftarrow{x} : x \in X\}$ is the *reversed code* of X. Similarly, the *complementing map* $c\colon \{A, C, G, T\} \to \{A, C, G, T\}$ that exchanges A and T as well as C and G induces the *complemented code*

$c(X) = \{c(x) : x \in X\}$ where $c(N_1N_2 \cdots N_{\ell-1}N_\ell) = c(N_1)c(N_2)\cdots c(N_{\ell-1})c(N_\ell)$ for any ℓ-nucleotide $x \in \mathcal{B}^\ell$. Note that for a trinucleotide $x = N_1N_2N_3$, the anti-trinucleotide of x is exactly $\overleftarrow{c(x)}$.

Example 2. $\overleftarrow{c(\{AC, AT\})} = \{AT, GT\}$ and $\overleftarrow{c(\{CGA, GAT\})} = \{ATC, TCG\}$.

Definition 3.

(1) A code $X \subseteq \mathcal{B}^*$ is *self-complementary* if $X = \overleftarrow{c(X)}$.
(2) A trinucleotide circular code $X \subseteq \mathcal{B}^3$ is called a C^3*-code* if both the circular shifted codes $X' := \{N_2N_3N_1 : N_1N_2N_3 \in X\}$ and $X'' := \{N_3N_1N_2 : N_1N_2N_3 \in X\}$ are circular as well.

Let us remark that the C^3-property means that the code is circular in all three reading frames of the ribosome, hence it may retrieve the correct frame in each of the three frames. For instance, the trinucleotide circular code that was found in genes (1) is a C^3-self-complementary code.

Remark 3. The genetic code $\mathcal{B}^3$ is a self-complementary trinucleotide code.

Definition 4. Let C be a class of circular codes (for instance, circular, or circular and self-complementary, or C^3-code) defined over $\mathcal{B}$.

(1) A code $X \in C$ is *maximal* (in C) if

$$\text{for all } Y \in C, \text{ we have } X \subseteq Y \Rightarrow X = Y.$$

(2) A code $X \in C$ is *maximum* (in C) if

$$\text{for all } Y \in C, \text{ we have } |Y| \leq |X|.$$

In other words, a circular code X of a class C is maximal if it is not properly contained in any other circular code from the same class C. Moreover, it is a maximum circular code if there is no circular code from the same class C that has a strictly larger size than X. Thus, a maximum circular code is already maximal but the converse is not true in general, as we will see in the sequel.

Following earlier works (Fimmel *et al.*, 2016), we define the directed graph $\mathcal{G}(X)$ of some code $X \subseteq \mathcal{B}^\ell$ as follows: for every word $w = N_1 \cdots N_l \in X$ of length ℓ, and for every $i \in \{1, \ldots, \ell-1\}$, we add an edge from the vertex labelled $N_1 \cdots N_i$ to the vertex labelled $N_{i+1} \cdots N_\ell$, creating such vertices if they do not exist already.

Definition 5. Fix $\ell \in \mathbf{N}$ and let $X \subseteq \mathcal{B}^\ell$ be an ℓ-letter code. We define a directed graph $\mathcal{G}(X) = (V(X), E(X))$ with vertex set $V(X)$ and edge set $E(X)$ as follows.

(1) $V(X) := \{N_1 \cdots N_j, N_{j+1} \cdots N_\ell : N_1N_2 \cdots N_\ell \in X, 1 \leq j \leq \ell - 1\}$; and
(2) $E(X) := \{[N_1 \cdots N_j, N_{j+1} \cdots N_\ell] : N_1N_2 \cdots N_\ell \in X, 1 \leq j \leq \ell - 1\}$.

The graph $\mathcal{G}(X)$ is said to be *associated with* X. An example of graph $\mathcal{G}(X)$ is given in Figure 3.

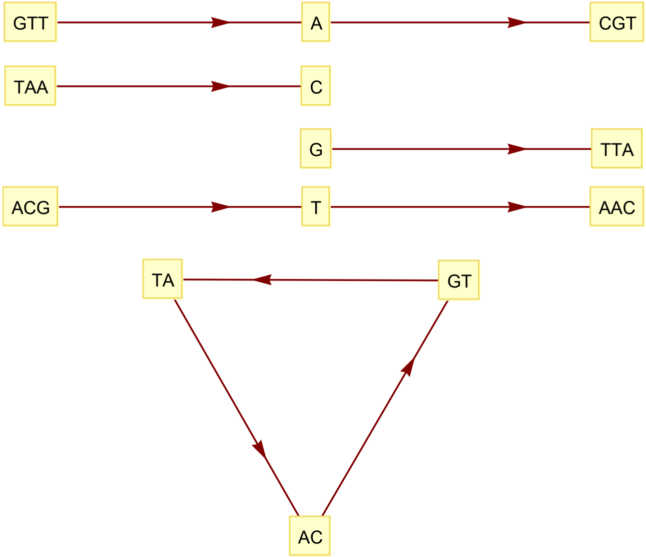

FIGURE 3. Graph $\mathcal{G}(X)$ of the tetranucleotide code $X = \{ACGT, GTTA, TAAC\}$.

The relevance of Definition 5 is witnessed by the following theorem.

Theorem 1 (Fimmel *et al.*, 2016)**.** *The following statements hold for every integer $\ell \geq 2$.*

(1) A code $X \subseteq \mathcal{B}^\ell$ is circular if and only if its associated graph $\mathcal{G}(X)$ is acyclic.

(2) A circular code $X \subseteq \mathcal{B}^\ell$ is comma-free if and only if the longest directed path in its associated (acyclic) graph $\mathcal{G}(X)$ has length at most 2.

Definition 6. For all positive integers ℓ and i such that $1 \leq i \leq \ell$, we define the projection on the ith coordinate $\pi_i \colon \mathcal{B}^\ell \to \mathcal{B}$ by

$$\pi_i(N_1 \cdots N_\ell) = N_i.$$

The projections on two coordinates $\pi_{ij} : \mathcal{B}^\ell \to \mathcal{B}^2$ are defined in a similar way whenever it makes sense.

Example 3. Let $X = \{ACG, AGA, TCT\}$. Then, the projection π_1 on the 1st component yields the set $\pi_1(X) = \{A, T\}$, the projection π_2 on the 2nd component leads to $\pi_2(X) = \{C, G\}$ and the projection π_3 on the 3rd component yields $\pi_3(X) = \{A, G, T\}$.

We investigate the *mixed codes* $X \subseteq \mathcal{B}^2 \cup \mathcal{B}^3 \cup \mathcal{B}^4$ where the elements are dinucleotides, trinucleotides and tetranucleotides. We will also consider mixed codes that contain (di,tri)-nucleotides or (tri,tetra)-nucleotides only.

The first obstacle that appears when considering proper mixed codes is the fact that sets of ℓ-nucleotides of various values of ℓ are not necessarily a code, *i.e.* they may allow words that have two different decompositions over the elements of the code. This is in contrast to ℓ-nucleotide sets (for a fixed ℓ) which are codes necessarily. The next subsection deals with examples showing that a mixing of circular codes over $\mathcal{B}^2$ and $\mathcal{B}^3$ can be a code but not circular, not a code but circular, or neither a code nor circular.

2.1. **Illustrative examples of mixed codes.** Let us consider a mixed code $X = X_2 \cup X_3 \subseteq \mathcal{B}^2 \cup \mathcal{B}^3$ where $X_2 \subseteq \mathcal{B}^2$ and $X_3 \subseteq \mathcal{B}^3$ are both circular codes. The following examples show that the mixed set X can be not circular and even not a code. In a similar way, one can also construct such examples inside $\mathcal{B}^3 \cup \mathcal{B}^4$ or $\mathcal{B}^2 \cup \mathcal{B}^4$ but we will restrict ourselves to (di,tri)-nucleotides to show the ideas.

Example 4. Consider the two circular codes $X_2 = \{GT\}$ and $X_3 = \{ACG, TAC\}$. The mixed set $X = \{GT, ACG, TAC\}$ is a code that is not circular. Indeed, the word $TACGTACG$ admits the two circular decompositions $TAC|GT|ACG$ and $T|ACG|TAC|G$. The graph $\mathcal{G}(X)$ of X (defined naturally for mixed codes; see the next section for a formal definition) contains the cycle $G \to T \to AC \to G$ (Fig. 4).

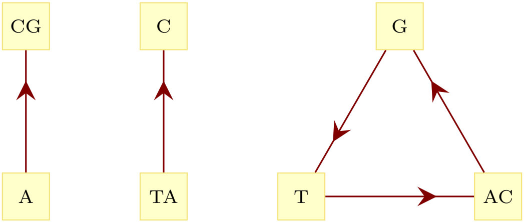

FIGURE 4. The cyclic graph $\mathcal{G}(X)$ of the non-circular mixed code X from Example 4.

Example 5. Consider the two circular codes $X_2 = \{AC, AT, GA\}$ and $X_3 = \{GAA, TAC\}$. The mixed set $X = \{AC, AT, GA, GAA, TAC\}$ is not a code as the word $GAATAC$ has two different decompositions into words in X, namely $GA|AT|AC$ and $GAA|TAC$. However, the graph $\mathcal{G}(X)$ of X (naturally defined; see the next section for a formal definition) is acyclic (Fig. 5).

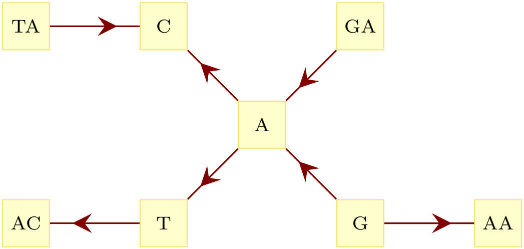

FIGURE 5. The acyclic graph $\mathcal{G}(X)$ of the mixed set X that is not a code from Example 5.

Example 6. Consider the two circular codes $X_2 = \{AC\}$ and $X_3 = \{ACA, CGT, CTG, GAC, GTT, TCT\}$. The mixed set $X = \{AC, ACA, CGT, CTG, GAC, GTT, TCT\}$ is not a code as the word $ACACGTTCTGAC$ has two different decompositions into words in X, namely $AC|AC|GTT|CTG|AC$ and $ACA|CGT|TCT|GAC$. Moreover, the graph $\mathcal{G}(X)$ of X (naturally defined; see the next section for a formal definition) contains the cycle $AC \to A \to C \to GT \to T \to CT \to G \to AC$ (Fig. 6).

In the next section, we give a handy criterion for a mixed set to be a code. Moreover, we construct various examples of maximum and hence maximal mixed circular codes.

3. CONSTRUCTION OF MAXIMUM MIXED CIRCULAR CODES

Let $X_2 \subseteq \mathcal{B}^2$, $X_3 \subseteq \mathcal{B}^3$ and $X_4 \subseteq \mathcal{B}^4$ be circular codes. We want to know whether the mixed set $X := X_2 \cup X_3$, respectively $X := X_3 \cup X_4$, is a circular code. The associated graph $\mathcal{G}(X)$ of X is $\mathcal{G}(X) = \mathcal{G}(X_2) \cup \mathcal{G}(X_3)$, respectively $\mathcal{G}(X) = \mathcal{G}(X_3) \cup \mathcal{G}(X_4)$. The next theorem provides a graph certificate to the fact that the mixed set X is a circular code. Its proof can be found in Appendix 6.1.

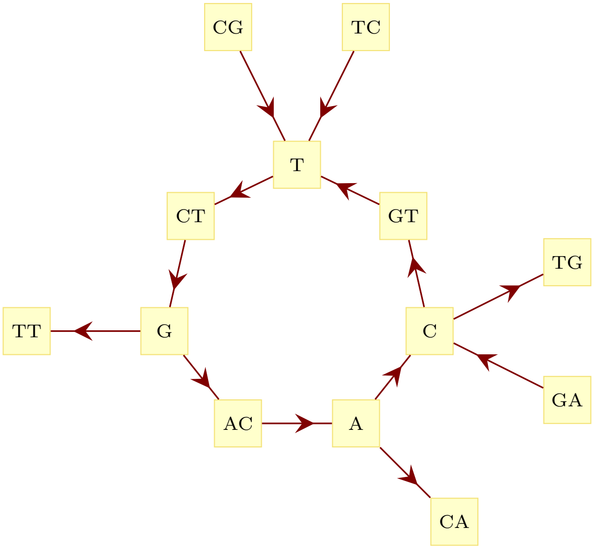

FIGURE 6. The cyclic graph $\mathcal{G}(X)$ of the mixed set X that is not a code from Example 6.

Theorem 2. *For $i \in \{2,3,4\}$, let $X_i \subseteq \mathcal{B}^i$ and set $X := X_2 \cup X_3$, $\tilde{X} := X_3 \cup X_4$.*

(1) The mixed set X is a code if and only if there exists no directed path in $\mathcal{G}(X)$ between a pair of vertices with labels in X_2.

(2) The mixed set $\tilde{X}$ is a code if and only if there exists no directed path in $\mathcal{G}(\tilde{X})$ between a pair of vertices with labels in X_3.

(3) The mixed sets X and $\tilde{X}$ are circular codes if and only if X and $\tilde{X}$ are codes and $\mathcal{G}(X)$ and $\mathcal{G}(\tilde{X})$ are acyclic.

Theorem 2 directly implies the following statement.

Corollary 1. For $i \in \{1,2\}$, let X_i be a subset of $\mathcal{B}^i$ and set $X := X_1 \cup X_2$. Let $\tilde{\mathcal{G}}(X)$ be the digraph (Bang-Jensen and Gutin, 2009) obtained from $\mathcal{G}(X)$ by identifying all vertices of $\mathcal{G}(X)$ corresponding to the dinucleotides in X_2 (keeping loops and multiple edges should they arise). The mixed set X is a circular code if and only if the digraph $\tilde{\mathcal{G}}(X)$ is acyclic. A similar statement holds for mixed subsets $\tilde{X} \subseteq \mathcal{B}^3 \cup \mathcal{B}^4$.

3.1. **Construction of maximum mixed circular codes of dinucleotides and trinucleotides.** In this section, we address the question of the existence of maximum (and hence maximal) mixed circular codes $X \subseteq \mathcal{B}^2 \cup \mathcal{B}^3$ of dinucleotides and trinucleotides. We point out that, as it has been noticed already (Fimmel *et al.*, 2017a), the notions of maximum and maximal coincide for dinucleotide circular codes. It is straightforward to see that the cardinality of a maximum mixed circular code $X \subseteq \mathcal{B}^2 \cup \mathcal{B}^3$ is less than or equal to 26, because a subset of a circular code is a circular code, the cardinality of a maximum dinucleotide circular code is 6 and the cardinality of a maximum trinucleotide circular code is 20. In fact, the following theorem shows that this simple upper bound is reached. Even more, we show that any maximal dinucleotide circular code can be embedded into a maximum mixed circular code of cardinality 26 such that the corresponding trinucleotide part is C^3-comma-free.

Theorem 3. *Let $D \subseteq \mathcal{B}^2$ be a maximal dinucleotide circular code. Then D can be embedded into a mixed code $X \subseteq \mathcal{B}^2 \cup \mathcal{B}^3$ such that*

(1) The cardinality of X is 26 (hence maximum and so maximal);
(2) X is a circular code; and
(3) $X \cap \mathcal{B}^3$ is a C^3-comma-free code.

Proof. Let $D \subseteq \mathcal{B}^2$ be a maximal dinucleotide circular code (which is thus also maximum in this case, as pointed out above). The idea of the construction of the desired mixed circular code X is to use the structure of D to construct $X_3 \subseteq \mathcal{B}^3$ by adding letters to the dinucleotides of D in the first position ensuring at the same time that $\pi_{12}(X_3) \cap \pi_{23}(X_3) = \varnothing$. This construction will then force X_3 to be comma-free and also $X = D \cup X_3$ to be a code.
As proved by Michel and Pirillo (2013) and by Fimmel *et al.* (2015, 2017a), any maximal dinucleotide circular code is of the following form:

$$D = \{N_i N_j \,:\, 1 \leq i < j \leq 4\}$$

writing $\mathcal{B} = \{N_1, N_2, N_3, N_4\}$. The code is thus, in particular, maximum. We now set

$$X_3 := \{N_k N_i N_j \,:\, N_i N_j \in D \text{ and } i \leq k \leq 4\}.$$

Explicitly,

$$X_3 = \begin{Bmatrix} N_1N_1N_2, N_1N_1N_3, N_1N_1N_4, N_2N_1N_2, N_2N_1N_3, \\ N_2N_1N_4, N_2N_2N_3, N_2N_2N_4, N_3N_1N_2, N_3N_1N_3, \\ N_3N_1N_4, N_3N_2N_3, N_3N_2N_4, N_3N_3N_4, N_4N_1N_2, \\ N_4N_1N_3, N_4N_1N_4, N_4N_2N_3, N_4N_2N_4, N_4N_3N_4 \end{Bmatrix}.$$

Consequently, $|D \cup X_3| = 26$ and $\pi_{23}(X_3) = D$ while $\pi_{12}(X_3) = \{N_i N_j \,:\, 1 \leq j \leq i \leq 4\}$, and hence the two sets $\pi_{12}(X_3)$ and $\pi_{23}(X_3)$ are disjoint. We set $X := D \cup X_3$.
The key is the following remark. Let v be a vertex of $\mathcal{G}(X)$ labelled by a dinucleotide $d = N_i N_j$. If v has a positive out-degree, then by the definition, $d \in \pi_{12}(X_3)$, that is, $i \geq j$ and therefore $d \notin D$. Similarly, if v has a positive in-degree, then $d \in \pi_{23}(X_3) = D$. Consequently, no vertex of $\mathcal{G}(X)$ labelled by a dinucleotide has both positive in-degree and positive out-degree. Several consequences readily follow. First, $\mathcal{G}(X)$ cannot contain a path starting at a vertex in D, so Theorem 2.1 implies that X is a code. Second, $\mathcal{G}(X_3)$ contains no path of length greater than 2 and no cycle of length 2. The graph $\mathcal{G}(X_3)$ is thus acyclic and X_3 is a comma-free circular code. Further, no cycle in $\mathcal{G}(X)$ can contain a vertex labelled by a dinucleotide. Since D itself is circular, and hence $\mathcal{G}(D)$ is acyclic, we infer that $\mathcal{G}(X)$ is acyclic and hence X is a circular code by Theorem 2.
It remains to show that X is a C^3-code. Let $X_3' := \{M_2 M_3 M_1 \,:\, M_1 M_2 M_3 \in X_3\}$ and $X_3^{''} := \{M_3 M_1 M_2 \,:\, M_1 M_2 M_3 \in X_3\}$ be the two circular shifts of X_3. We show that X_3' is a circular code, the argument being analogue for X_3''.
It follows from the definition of X_3 that

$$X_3' = \{N_i N_j N_k \,:\, 1 \leq i < j \leq 4 \text{ and } i \leq k \leq 4\}.$$

First, observe that a 2-cycle in $\mathcal{G}(X_3')$ would readily yield a 2-cycle in $\mathcal{G}(X)$, and hence $\mathcal{G}(X_3')$ does not contain any 2-cycle. As a result, any cycle in $\mathcal{G}(X_3')$ has length at least 4, and thus contains a directed path of the form

$$N_{i_1} \to N_{i_2} N_{i_3} \to N_{i_4} \to N_{i_5} N_{i_6} \to N_{i_7},$$

where $N_{i_j} \in \mathcal{B}$ for each $j \in \{1, \ldots, 7\}$. The form of X_3' then implies that

$$4 \geq i_6 > i_5 > i_4 \geq i_2 > i_1 \geq 1,$$

leading to $i_6 = 4$, $i_5 = 3$, $i_4 = i_2 = 2$ and $i_1 = 1$. In particular, the cycle has length more than 4, and every directed path of length 4 on it starting at a nucleotide must satisfy the above property, which yields a contradiction. Consequently, X'_3 is a circular code and we infer that X is a C^3-code. □

The next corollary is now immediate.

Corollary 2. The cardinality of a maximum mixed circular code $X \subseteq \mathcal{B}^2 \cup \mathcal{B}^3$ is 26.

In fact, we will show in Section 4 that there are exactly 32 self-complementary maximum mixed circular codes $X \subseteq \mathcal{B}^2 \cup \mathcal{B}^3$ of cardinality 26.

Remark 4. If D is a dinucleotide comma-free circular code then we can embed it into a comma-free code X but X will not have cardinality 26 as the cardinality of a maximum dinucleotide comma-free code is 5, as shown by Table 2 in Fimmel *et al.*, 2017a.

3.2. **Construction of maximum mixed circular codes of trinucleotides and tetranucleotides.** As in Theorem 3, we can also construct maximum (and hence maximal) mixed circular codes $X \subseteq \mathcal{B}^3 \cup \mathcal{B}^4$ of trinucleotides and tetranucleotides. These codes have a cardinality of 80 for the following reason. Let $X \subseteq \mathcal{B}^3 \cup \mathcal{B}^4$ be a mixed circular code. Then $X_3 := X \cap \mathcal{B}^3$ and $X_4 := X \cap \mathcal{B}^4$ are also circular codes. Therefore, X_3 has cardinality at most 20 and X_4 at most 60 as there are 256 tetranucleotides among which the 16 tetranucleotides of the form $N_1N_2N_1N_2$ for some $N_1, N_2 \in \mathcal{B}$ cannot be part of a circular code. Any other tetranucleotide has a conjugacy class of cardinality 4, so we have $60 = (256 - 16)/4$ conjugacy classes[1].

Theorem 4. *There are maximum (and hence maximal) mixed circular codes $X \subseteq \mathcal{B}^3 \cup \mathcal{B}^4$ of trinucleotides and tetranucleotides of cardinality* 80.

Proof. We shall construct a maximum mixed circular code $X \subseteq \mathcal{B}^3 \cup \mathcal{B}^4$ of cardinality 80 by taking the union of

$$X_3 := \begin{Bmatrix} AAC, AAG, AAT, CAC, CAG, CAT, CCG, CCT, GAC, GAG, \\ GAT, GCG, GCT, GGT, TAC, TAG, TAT, TCG, TCT, TGT \end{Bmatrix}$$

and

$$X_4 := \begin{Bmatrix} AAAT, AAAC, AAAG, ACAG, ACAT, ACCG, ACCT, ACGT, AGAT, AGCG, \\ AGCT, AGGT, ATCG, ATCT, ATGT, CAAC, CAAG, CAAT, CCAC, CCAG, \\ CCAT, CCCG, CCCT, CGCT, CGGT, CTGT, GAAC, GAAG, GAAT, GCAC, \\ GCAG, GCAT, GCCG, GCCT, GGAC, GGAG, GGAT, GGCG, GGCT, GGGT, \\ TAAC, TAAG, TAAT, TCAC, TCAG, TCAT, TCCG, TCCT, TGAC, TGAG, \\ TGAT, TGCG, TGCT, TGGT, TTAC, TTAG, TTAT, TTCG, TTCT, TTGT \end{Bmatrix}.$$

This construction keeps the property that $\pi_{123}(X_4) \cap \pi_{234}(X_4) = \varnothing$. Since $X_3 \subset \pi_{234}(X_4)$, there is no directed path in $\mathcal{G}(X)$ between two vertices labelled by elements of X_3. Thus, Theorem 2.2 implies that X is indeed a code. In addition, the bipartite digraph induced by $\mathcal{B} \cup \mathcal{B}^3$ is acyclic. It remains to check the component induced by $\mathcal{B}^2$, which by the construction is a transitive tournament[2], and hence acyclic too. □

[1] A *conjugacy class* of an ℓ-nucleotide $N_1 \cdots N_\ell \in \mathcal{B}^\ell$ is defined as the set $\{N_k \cdots N_\ell N_1 \cdots N_{k-1} : 1 \leq k \leq \ell\}$; clearly an ℓ-nucleotide circular code can contain at most one ℓ-nucleotide from each conjugacy class.

[2] A tournament is an orientation of a complete graph: every pair of distinct vertices is connected by a single directed edge. The reader is referred to the book by Bang-Jensen and Gutin (2009) for the notion of tournaments and to the work by Fimmel *et al.* (2016).

For the reader's convenience, we state the following remark, which was checked by computer calculations but can also be inferred by a construction similar to those in the proofs of Theorems 3 and 4.

Remark 5. Maximum mixed circular codes $X \subseteq \mathcal{B}^2 \cup \mathcal{B}^4$ of dinucleotides and tetranucleotides have cardinality 66. It is the best possible since the cardinality of a maximum dinucleotide circular code is 6 and the cardinality of a maximum tetranucleotide circular code is 60.

We finally approach the construction of mixed circular codes containing dinucleotides, trinucleotides and tetranucleotides in the next section.

3.3. **Construction of maximal mixed circular codes of (di,tri,tetra)-nucleotides.**

We present a construction of a maximal, yet not maximum, mixed circular code of (di,tri,tetra)-nucleotides that has cardinality 71. The code is not maximum as there exist mixed circular codes of (di,tri,tetra)-nucleotides of cardinality 74 that contain a maximum circular code of dinucleotides (hence of size 6) and a maximum circular code of trinucleotides (hence of size 20). There also exist larger mixed circular codes of (di,tri,tetra)-nucleotides, for instance of cardinality 81, that contain only one dinucleotide and a maximum circular code of trinucleotides. We do not know the size of a maximum mixed circular code of (di,tri,tetra)-nucleotides and this combinatorial problem remains open.

The following theorem exhibits the structure of some maximal mixed circular codes. Its proof can be found in Appendix 6.2.

Theorem 5. *Let $M_1 < M_2 < M_3 < M_4$ be an ordering of the alphabet $\mathcal{B}$. For a natural number $n > 1$, the set $X^{(n)} \subseteq \mathcal{B}^{\leq n}$ defined by*

$$X^{(n)} := \{N_m \cdots N_1 \ : \ N_i \in \mathcal{B},\ 2 \leq m \leq n \text{ and } N_m \geq N_{m-1} \geq \cdots \geq N_2 \text{ but } N_2 < N_1\}$$

is a maximal mixed circular code.

Theorem 5 allows us to construct the following maximal mixed circular code.

Corollary 3. There are maximal mixed circular codes in $\mathcal{B}^2 \cup \mathcal{B}^3 \cup \mathcal{B}^4$ of dinucleotides, trinucleotides and tetranucleotides of cardinality 71.

Proof. A maximal mixed circular code $X \subseteq \mathcal{B}^2 \cup \mathcal{B}^3 \cup \mathcal{B}^4$ is constructed by removing from the maximal mixed circular code $X' \subseteq \mathcal{B}^3 \cup \mathcal{B}^4$ described in the proof of Theorem 4 all tetranucleotides that are built from combining two dinucleotides from the corresponding maximal dinucleotide circular code D. We now construct a maximal mixed circular code $X \subseteq \mathcal{B}^2 \cup \mathcal{B}^3 \cup \mathcal{B}^4$ of cardinality 71: $X := X_2 \cup X_3 \cup X_4$ where

$$X_2 := \Big\{AC, AG, AT, CG, CT, GT\Big\}$$

$$X_3 := \begin{Bmatrix} AAC, AAG, AAT, CAC, CAG, CAT, CCG, CCT, GAC, GAG, \\ GAT, GCG, GCT, GGT, TAC, TAG, TAT, TCG, TCT, TGT \end{Bmatrix}$$

and

$$X_4 := \begin{Bmatrix} AAAC, AAAG, AAAT, CAAC, CAAG, CAAT, CCAC, CCAG, CCAT, \\ CCCG, CCCT, GAAC, GAAG, GAAT, GCAC, GCAG, GCAT, GCCG, \\ GCCT, GGAC, GGAG, GGAT, GGCG, GGCT, GGGT, TAAC, TAAG, \\ TAAT, TCAC, TCAG, TCAT, TCCG, TCCT, TGAC, TGAG, TGAT, \\ TGCG, TGCT, TGGT, TTAC, TTAG, TTAT, TTCG, TTCT, TTGT \end{Bmatrix}.$$

From Theorem 5, we deduce that $X = X_2 \cup X_3 \cup X_4$ is indeed a mixed circular code using the order $A < C < G < T$. Moreover, it is maximal since any combination of dinucleotides in X_2 has to be removed from X_4, so a total of $\binom{6}{2} = 15$ tetranucleotides. Thus $71 = 6 + 20 + 60 - 15$ is the maximal cardinality for X. □

Theorem 5 also readily yields the following corollary.

Corollary 4. Let $M_1 < M_2 < M_3 < M_4$ be any ordering of the alphabet $\mathcal{B}$. Then the following set $X \subseteq \mathcal{B}^*$ where

$$X := \{N_m N_{m-1} \cdots N_2 N_1 \,:\, N_i \in \mathcal{B},\ m \in \mathbf{N} \textit{ and } N_m \geq N_{m-1} \geq \cdots \geq N_2 \textit{ but } N_2 < N_1\}$$

is an infinite mixed circular code.

After having constructed several examples of mixed circular codes of (di,tri,tetra)-nucleotides, we concentrate on the biologically important classes of self-complementary circular (and comma-free) codes of dinucleotides and trinucleotides in the next sections.

4. Self-complementary mixed circular codes of dinucleotides and trinucleotides

This section is devoted to the class of self-complementary mixed circular codes, which are of importance since they can be found on both strands of the double helix of DNA simultaneously. In particular, we completely determine the maximum self-complementary mixed circular codes over $\mathcal{B}^2 \cup \mathcal{B}^3$ that contain the maximal C^3-self-complementary circular code (1) found in nature.

4.1. Growth function of self-complementary mixed circular codes of dinucleotides and trinucleotides.

The following statement has been established by computer calculus.

Proposition 1. The growth function of self-complementary mixed circular codes of dinucleotides and trinucleotides varies from cardinality 3 to 26. Its maximum is reached with 121,792 self-complementary mixed circular codes of cardinality 15 (see Table 1).

Table 1 displays the number of self-complementary mixed circular codes $X \subseteq \mathcal{B}^2 \cup \mathcal{B}^3$ as a function of the length of a longest directed path in their associated acyclic graphs. These lengths vary from 1 to 8 by Theorem 4.2 in Fimmel *et al.* (2018) and by Theorem 2 which naturally extends to mixed circular codes. A mixed circular code whose maximal directed path length is less than or equal to 2, 1 respectively, is comma-free, strong comma-free respectively (Fimmel *et al.* 2017b). Thus there are exactly 164 self-complementary mixed strong comma-free codes and $11,788 + 164 = 11,952$ self-complementary mixed comma-free codes over $\mathcal{B}^2 \cup \mathcal{B}^3$. We will come back later to the classes of self-complementary (strong) comma-free codes.
Before we move on to the self-complementary comma-free codes, we state two propositions proving the entries of Table 1 for cardinalities 3 and 4.

Proposition 2. The number of self-complementary mixed circular codes X of dinucleotides and trinucleotides of cardinality 3 is equal to 100.

Proof. Clearly, such a code X can only be the union of a self-complementary dinucleotide $N_1N_2 = \overleftarrow{c}(N_1N_2)$ and a self-complementary trinucleotide circular code of cardinality 2 since self-complementary trinucleotides do not exist. There are 4 self-complementary dinucleotides AT, TA, CG and GC and 28 self-complementary trinucleotide circular codes of cardinality 2 since from the 32 trinucleotide-anti-trinucleotide pairs, one should exclude the two trivial pairs (AAA, TTT) and (CCC, GGG) as well as the two obvious pairs (ATA, TAT) and (CGC, GCG). Consequently,

Table 1. Growth function of self-complementary mixed circular codes $X \subseteq \mathcal{B}^2 \cup \mathcal{B}^3$ (cardinality between 3 and 26) as a function of the maximal path length ℓ (from 1 to 8) in their associated graph $\mathcal{G}(X)$.

ℓ $\lvert X\rvert$	1	2	3	4	5	6	7	8	**Total**
3	28	36	36	0	0	0	0	0	100
4	32	80	64	16	0	8	0	0	200
5	48	392	516	324	80	16	16	0	1392
6	32	560	760	564	88	216	16	0	2236
7	16	1280	2088	3940	760	548	288	112	9032
8	8	1248	2608	4772	952	1960	288	104	11940
9	0	2020	3812	17484	2848	4612	1568	1744	34088
10	0	1400	3880	16572	3640	8724	1568	1560	37344
11	0	1788	3640	38812	5696	16988	3728	9424	80076
12	0	956	3056	30720	6456	21828	3728	7704	74448
13	0	940	1948	49904	6456	33732	4472	24192	121644
14	0	472	1376	35172	6168	33104	4472	18032	98796
15	0	316	560	40396	4208	39728	2872	33712	121792
16	0	176	328	26520	3408	32176	2872	23168	88648
17	0	72	64	21356	1592	29112	968	27120	80284
18	0	40	32	13360	1088	20504	968	17360	53352
19	0	8	0	7404	336	13404	152	12672	33976
20	0	4	0	4432	192	8456	152	7600	20836
21	0	0	0	1648	32	3736	8	3264	8688
22	0	0	0	936	16	2144	8	1832	4936
23	0	0	0	224	0	560	0	400	1184
24	0	0	0	120	0	296	0	208	624
25	0	0	0	16	0	32	0	16	64
26	0	0	0	8	0	16	0	8	32
Total	164	11788	24768	314700	44016	271900	28144	190232	885712

there are at most $4 \cdot 28 = 112$ self-complementary circular mixed codes of cardinality 3. Each such code X has the form $\{N_1N_2N_3, c(N_3)c(N_2)c(N_1), Nc(N)\}$ where $N_1, N_2, N_3, N \in \mathcal{B}$ such that N_1, N_2 and N_3 are not all equal and if $N_1 = N_3$ then $N_2 \neq c(N_1)$.

We assert that X is not circular if and only if $X = \{N'c(N')N, c(N)N'c(N'), Nc(N)\}$ where $N \in B$ and $N' \in \mathcal{B} \setminus \{N\}$. This will then imply that exactly $4 \cdot 3 = 12$ additional self-complementary mixed circular codes of cardinality 3 have to be removed, and we will hence obtain $112 - 4 \cdot 3 = 100$ mixed self-complementary circular codes of cardinality 3 overall. It remains to prove the equivalence, one direction being trivial. For the converse, if X is not circular, then $\mathcal{G}(X)$ must have a cycle that contains the edge $N \to c(N)$. Necessarily, the vertex preceding N on the cycle is labelled by a dinucleotide d, and has both positive in-degree and positive out-degree. The definition of X thus yields that either $d = N_1N_2 = c(N_2)c(N_1)$, or $d = N_2N_3 = c(N_3)c(N_2)$; in particular, d is of the form $N'c(N')$. This further implies that $N_3 = N$ or $N_1 = c(N)$, respectively. Consequently, we conclude that $X = \{N'c(N')N, c(N)N'c(N'), Nc(N)\}$, which finishes the proof. □

Proposition 3. The number of self-complementary mixed circular codes in $\mathcal{B}^2 \cup \mathcal{B}^3$ of cardinality 4 is 200.

Proof. Analogously to the proof of Proposition 2, such a code X can only be the union of a self-complementary dinucleotide circular code of cardinality 2 and a self-complementary trinucleotide circular code of cardinality 2. Therefore

$$X = \{N_1N_2N_3, c(N_3)c(N_2)c(N_1), M_1M_2, M_3M_4\},$$

where $N_1, N_2, N_3, M_1, M_2, M_3, M_4 \in \mathcal{B}$. Moreover, there are 28 self-complementary trinucleotide circular codes of cardinality 2 (see the proof of Proposition 2) and 8 self-complementary dinucleotide circular codes of cardinality 2, namely

$$\{AT, CG\}, \{AT, GC\}, \{TA, GC\}, \{TA, CG\}, \{AC, GT\}, \{AG, CT\}, \{TC, GA\}, \{TG, CA\}.$$

This means that there are at most $8 \cdot 28 = 224$ mixed self-complementary circular codes of cardinality 4.

Now observe that X is not circular if and only if $\mathcal{G}(X)$ has a cycle containing a path of the form $d_1 \to L_1 \to L_2 \to d_2$ where $d_1, d_2 \in \mathcal{B}^2$ are dinucleotides and $L_1L_2 \in \{M_1M_2, M_3M_4\}$. As in the proof of Proposition 2, since every vertex on a cycle has both positive in-degree and positive out-degree, it follows that $d_1 = Nc(N)$ and $d_2 = N'c(N')$ with $N, N' \in \mathcal{B}$. Moreover, since $Nc(N)L_1 \in X$, we infer that $L_2N'c(N')$ must be equal to $c(L_1)Nc(N)$ and hence $L_2 = c(L_1)$ and $N = N'$. Therefore, if X is not circular then $\{M_1M_2, M_3M_4\}$ is one of the 4 codes consisting of self-complementary dinucleotides, and X has the form

$$\{Nc(N)L_1, c(L_1)Nc(N), L_1c(L_1), L_2c(L_2)\},$$

with $L_1 \in B$, $L_2 \in \mathcal{B} \setminus \{L_1, c(L_1)\}$ and $N \in \mathcal{B}$. Because the codes containing a trinucleotide of the form $Mc(M)M$ for $M \in \mathcal{B}$ have already been excluded, we further deduce that $N \neq L_1$, and we thus have to exclude exactly $4 \cdot 3 \cdot 2 = 24$ additional codes. We obtain $224 - 24 = 200$ self-complementary mixed circular codes of cardinality 4 overall. □

We finally show that there are exactly 32 maximum self-complementary mixed circular codes over $\mathcal{B}^2 \cup \mathcal{B}^3$ that have a maximal cardinality of 26. However, none of them contains the maximal C^3-self-complementary circular code observed in genes (see Proposition 5). The proof of the following result is contained in Appendix 6.3 and the list appears in Appendix 6.5.

Proposition 4. The cardinality of a maximum self-complementary mixed circular code over $\mathcal{B}^2 \cup \mathcal{B}^3$ is 26 and there are 32 of them.

We now consider the maximal C^3-self-complementary trinucleotide circular code observed in genes. From an evolutionary point of view, it seems relevant to point out how dinucleotides can be added to this particular code without compromising its properties.

Proposition 5. Let Y be a self-complementary mixed circular code over $\mathcal{B}^2 \cup \mathcal{B}^3$ that contains the maximal C^3-self-complementary trinucleotide circular code X of cardinality 20 observed in genes, defined by (1). Then $Y \cap \mathcal{B}^2 \in \{\{AT\}, \{GC\}, \{AT, GC\}\}$. In other words,

$$\begin{aligned} Y \subseteq \{&AT, GC, AAC, GTT, GAA, TTC, AAT, ATT, ACC, GGT, GAC, GTC,\\ &CAG, CTG, GTA, TAC, ATC, GAT, GCC, GGC, CTC, GAG\}. \end{aligned}$$

Proof. Let Y be as stated in the proposition. We systematically exclude all dinucleotides except for AT and GC. Let us start with the dinucleotides that are not self-complementary, ignoring the 4 trivial ones (NN for $N \in \mathcal{B}$). If one of them is contained in Y, then the complementary dinucleotide must be in Y as well. There are 4 pairs of this kind: $\{GA, TC\}$, $\{AC, GT\}$, $\{AG, CT\}$ and $\{CA, TG\}$. We first handle the first two cases. Because both GAT and TTC belong to X, the digraph $\mathcal{G}(X)$ contains a directed path from GA to TC, and hence the pair $\{GA, TC\}$ cannot belong to Y. Similarly, the fact that GTT and TAC belong to X forbids the pair $\{AC, GT\}$. For the latter two cases, if $\{AG, CT\} \subseteq Y$, then as GTA and GGT also belong to Y the digraph $\mathcal{G}(Y)$ would contain a directed cycle, namely

$$GT \to A \to G \to GT.$$

Similarly, the pair $\{CA, TG\}$ cannot belong to Y as both TTC and ATT do.
Finally, a given self-complementary dinucleotide $Nc(N)$ cannot be combined with any code that contains the 2 complementary trinucleotides $N'c(N')N, c(N)N'c(N')$ for some $N' \in \mathcal{B}$, since the obtained code would not be circular. Indeed, the associated digraph would contain the cycle

$$N'c(N') \to N \to c(N) \to N'c(N').$$

It now suffices to remark that X contains the complementary pairs (ATT, AAT) and (TAC, GTA), which thus forbids the self-complementary dinucleotides TA and CG.
We finally note that $Y \cap \mathcal{B}^2$ can indeed be equal to $\{AT, GC\}$, as the obtained code is then circular (and self-complementary). The digraph $\mathcal{G}(Y)$ associated with Y in this case is given in Figure 7. □

To close this section, we finally consider the parameters for self-complementary mixed (strong) comma-free codes determined in Table 1 and prove their correctness. We also determine explicitly the 4 maximum (hence maximal) self-complementary mixed comma-free codes of dinucleotides and trinucleotides. The following result was obtained by an extensive computer calculation — see Table 1.

Proposition 6. The following statements hold.

(1) The growth function of self-complementary mixed strong comma-free codes of dinucleotides and trinucleotides varies from cardinality 3 to 8.
(2) The growth function of self-complementary mixed comma-free codes of dinucleotides and trinucleotides varies from cardinality 3 to 20.

As it turns out, there are only a few maximum self-complementary mixed comma-free codes over $\mathcal{B}^2 \cup \mathcal{B}^3$. See Appendix 6.4 for a proof.

Proposition 7. The cardinality of a maximum self-complementary mixed comma-free code of dinucleotides and trinucleotides is 20 and there are exactly 4 of them:

- $\{AC, GT, AT, GC, AAC, GTT, AAT, ATT, ACC, GGT, GAC, GTC, ACT, AGT, AGC, GCT,$ $ATC, GAT, GCC, GGC\}$;

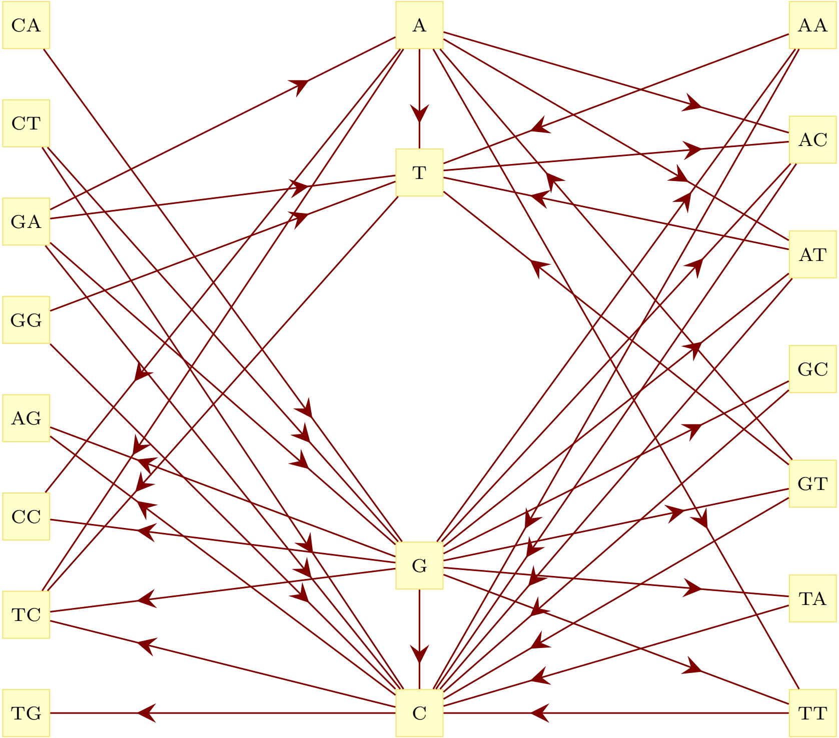

Figure 7. The graph $\mathcal{G}(Y)$ of the maximum self-complementary mixed circular code of dinucleotides and the maximal C^3-self-complementary trinucleotide circular code X of cardinality 20 observed in genes: $Y = \{AT, GC, AAC, GTT, GAA, TTC, AAT, ATT, ACC, GGT, GAC, GTC, CAG, CTG, GTA, TAC, ATC, GAT, GCC, GGC, CTC, GAG\}$. The vertices labelled by a nucleotide A, C, G, T have both ingoing and outgoing edges, the edges between vertices labelled by nucleotides being associated with the dinucleotides AT and GC. The vertices labelled by one of the dinucleotides AG, CC, TC, TG have no outgoing edge, while those labelled by one of the dinucleotides CA, CT, GA, GG have no ingoing edge. The vertices labelled by one of the 7 remaining dinucleotides $AA, AC, AT, GC, GT, TA, TT$ all have both ingoing and outgoing edges.

- $\{CA, TG, TA, CG, CAA, TTG, TAA, TTA, CCA, TGG, CGA, TCG, CTA, TAG, CAG, CTG, TCA, TGA, CCG, CGG\}$;
- $\{AG, CT, AT, CG, AAG, CTT, AAT, ATT, ACG, CGT, ACT, AGT, CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG\}$;
- $\{GA, TC, TA, GC, GAA, TTC, TAA, TTA, GAC, GTC, GCA, TGC, GGA, TCC, GTA, TAC, TCA, TGA, GCC, GGC\}$.

5. Discussion and Conclusion

In this section, some possible biological implications of the theory of mixed codes are discussed. Based on the new insights and hypotheses regarding the genesis of the genetic code, we suggest some potential biological functions of mixed circular codes in primitive genetic processes. As already mentioned in the introduction, various scientists assume that the coding of amino acids was first done by dinucleotides or tetranucleotides, and only later by trinucleotides. Therefore it is plausible to assume that the mixed codes could have had a function in a transitional coding process.

Self-complementary mixed circular codes, in particular of dinucleotides and trinucleotides, could have operated in the primitive soup for constructing the modern genetic code and the genes. They could be involved in two stages: a first stage directly without anticodon-amino acid interactions to form peptides from prebiotically amino acids, and a second stage using these interactions (Johnson and Wang, 2010). The absence of a code, as proposed by Noller (2004) and Krupkin *et al.* (2011), could be explained, from our point of view, by the lack of knowledge in the variety and the complexity of codes. We suggest some potential functional implications of mixed circular codes in these two stages.

A mixing of dinucleotides and trinucleotides with its property of reading frame retrieval could have been involved in the Implicated Site Nucleotides (ISN) of RNA interacting with the amino acids at the primitive step of life (review in Yarus, 2017). According to a great number of biological experiments, the ISN structure contains nucleotides in fixed and variable positions, as well as an important trinucleotide for interacting with the amino acid. The general structure of the aptamers binding amino acids, in particular its nucleotide length, its amino acid binding loop and its nucleotide position, is still an open problem. Likewise, aptamers without enrichment for Gln, Leu and Val are not explained so far. Similar arguments could hold for the ribonucleopeptides which could be involved in a primitive T box riboswitch functioning as an aminoacyl-tRNA synthetase and a peptidyl-transferase ribozyme (Saad, 2018). Circularity could have been necessary for determining the position of nucleotide motifs in the primitive construction of the genetic code (Kun and Radványi, 2018).

For the second stage, many schemes have suggested that a simpler code based on dinucleotides preceded the modern trinucleotide genetic code (see, for instance, Baranov *et al.*, 2009; Patel, 2005; Wu *et al.*, 2005), the two first codon sites being associated with the dinucleotide code. One scheme is the GC code: GG coding for Gly, CC coding for Pro, GC coding for Ala and CG coding for Arg (Hartman, 1995). A second dinucleotide code coding 14 amino acids, the 20 amino acids except the six structurally and synthetically complex amino acids His, Lys, Met, Phe, Trp and Tyr, was proposed based on a strong correlation between the first codon sites and the biosynthetic pathways of the amino acids they encode, as well as a strong relationship between the second codon sites and the hydrophobicity of the amino acids (Copley *et al.*, 2005). Such approaches have been much criticized and abandoned. The central argument is that the reading frame with dinucleotides is incompatible with the reading frame with trinucleotides, leading to gene sequences unreadable, problem synthesized shortly with the following sentence "letters belonging to the first position of the next codon would be consistently misread as being the last letter of the preceding codon" (Frank and Froese, 2018). We prove in this paper the existence of mixed circular codes, in particular of dinucleotides and trinucleotides. Thus, there are sequences which can be constructed with dinucleotides and trinucleotides such that their frame can be read unambiguously. In other words, primitive genetic codes mixing dinucleotides and trinucleotides are possible and could be a transition code between an earlier dinucleotide code and the modern trinucleotide code.

Figure 8 proposes an evolutionary hypothesis of self-complementary mixed circular codes SMC_ℓ of dinucleotides and trinucleotides according to a hierarchy related to their combinatorial complexity with the maximal path length ℓ (from 1 to 8) in their associated graph $\mathcal{G}$ (see Table 1). As the maximal path length ℓ is related to the window nucleotide length of reading frame retrieval, the self-complementary mixed circular codes SMC_1 are more constraint than those SMC_8. The maximal C^3-self-complementary trinucleotide circular code X observed in genes (1) could have come from a self-complementary mixed circular code SMC_8 of dinucleotides and trinucleotides with the deletion of the two dinucleotides AT and GC.

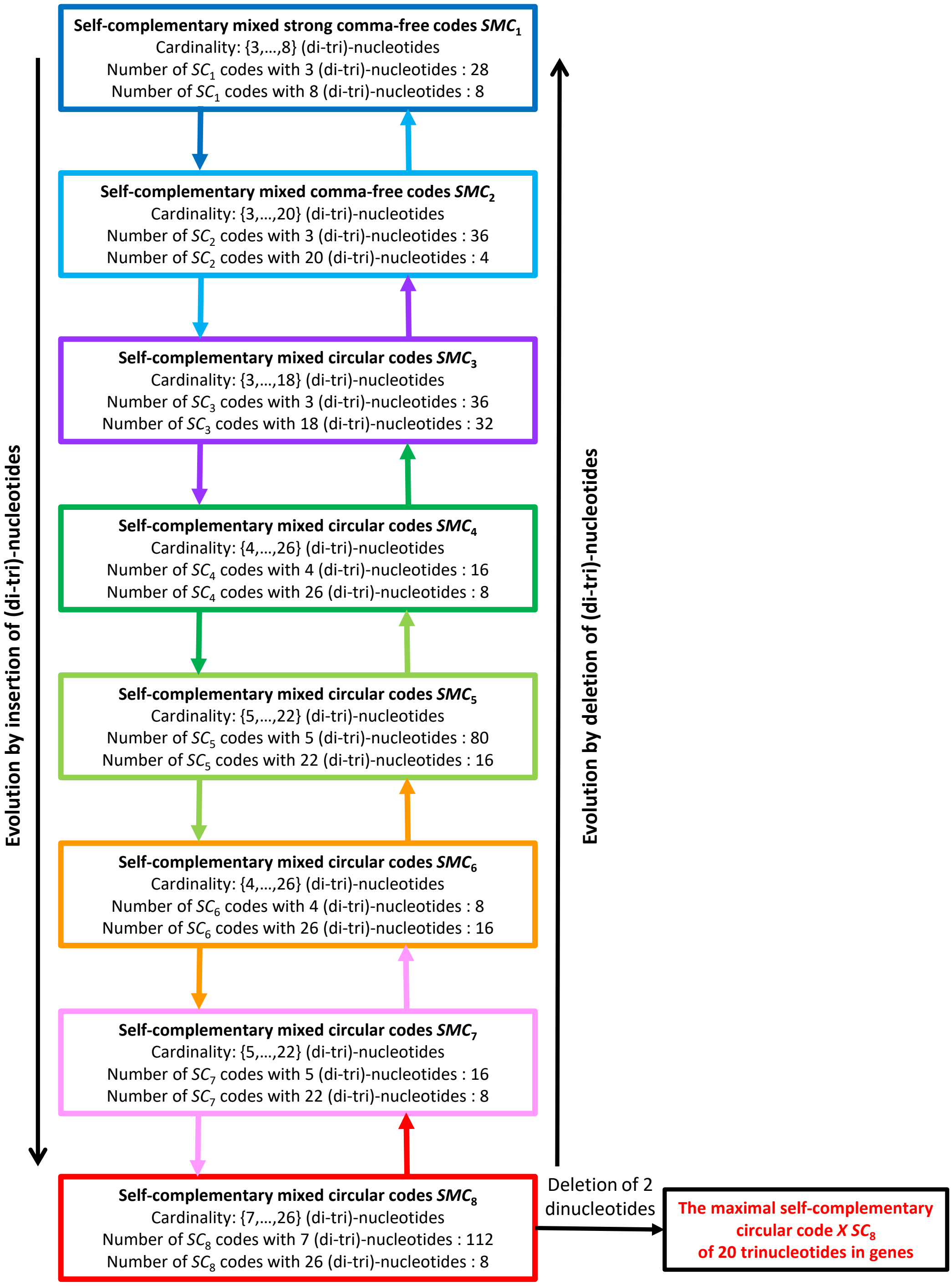
Self-complementary mixed strong comma-free codes *SMC*₁
Cardinality: {3,…,8} (di-tri)-nucleotides
Number of *SC*₁ codes with 3 (di-tri)-nucleotides : 28
Number of *SC*₁ codes with 8 (di-tri)-nucleotides : 8
Self-complementary mixed comma-free codes *SMC*₂
Cardinality: {3,…,20} (di-tri)-nucleotides
Number of *SC*₂ codes with 3 (di-tri)-nucleotides : 36
Number of *SC*₂ codes with 20 (di-tri)-nucleotides : 4
Self-complementary mixed circular codes *SMC*₃
Cardinality: {3,…,18} (di-tri)-nucleotides
Number of *SC*₃ codes with 3 (di-tri)-nucleotides : 36
Number of *SC*₃ codes with 18 (di-tri)-nucleotides : 32
Self-complementary mixed circular codes *SMC*₄
Cardinality: {4,…,26} (di-tri)-nucleotides
Number of *SC*₄ codes with 4 (di-tri)-nucleotides : 16
Number of *SC*₄ codes with 26 (di-tri)-nucleotides : 8
Self-complementary mixed circular codes *SMC*₅
Cardinality: {5,…,22} (di-tri)-nucleotides
Number of *SC*₅ codes with 5 (di-tri)-nucleotides : 80
Number of *SC*₅ codes with 22 (di-tri)-nucleotides : 16
Self-complementary mixed circular codes *SMC*₆
Cardinality: {4,…,26} (di-tri)-nucleotides
Number of *SC*₆ codes with 4 (di-tri)-nucleotides : 8
Number of *SC*₆ codes with 26 (di-tri)-nucleotides : 16
Self-complementary mixed circular codes *SMC*₇
Cardinality: {5,…,22} (di-tri)-nucleotides
Number of *SC*₇ codes with 5 (di-tri)-nucleotides : 16
Number of *SC*₇ codes with 22 (di-tri)-nucleotides : 8
Self-complementary mixed circular codes *SMC*₈
Cardinality: {7,…,26} (di-tri)-nucleotides
Number of *SC*₈ codes with 7 (di-tri)-nucleotides : 112
Number of *SC*₈ codes with 26 (di-tri)-nucleotides : 8
Evolution by insertion of (di-tri)-nucleotides
Evolution by deletion of (di-tri)-nucleotides
Deletion of 2 dinucleotides
The maximal self-complementary circular code *X SC*₈ of 20 trinucleotides in genes

FIGURE 8. An evolutionary hypothesis of self-complementary mixed circular codes of dinucleotides and trinucleotides leading to the maximal C^3-self-complementary trinucleotide circular code X of cardinality 20 observed in genes (1). Here SMC_ℓ means a self-complementary mixed circular code of maximal path length ℓ (see Table 1), while SC_8 means a self-complementary circular code of maximal path length 8.

Proposition 5 proves that the maximal C^3-self-complementary trinucleotide circular code X observed in genes can be mixed with self-complementary dinucleotide circular codes. It can be mixed with at most the two dinucleotides $\{AT, GC\}$, each dinucleotide being self-complementary and beginning with a purine nucleotide. The other two self-complementary dinucleotides TA and CG are excluded from such a mixing with X. We have no biological explanation of this new combinatorial property of X so far.

After having initiated the circular code theory in genes in 1996, we have opened here a new line of mathematical research on circular codes. We have proved that mixed codes of dinucleotides, trinucleotides and tetranucleotides can be circular. Furthermore, we have identified several new combinatorial properties with the self-complementary mixed circular codes of dinucleotides and trinucleotides. These results have been related to some potential biological functions of mixed circular codes in primitive genetic processes. We are currently extending this research work to different finite alphabets.

6. Appendix

6.1. **Proof of Theorem 2.**

Proof. First note that for every edge $e = ww' \in E(X)$, the concatenation ww' of the labels of the end-vertices of e yields an element of X, by the definition of $\mathcal{G}(X)$.
(*1*) Assume that there exists a directed path $w_1, \dots, w_r$ in $\mathcal{G}(X)$ such that both w_1 and w_r belong to X_2. Then the word $w := w_1 \cdots w_r$ has an ambiguous decomposition. Indeed, if r is odd then both $w_1|w_2w_3|\dots|w_{2i}w_{2i+1}|\dots|w_{r-1}w_r$ and $w_1w_2|\dots|w_{2i-1}w_{2i}|\dots|w_{r-2}w_{r-1}|w_r$ are decompositions of w into elements of X. If r is even then both $w_1|w_2w_3|\dots|w_{2i}w_{2i+1}|\dots|w_{r-2}w_{r-1}|w_r$ and $w_1w_2|\dots|w_{2i-1}w_{2i}|\dots|w_{r-1}w_r$ are decompositions of w into elements of X.
Conversely, assume that the mixed set X is not a code. So there are elements of X^* admitting several decompositions into elements of X, and we can choose a minimal one w (meaning that no substring of w has more than one decomposition into elements of X). Let $w_1|\dots|w_r \in X^r$ and $w'_1|\dots|w'_{r'} \in X^{r'}$ be two such decompositions of w. Since w is minimal, $w_1 \neq w'_1$ and $w_r \neq w'_{r'}$, and hence we may assume without loss of generality that $w_1 \in X_2$ and $w'_1 \in X_3$. Similarly, either $w_r \in X_2$ and $w'_{r'} \in X_3$, or $w_r \in X_3$ and $w'_{r'} \in X_2$. Both cases being similar to analyze, let us suppose that the latter one occurs. We show that w_i overlaps w'_i for each $i \in \{1, \dots, r\}$, and that w'_i overlaps w_{i+1} for each $i \in \{1, \dots, r-1\}$. We set $x_{2i-1} := w_i \sqcap w'_i$ and $x_{2i} := w'_i \sqcap w_{i+1}$, and we also show that there is a path in $\mathcal{G}(X)$ from $w_1 = x_1$ to x_{2i-1}. We proceed by a (finite) induction on the index i.

(i) If $i = 1$ then $w'_1 = w_1N$ for some $N \in \mathcal{B}$, so w_1 overlaps w'_1. Further, $w'_1 \sqcap w_2 = N$, and hence w'_1 overlaps w_2. By definition, $w_1 = x_1$ and so $\mathcal{G}(X)$ contains a (trivial) path from w_1 to x_1.

(ii) Fix $i \in \{2, \dots, r\}$ and assume, by induction, that w_{i-1} overlaps w'_{i-1}, that w'_{i-1} overlaps w_i, and also that $\mathcal{G}(X)$ contains a path from x_1 to x_{2i-3}. If w_i does not overlap w'_i, then w_i is contained in w'_{i-1}. Since each of them has length either 2 or 3, and moreover w'_{i-1} overlaps w_i, it follows that w_i is a suffix of w'_{i-1}. Consequently, the word $w'_1 \cdots w'_{i-1}$ admits two decompositions into elements of X, which is forbidden by the minimality of w. As a result, w_i overlaps w'_i, and for the same reason if $i < r$ then w'_i overlaps w_{i+1}. It follows that $w'_{i-1} = x_{2i-3}x_{2i-2}$ and $w_i = x_{2i-2}x_{2i-1}$. Therefore, $\mathcal{G}(X)$

contains an edge from x_{2i-3} to x_{2i-2} and an edge from x_{2i-2} to x_{2i-1}, which can be used to extend the path from x_1 to x_{2i-3} into a path from x_1 to x_{2i-1}.

Applying the above with $i = r$, we deduce that $\mathcal{G}(X)$ contains a path from w_1 to $x_{2r-1} = w'_{r'}$, which are two dinucleotides contained in X_2.
(2) The proof of the statement of the theorem for $\tilde{X}$ is analogous to *(1)*.
(3) Assume that $\mathcal{G}(X)$ contains a directed cycle $w_1, \ldots, w_r$. If r is odd then the cyclic word $w_1 \cdots w_r w_1 \cdots w_r$ has two circular decompositions into words on X^*, namely

$$w_1w_2|\ldots|w_rw_1|\ldots|w_{r-1}w_r| \quad \text{and} \quad w_1|\ldots|w_{r-1}w_r|w_1w_2|\ldots|w_{r-2}w_{r-1}|w_r.$$

The case where r is even is similar, the word $w_1 \cdots w_r$ admitting the two circular decompositions $w_1w_2|\ldots|w_{r-1}w_r|$ and $w_1|w_2w_3|\ldots|w_{r-2}w_{r-1}|w_r$.
Conversely, assume that X is a code that contains a word with two circular decompositions $w_1|\ldots|w_r$ and $w'_1|\ldots|w'_{r'}$ into elements of X. Assume without loss of generality, that $|w_1| \leq |w'_1|$. Then we may assume that for every index i, the word w_i overlaps w'_i and w'_i overlaps w_{i+1}. Indeed, otherwise one would be either the suffix or the prefix of the other, which would give a word that has two (non-circular) decompositions into elements of X, thereby contradicting the fact that X is a code. Setting $x_{2i-1} := w_i \sqcap w'_i$ and $x_{2i} := w'_i \sqcap w_{i+1}$, it follows that $x_1 \to \cdots \to x_{2r} \to x_1$ is a directed cycle in $\mathcal{G}(X)$, which ends the proof. □

6.2. **Proof of Theorem 5.**

Proof. The following property, coined Property O for future references, directly follows from the definition of $X^{(n)}$.

(O): Let w and w' be two words in $X^{(n)}$. If w overlaps w', then either w' is a suffix of w, or $w \sqcap w' = N \in \mathcal{B}$.

In particular, Property O implies that a word in $X^{(n)}$ cannot be a prefix of another word in $X^{(n)}$. This readily implies that $X^{(n)}$ is code. Indeed, suppose on the contrary that there is a concatenation of words $w = w_1 \cdots w_k$ from $X^{(n)}$ that has a second decomposition $w = w'_1 \cdots w'_l$ with $w'_i \in X^{(n)}$. Up to considering such a word w of minimal length, we may assume that $w_1 \neq w'_1$ and $|w_1| < |w'_1|$. However, this implies that w_1 is a prefix of w'_1, a contradiction. Thus $X^{(n)}$ is a code.
Property O also allows us to show that $X^{(n)}$ is a circular code. Suppose, on the contrary, that w is a word admitting two different circular decompositions $w_1 \cdots w_\ell$ and $w'_1 \cdots w'_k$ into words in $X^{(n)}$, with $\ell \geq 2$. The fact that $X^{(n)}$ is a code implies that if w_i overlaps w'_j, then w'_j cannot be a suffix of w_i. Consequently, Property O implies that $w_i \sqcap w'_j$ is a one-letter word. In particular, w_1 overlaps w'_1, which overlaps w_2, from which it follows that w'_1 has length 2. We therefore infer that $|w'_i| = 2$ for each $i \in \{1, \ldots, k\}$ and $|w_i| = 2$ for each $i \in \{1, \ldots, \ell\}$. This implies that w is a strictly increasing sequence of nucleotides, a contradiction.
It remains to show that $X^{(n)}$ is a maximal circular code. We proceed by induction on the integer $n \geq 2$, the statement following from Theorem 3 when $n \in \{2, 3\}$. Assume that $X^{(k)}$ is maximal for all $k \leq n$ and set $X := X^{(n+1)} \subseteq \mathcal{B}^{\leq n+1}$. Assume furthermore that X is not maximal, *i.e.* there is a word $w \in \mathcal{B}^{\leq n+1}$ such that $X \cup \{w\}$ is a circular code. Since $X \cap \mathcal{B}^{\leq n} = X^{(n)}$ is a maximal mixed circular code by the induction hypothesis, $w = L_{n+1} \cdots L_1 \in \mathcal{B}^{n+1}$. The construction of X and the fact that $w \notin X$ imply that w can be decomposed into words from $X \cap \mathcal{B}^{\leq n}$ and a word of the form $v = N_\ell \cdots N_1$ where $N_\ell \geq \cdots \geq N_1$ and $\ell \geq 0$, with $v = \varepsilon$ if $\ell = 0$.

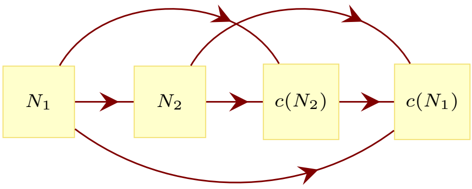

Figure 9. The directed graph associated with each of the 8 maximum self-complementary dinucleotide circular codes in the proof of Proposition 4. For instance, for the first listed code, one has $N_1 = A$ and $N_2 = C$.

Let $w = w_1 \cdots w_t v$ be such a decomposition, where t can be zero — implying that $w = v$. If $v = \varepsilon$, then w and $w_1 \cdots w_t$ are two decompositions of w into words from $X \cup \{w\}$, which contradicts that $X \cup \{w\}$ is a code. We deduce that $v \neq \varepsilon$. We now distinguish two cases.

(i) If $t = 0$, then $w = v = N_{n+1} \cdots N_1$ with $N_{n+1} \geq N_n \geq \cdots \geq N_1$. Clearly, we cannot have $N_{n+1} = \cdots = N_1$ as this would contradict the circularity of $X \cup \{w\}$. Therefore, $N_1 < N_{n+1}$. As a result, $N_n \cdots N_1 N_{n+1} \in X$, which contradicts the circularity of $X \cup \{w\}$.

(ii) If $t \neq 0$, then $v = N_\ell \cdots N_1$ with $\ell \leq n - 1$. Consequently the word wM_1M_2 has two different decompositions over $X \cup \{w\}$, namely $wM_1M_2 = w_1 \mid \ldots \mid w_t \mid vM_1M_2$ since $M_1M_2 \in X$ and vM_1M_2 has length at most $n + 1$ and hence belongs to X as well, a contradiction to the assumption that $X \cup \{w\}$ is a code.

□

6.3. Proof of Proposition 4.

Proof. A circular code over $\mathcal{B}^\ell$ intersects each conjugacy class of length ℓ in at most one ℓ-nucleotide. We proceed by finding all self-complementary mixed circular codes contained in $\mathcal{B}^2 \cup \mathcal{B}^3$ that intersect each such conjugacy class in exactly one element, hence having cardinality $20 + 6 = 26$. There are exactly 32 of them.

Let X be such a self-complementary mixed circular code over $\mathcal{B}^2 \cup \mathcal{B}^3$. It follows that $X_2 := X \cap \mathcal{B}^2$ is one of the 8 maximum (of cardinality 6) self-complementary dinucleotide circular codes:

$$\{AC, GT, AG, CT, AT, CG\}, \{AC, GT, AG, CT, AT, GC\}, \{AC, GT, GA, TC, AT, GC\},$$
$$\{AC, GT, GA, TC, TA, GC\}, \{CA, TG, AG, CT, AT, CG\}, \{CA, TG, AG, CT, TA, CG\},$$
$$\{CA, TG, GA, TC, TA, CG\}, \{CA, TG, GA, TC, TA, GC\}.$$

We note that the directed graphs associated with all these 8 codes are isomorphic to the directed graph depicted in Figure 9. In particular, X_2 yields a total order $<$ on $\mathcal{B}$ such that $X < Y$ implies that $c(Y) < c(X)$. Let us write $\mathcal{B} = \{N_1, N_2, N_3, N_4\}$ with $N_i < N_j$ if $i < j$ (in particular, $c(N_i) = N_{5-i}$). We set $D := \{N_1, N_2\}$ and $F := c(D)$. Notice that $D < F$, that is, $d < f$ for each $(d, f) \in D \times F$, and hence $df \in X_2$ for each $(d, f) \in D \times F$, as this will be implicitly used in the forthcoming arguments. Since X is self-complementary, it contains no element in $X_2 D \cup F X_2$ for if $w \in X \cap FX_2$, then $c(w) \in X \cap X_2 D$ and thus $\mathcal{G}(X)$ would contain a directed path (of length 3) between two vertices labelled by elements of X_2[3].

We shall prove that there are exactly four valid ways to extend X_2, which in total will give the 32 different codes X. Recall that X has to contain exactly one element in each conjugacy class.

[3] Recall that if I and J are two sets, then $IJ = \{ij \,:\, i \in I, j \in J\}$ is the set of all concatenations of words $i \in I$ and $j \in J$.

We proceed by finding necessary conditions implying which element must be chosen in each conjugacy class. It will remain 6 conjugacy classes with potentially 2 elements that can be chosen. We shall see that only four choices are valid. At the same time, all choices ensure the following condition, which allows one to see that $\mathcal{G}(X)$ has no directed cycle: if a dinucleotide $w \in \mathcal{B}^2$ has an in-neighbour N_i and an out-neighbour N_j, then $N_j > N_i$. It is obviously a necessary condition, for otherwise $\mathcal{G}(X)$ would contain the directed cycle $N_i \to w \to N_j \to N_i$, and is also sufficient as one readily checks. This property also helps checking that $\mathcal{G}(X)$ has no directed path between two dinucleotides in X_2.
Let $\mathcal{A}$ be the set of self-complementary dinucleotides in $\mathcal{B}^2$, *i.e.* $\mathcal{A} = \{AT, TA, CG, GC\}$. The self-complementarity of X implies that it contains no element in $\mathcal{A}D \cup F\mathcal{A}$, because $\mathcal{G}(X)$ contains no directed cycle of length 3.

The words in $D\mathcal{B}F$ yield 16 conjugacy classes, and the two above properties directly determine the element that belongs to X for 10 of them, and leave exactly two choices for each of the remaining classes: the code X must contain

$$N_1N_1N_4, N_1N_2N_3, N_1N_2N_4, N_1N_3N_4, N_1N_4N_4, N_2N_1N_3, N_2N_2N_3, N_2N_3N_3, N_2N_3N_4, N_2N_4N_3.$$

In the conjugacy class of $N_1N_4N_3$, the code X must contain $N_1N_4N_3$ as otherwise it would contain $N_4N_3N_1$ thus creating the cycle $N_2 \to N_4N_3 \to N_1 \to N_2$. This implies that X contains $N_2N_1N_4$. Similarly, X must contain $N_2N_3N_3$. This implies that $N_3N_3N_1 \notin X$ and hence $N_1N_3N_3 \in X$. It thus remains 2 complementary conjugacy classes in $D\mathcal{B}F$ for which we have two choices, namely that of $N_1N_1N_3$ and that of $N_2N_4N_4$.
There are 4 classes that have not been mentioned yet: those of elements in $D^3 \cup F^3$. Similarly as above, those 4 classes are composed of 2 pairs of complementary classes, and only 2 elements in each class can be valid choices. In total, the 6 remaining classes are that of $N_1N_1N_3$, of $N_1N_1N_2$, of $N_2N_2N_1$ and of their anti-trinucleotides. Note that all eight choices are valid. Indeed, since every dinucleotide that appears both as a prefix and as a suffix cannot have an in-neighbour N_i and an out-neighbour N_j with $N_i > N_j$, we infer that one of $N_2N_2N_1$ and $N_2N_1N_1$ does not belong to X. Similarly, one of $N_3N_1N_1$ and $N_1N_1N_2$ does not belong to X. Using that X is self-complementary, we deduce that only four choices can be valid:

$$\begin{aligned}
&(N_1N_1N_3, N_2N_4N_4, N_1N_1N_2, N_3N_4N_4, N_2N_2N_1, N_4N_3N_3),\\
&(N_1N_1N_3, N_2N_4N_4, N_1N_1N_2, N_3N_4N_4, N_2N_1N_2, N_3N_4N_3),\\
&(N_1N_1N_3, N_2N_4N_4, N_2N_1N_1, N_4N_4N_3, N_2N_1N_2, N_3N_4N_3),\\
&(N_3N_1N_1, N_4N_4N_2, N_2N_1N_1, N_4N_4N_3, N_2N_1N_2, N_3N_4N_3).
\end{aligned}$$

Each of these choices ensures the aforementioned property of in-neighbours and out-neighbours of dinucleotides in $\mathcal{G}(X)$. It follows that we obtain precisely 4 different codes for each of the 8 different orders on $\mathcal{B}$, for a total of 32 different maximum self-complementary mixed circular codes in $\mathcal{B}^2 \cup \mathcal{B}^3$. □

6.4. **Proof of Proposition 7.**

Proof. There are 4 maximum self-complementary trinucleotide comma-free codes of 16 trinucleotides (Tables 3a and 3b in Michel et al., 2008). There are 4 maximum dinucleotide comma-free

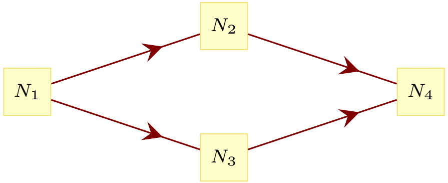

FIGURE 10. The directed graph associated with the 4 codes in Case 1. For instance, for X^1, one has $N_1 = A$, $N_2 = C$, $N_3 = G$ and $N_4 = T$.

codes of 5 dinucleotides (Fimmel and Strüngmann, 2016):

$$(*) \qquad \begin{array}{ll} \{AC, GT, AG, CT, AT\}, & \{AC, GT, GA, TC, GC\}, \\ \{AG, CT, CA, TG, CG\}, & \{GA, TC, CA, TG, TA\}. \end{array}$$

All of them consist of 2 complementary pairs of dinucleotides and 1 self-complementary dinucleotide, *e.g.* $\{AC, GT, AG, CT, AT\} = \{AC, \overleftarrow{c(AC)}, AG, \overleftarrow{c(AG)}, AT = \overleftarrow{c(AT)}\}$. In Case 1 below, we prove that a dinucleotide comma-free code containing 2 complementary pairs of dinucleotides can be mixed with a self-complementary trinucleotide comma-free code of cardinality at most 14. Thus all mixed circular codes containing one of the maximal codes (*) above have cardinality at most 19.

Let us now consider self-complementary dinucleotide comma-free codes of cardinality 4. It is clear due to the comma-freeness that either 2 dinucleotides in the code must be self-complementary, *i.e.* of the form $N_1N_2 = \overleftarrow{c(N_1N_2)}$, and the other 2 complementary to each other, or none of the 4 is self-complementary, *i.e.* they form 2 complementary pairs. Thus a self-complementary mixed comma-free code of cardinality 20 can only be the union of a self-complementary dinucleotide comma-free code of cardinality 4 and a maximum (if possible) self-complementary trinucleotide comma-free code.

Case 1: Let us consider first the case that a self-complementary dinucleotide comma-free code of cardinality 4 contains no self-complementary dinucleotides. In this case, we obtain allover four possibilities:

$$\begin{array}{ll} X^1 = \{AC, GT, AG, CT\}, & X^2 = \{AC, GT, GA, TC\}, \\ X^3 = \{AG, CT, CA, TG\}, & X^4 = \{GA, TC, CA, TG\}. \end{array}$$

We notice that the directed graphs associated with these 4 codes are all isomorphic to the graph depicted in Figure 10. Without loss of generality, it thus suffices to consider the case of X^1.

We also note that the graph associated with X^1 contains 2 directed paths of length 2, namely $A \to C \to T$ and $A \to G \to T$. The existence of these paths means that the trinucleotide code X mixed with X^1 can neither intersect the set $\mathcal{B}^2A \cup T\mathcal{B}^2$, nor contains trinucleotides of the form

$$(+) \qquad dC,\ Cd; \quad \text{or} \quad dG,\ Gd$$

where $d \in \mathcal{B}^2$ is a self-complementary dinucleotide, as this would creates a directed path of length 3 in the associated graph. For instance, for dC ($\overleftarrow{c(dC)} = Gd$), one would obtain the directed path $A \to G \to d \to C$.

Since no trinucleotide in X can start with T or ends with A, there are allover $3 \cdot 4 \cdot 3 = 36$ trinucleotides remaining. After removing the 2 trivial trinucleotides and all trinucleotides

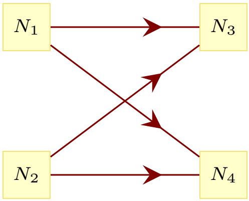

FIGURE 11. The directed graph associated with the 4 codes in Case 2. For instance, for X^1, one has $N_1 = A$, $N_2 = G$, $N_3 = C$ and $N_4 = T$.

satisfying (+)

$$(GCG, GGC, GAT, CCG, CGC, CAT, CGG, GCC, ATC, ATG),$$

we still have 24 trinucleotides remaining. However, among these there are 8 cyclically equivalent pairs ($CAG/AGC, GAC/ACG, ACC/CAC, CCT/CTC$ and their anti-trinucleotides): at most one member of each of these 8 pairs can be in X (because X is circular), so $|X| \leq 16$. The remaining 8 trinucleotides are AAC, AAG, AAT, ACT and their anti-trinucleotides GTT, CTT, ATT, AGT.

So X has cardinality 16 if and only if it contains an element in each of the 16 aforementioned conjugacy classes. We show that this does not happen, and hence X has cardinality at most 14. Specifically, we prove that if X intersects the conjugacy class of CCT, and hence also that of its complement AGG, then it cannot intersect all remaining classes — and hence avoids at least 2 of them. So suppose that X intersects the conjugacy class of CCT. Then either CCT or CTC belongs to X (as we already know that $TCC \notin X$). If $CCT \in X$ then, because $AC \in X^1$, the code X cannot contain trinucleotides starting with CT, which forbids CTT and its complement AAG. Similarly, if $CTC \in X$ then, because $CT \in X^1$, the code X cannot contain trinucleotides ending with CT, which forbids ACT and its complement AGT. Consequently, X has cardinality at most 14.

Case 2: Let us consider now the case that a self-complementary dinucleotide comma-free code of cardinality 4 contains 2 self-complementary dinucleotides. In this case, we obtain allover four possibilities:

$$X^1 = \{AC, GT, AT, GC\}, \quad X^2 = \{CA, TG, TA, CG\},$$
$$X^3 = \{AG, CT, AT, CG\}, \quad X^4 = \{GA, TC, TA, GC\}.$$

Let us point out that, similarly as in Case 1, the directed graphs associated with these 4 codes are isomorphic, this times to the directed graph depicted in Figure 11. Without loss of generality, it thus suffices to consider the case of X^1.

We first note that $c(\pi_1(X^1)) = \pi_2(X^1)$, in particular $\pi_1(X^1) \cap \pi_2(X^1) = \varnothing$. Furthermore we observe that the trinucleotide code, being self-complementarity, must be contained in $\pi_1(X^1)\mathcal{B}\pi_2(X^1)$. Indeed, suppose that the trinucleotide code contains αd with $\alpha \in \pi_2(X^1) = \{C, T\}$ and $d \in \mathcal{B}^2$. Being self-complementary, the trinucleotide code must contain $\overleftarrow{c(d)}c(\alpha)$. In addition, X^1 contains the self-complementary dinucleotide $c(\alpha)\alpha$. As as result, the directed graph associated to the mixed code would contain the directed path

$$\overleftarrow{c(d)} \to c(\alpha) \to \alpha \to d.$$

We infer that

$$X = \pi_1(X^1)\mathcal{B}\pi_2(X^1) = \{N_1N_2N_3 | N_1 \in \{A, G\}, N_3 \in \{C, T\}, N_2 \in \mathcal{B}\},$$

which is a maximal self-complementary comma-free code since

$$\pi_1(X) \cap \pi_3(X) = \varnothing, \quad \pi_1(X) = c(\pi_3(X))$$

(Fimmel *et al.*, 2017b). By mixing the dinucleotide and the trinucleotide codes, we obtain a mixed self-complementary comma-free code since $\pi_1(X) \cap \pi_2(X^1) = \varnothing = \pi_3(X) \cap \pi_1(X^1)$, hence no directed path in the directed graph associated with X can be extended using an edge from the directed graph associated with X^1. There is therefore exactly one way to extend X^1 into a self-complementary mixed comma-free code of cardinality 20, and consequently in total exactly 4 different maximum self-complementary mixed comma-free codes.

□

6.5. List of the 32 maximum self-complementary mixed circular codes of dinucleotides and trinucleotides with cardinality 26.

$$\{AC, GT, AG, CT, AT, CG, AAC, GTT, AAG, CTT, AAT, ATT, CAC, GTG, ACG, CGT, ACT, AGT,$$
$$CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG\},$$
$$\{AC, GT, AG, CT, AT, CG, AAC, GTT, AAG, CTT, AAT, ATT, CCA, TGG, ACG, CGT, ACT, AGT,$$
$$CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG\},$$
$$\{AC, GT, AG, CT, AT, CG, CAA, TTG, AAG, CTT, AAT, ATT, CAC, GTG, ACG, CGT, ACT, AGT,$$
$$CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG\},$$
$$\{AC, GT, AG, CT, AT, CG, CAA, TTG, GAA, TTC, AAT, ATT, CAC, GTG, ACG, CGT, ACT, AGT,$$
$$CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG\},$$
$$\{AC, GT, AG, CT, AT, GC, AAC, GTT, AAG, CTT, AAT, ATT, ACC, GGT, GAC, GTC, ACT, AGT,$$
$$AGC, GCT, ATC, GAT, GCC, GGC, CTC, GAG\},$$
$$\{AC, GT, AG, CT, AT, GC, AAC, GTT, AAG, CTT, AAT, ATT, ACC, GGT, GAC, GTC, ACT, AGT,$$
$$AGC, GCT, GGA, TCC, ATC, GAT, GCC, GGC\},$$
$$\{AC, GT, AG, CT, AT, GC, AAC, GTT, GAA, TTC, AAT, ATT, ACC, GGT, GAC, GTC, ACT, AGT,$$
$$AGC, GCT, ATC, GAT, GCC, GGC, CTC, GAG\},$$
$$\{AC, GT, AG, CT, AT, GC, CAA, TTG, GAA, TTC, AAT, ATT, ACC, GGT, GAC, GTC, ACT, AGT,$$
$$AGC, GCT, ATC, GAT, GCC, GGC, CTC, GAG\},$$
$$\{AC, GT, GA, TC, AT, GC, AAC, GTT, AGA, TCT, AAT, ATT, ACC, GGT, GAC, GTC, ACT, AGT,$$
$$AGC, GCT, GGA, TCC, ATC, GAT, GCC, GGC\},$$
$$\{AC, GT, GA, TC, AT, GC, AAC, GTT, AAG, CTT, AAT, ATT, ACC, GGT, GAC, GTC, ACT, AGT,$$
$$AGC, GCT, GGA, TCC, ATC, GAT, GCC, GGC\},$$
$$\{AC, GT, GA, TC, AT, GC, AAC, GTT, AGA, TCT, AAT, ATT, ACC, GGT, GAC, GTC, ACT, AGT,$$
$$AGC, GCT, AGG, CCT, ATC, GAT, GCC, GGC\},$$
$$\{AC, GT, GA, TC, AT, GC, AAC, GTT, AGA, TCT, AAT, ATT, CCA, TGG, GAC, GTC, ACT, AGT,$$
$$AGC, GCT, AGG, CCT, ATC, GAT, GCC, GGC\},$$
$$\{AC, GT, GA, TC, TA, GC, ACA, TGT, GAA, TTC, TAA, TTA, ACC, GGT, GAC, GTC, GCA, TGC,$$
$$GGA, TCC, GTA, TAC, TCA, TGA, GCC, GGC\},$$
$$\{AC, GT, GA, TC, TA, GC, ACA, TGT, GAA, TTC, TAA, TTA, CCA, TGG, GAC, GTC, GCA, TGC,$$
$$GGA, TCC, GTA, TAC, TCA, TGA, GCC, GGC\},$$

$$\{AC, GT, GA, TC, TA, GC, CAA, TTG, GAA, TTC, TAA, TTA, ACC, GGT, GAC, GTC, GCA, TGC, GGA, TCC, GTA, TAC, TCA, TGA, GCC, GGC\},$$

$$\{AC, GT, GA, TC, TA, GC, ACA, TGT, GAA, TTC, TAA, TTA, CCA, TGG, GAC, GTC, GCA, TGC, AGG, CCT, GTA, TAC, TCA, TGA, GCC, GGC\},$$

$$\{CA, TG, AG, CT, AT, CG, ACA, TGT, AAG, CTT, AAT, ATT, CCA, TGG, ACG, CGT, ACT, AGT, CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG\},$$

$$\{CA, TG, AG, CT, AT, CG, AAC, GTT, AAG, CTT, AAT, ATT, CCA, TGG, ACG, CGT, ACT, AGT, CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG\},$$

$$\{CA, TG, AG, CT, AT, CG, ACA, TGT, AAG, CTT, AAT, ATT, ACC, GGT, ACG, CGT, ACT, AGT, CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG\},$$

$$\{CA, TG, AG, CT, AT, CG, ACA, TGT, AAG, CTT, AAT, ATT, ACC, GGT, ACG, CGT, ACT, AGT, CAG, CTG, GGA, TCC, ATG, CAT, CCG, CGG\},$$

$$\{CA, TG, AG, CT, TA, CG, CAA, TTG, AGA, TCT, TAA, TTA, CCA, TGG, CGA, TCG, CTA, TAG, CAG, CTG, AGG, CCT, TCA, TGA, CCG, CGG\},$$

$$\{CA, TG, AG, CT, TA, CG, CAA, TTG, AGA, TCT, TAA, TTA, CCA, TGG, CGA, TCG, CTA, TAG, CAG, CTG, GGA, TCC, TCA, TGA, CCG, CGG\},$$

$$\{CA, TG, AG, CT, TA, CG, CAA, TTG, GAA, TTC, TAA, TTA, CCA, TGG, CGA, TCG, CTA, TAG, CAG, CTG, AGG, CCT, TCA, TGA, CCG, CGG\},$$

$$\{CA, TG, AG, CT, TA, CG, CAA, TTG, AGA, TCT, TAA, TTA, ACC, GGT, CGA, TCG, CTA, TAG, CAG, CTG, GGA, TCC, TCA, TGA, CCG, CGG\},$$

$$\{CA, TG, GA, TC, TA, CG, CAA, TTG, GAA, TTC, TAA, TTA, CCA, TGG, CGA, TCG, CTA, TAG, CAG, CTG, TCA, TGA, CCG, CGG, CTC, GAG\},$$

$$\{CA, TG, GA, TC, TA, CG, CAA, TTG, AAG, CTT, TAA, TTA, CCA, TGG, CGA, TCG, CTA, TAG, CAG, CTG, TCA, TGA, CCG, CGG, CTC, GAG\},$$

$$\{CA, TG, GA, TC, TA, CG, CAA, TTG, GAA, TTC, TAA, TTA, CCA, TGG, CGA, TCG, CTA, TAG, CAG, CTG, AGG, CCT, TCA, TGA, CCG, CGG\},$$

$$\{CA, TG, GA, TC, TA, CG, AAC, GTT, AAG, CTT, TAA, TTA, CCA, TGG, CGA, TCG, CTA, TAG, CAG, CTG, TCA, TGA, CCG, CGG, CTC, GAG\},$$

$$\{CA, TG, GA, TC, TA, GC, CAA, TTG, GAA, TTC, TAA, TTA, CAC, GTG, GAC, GTC, GCA, TGC, GGA, TCC, GTA, TAC, TCA, TGA, GCC, GGC\},$$

$$\{CA, TG, GA, TC, TA, GC, AAC, GTT, GAA, TTC, TAA, TTA, CAC, GTG, GAC, GTC, GCA, TGC, GGA, TCC, GTA, TAC, TCA, TGA, GCC, GGC\},$$

$$\{CA, TG, GA, TC, TA, GC, CAA, TTG, GAA, TTC, TAA, TTA, ACC, GGT, GAC, GTC, GCA, TGC, GGA, TCC, GTA, TAC, TCA, TGA, GCC, GGC\},$$

$$\{CA, TG, GA, TC, TA, GC, AAC, GTT, AAG, CTT, TAA, TTA, CAC, GTG, GAC, GTC, GCA, TGC, GGA, TCC, GTA, TAC, TCA, TGA, GCC, GGC\}.$$

References

[1] Arquès D.G., Michel C.J. 1996. A complementary circular code in the protein coding genes. Journal of Theoretical Biology 182, 45–58.

[2] Arquès D.G., Michel C.J. 1987. Periodicities in introns. Nucleic Acids Research 15, 7581–7592.

[3] Bang-Jensen J., Gutin, G. 2009. Digraphs. 2nd Ed, Springer Monographs in Mathematics. London: Springer-Verlag London, Ltd.

[4] Baranov P.V., Venin M., Provan G. 2009. Codon size reduction as the origin of the triplet genetic code. PLOS ONE 4(5), 2009: e5708.

[5] Bobay L.-M., Ochman H. 2017. The evolution of bacterial genome architecture. Frontiers in Genetics 8, 1–6.

[6] 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 Research 10, 297–304.

[7] Copley S.D., Smith E., Morowitz H.J. 2005. A mechanism for the association of amino acids with their codons and the origin of the genetic code. Proceedings of the National Academy of Sciences U.S.A. 102, 4442–4447.

[8] Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics 5, 435–445.

[9] El Soufi K., Michel C.J. 2017. Unitary circular code motifs in genomes of eukaryotes. Biosystems 153, 45–62.

[10] Fickett J.W. 1982. Recognition of protein coding regions in DNA sequences. Nucleic Acids Research 10, 5303–5318.

[11] Fimmel E., Giannerini S., Gonzalez D., Strüngmann L. 2015. Dinucleotide circular codes and bijective transformations. Journal of Theoretical Biology 386:159–165.

[12] Fimmel E., Michel C.J., Strüngmann L. 2016. n-nucleotide circular codes in graph theory. Philosophical Transactions of the Royal Society A 374, 20150058.

[13] Fimmel E., Michel C.J., Strüngmann L. 2017a. Diletter circular codes over finite alphabets. Mathematical Biosciences 294 120–129.

[14] Fimmel E., Michel C.J., Strüngmann L. 2017b. Strong comma-free codes in genetic information. Bulletin of Mathematical Biology 79, 1796–1819.

[15] Fimmel E., Michel C.J., Starman M., Strüngmann L. 2018. Self-complementary circular codes in coding theory. Theory in Biosciences 137, 51–65.

[16] Fimmel E., Strüngmann L. 2016. Maximal dinucleotide comma-free codes. Journal of Theoretical Biology 21, 206–13.

[17] Fimmel E., Strüngmann L. 2018. Mathematical fundamentals for the noise immunity of the genetic code. Biosystems 164, 186–198.

[18] Frank A, Froese T. 2018. The standard genetic code can evolve from a two-letter GC code without information loss or costly reassignments. Origins of Life and Evolution of Biospheres 48, 259–272.

[19] Johnson D.B.F., Wang L. 2010. Imprints of the genetic code in the ribosome. Proceedings of the National Academy of Sciences U.S.A. 107, 8298–8303.

[20] Gemayel R., Vinces M.D., Legendre M., Verstrepen K.J. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annual Review of Genetics 44, 445–77.

[21] Gonzalez D., 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.

[22] Hartman H. 1995. Speculations on the origin of the genetic code. Journal of Molecular Evolution 40, 541–544.

[23] Konopka A.K., Smythers G.W. 1987. DISTAN - A program which detects significant distances between short oligonucleotides. Bioinformatics 3, 193–201.

[24] Krupkin M., Matzov D., Tang H., Metz M., Kalaora R., Belousoff M.J., Zimmerman E., Bashan A., Yonath A. 2011. A vestige of a prebiotic bonding machine is functioning within the contemporary ribosome. Philosophical Transactions of the Royal Society B 366, 2972–2978.

[25] Kun Á., Radványi Á. 2018. The evolution of the genetic code: impasses and challenges. Biosystems 164, 217–225.

[26] Michel C.J. 1986. New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its evaluation. Journal of Theoretical Biology 120, 223–236.

[27] Michel C.J. 2008. A 2006 review of circular codes in genes. Computer and Mathematics with Applications 55, 984–988.

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

[29] Michel C.J. 2017. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, archaea, eukaryotes, plasmids and viruses. Life 7(20), 1–16.

[30] Michel C.J., Pirillo G. 2013. Dinucleotide circular codes. ISRN Biomathematics 2013, Article ID 538631, 1–8.

[31] Michel C.J., Pirillo G, Pirillo M.A. 2008. Varieties of comma free codes. Computer and Mathematics with Applications 55, 989–996.

[32] Noller H.F. 2004. The driving force for molecular evolution of translation. RNA 10, 1833–1837.

[33] Patel A. 2005. The triplet genetic code had a doublet predecessor. Journal of Theoretical Biology 233, 527–532.

[34] Saad N.Y. 2018. A ribonucleopeptide world at the origin of life. Journal of Systematics and Evolution 56, 1–13.

[35] Seligmann H. 2014. Putative anticodons in mitochondrial tRNA sidearm loops: Pocketknife tRNAs? Journal of Theoretical Biology 340, 155–63.

[36] Shepherd J.C.W. 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proceedings of the National Academy of Sciences U.S.A. 78, 1596–1600.

[37] Wilhelm T., Nikolajewa S. 2004. A new classification scheme of the genetic code. Journal of Molecular Evolution 59, 598–605.

[38] Wu H.L., Bagby S. van den Elsen J.M. 2005. Evolution of the genetic triplet code via two types of doublet codons. Journal of Molecular Evolution 61, 54–64.

[39] Yáñez-Cuna J.O., Arnold C.D., Stampfel G., Boryń Ł.M., Gerlach D., Rath M., A. Stark. 2014. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. Genome Research 24, 1147–1156.

[40] Yarus M. 2017. The genetic code and RNA-amino acid affinities. Life 7, 1–16.