Mixed circular codes

Elena Fimmel\textsuperscript{a}, Christian J. Michel\textsuperscript{b,c}, François Pirot\textsuperscript{b,c}, Jean-Sébastien Sereni\textsuperscript{b}, Lutz Strüngmann\textsuperscript{a}

\textsuperscript{a} Institute of Mathematical Biology, Faculty for Computer Sciences, Mannheim University of Applied Sciences, 68163 Mannheim, Germany
\textsuperscript{b} Theoretical Bioinformatics, K\&i, C.N.R.S., University of Strasbourg, 300 Boulevard Sébastien Brant, 67400 Illkirch, France
\textsuperscript{c} LORIA (Orpailleur) and Department of Mathematics, University of Lorraine and Radboud University, Vandœuvre-lès-Nancy, France and Nijmegen, Netherlands

A R T I C L E I N F O

Keywords:
Mixed circular codes
Dinucleotide circular codes
Trinucleotide circular codes
Tetranucleotide circular codes
Graph properties

A B S T R A C T

By an extensive statistical analysis in genes of bacteria, archaea, eukaryotes, plasmids and viruses, a maximal C\textsuperscript{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 \( (\text{di,tri}) \)-nucleotides, \( (\text{tri,tetra}) \)-nucleotides and \( (\text{di,tri,tetra}) \)-nucleotides are constructed, respectively. In particular, we show that any maximal dinucleotide circular code of size 6 can be embedded into a maximal \( (\text{di,tri}) \)-nucleotide circular code such that its trinucleotide component is a maximal C\textsuperscript{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, \textit{i.e.} archaea, bacteria, eukaryotes, viruses, plasmids, mitochondria and chloroplasts, contain regions for coding proteins, \textit{i.e.} a series of amino acids. These genomic regions are called (protein coding) genes. A gene is a series of words of the same length equal to 3 nucleotides, \textit{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 \cite{1,2} and \cite{3,Figure 1}. Ten years later, this periodicity modulo 3 has been explained by a maximal C\textsuperscript{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 \cite{4-6}. It contains the following 20 trinucleotides

\begin{equation}
X = \{ \text{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC} \}.
\end{equation}

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

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

\{ \text{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 \cite{7,8}. 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 \cite{9}. 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),

\* Corresponding author.

E-mail addresses: e.fimmel@hs-mannheim.de (E. Fimmel), c.michel@unistra.fr (C.J. Michel), francois.pirot@loria.fr (F. Pirot), sereni@kam.mff.cuni.cz (J.-S. Sereni), l.struengmann@hs-mannheim.de (L. Strüngmann).

https://doi.org/10.1016/j.mbs.2019.108231

Received 7 September 2018; Received in revised form 16 July 2019; Accepted 17 July 2019
Available online 17 July 2019

0025-5564/© 2019 Elsevier Inc. All rights reserved.
DNA transposons, etc. Introns of eukaryotes have no nucleotide periodicity modulo 3 (Figure 2 in [3]) but a nucleotide periodicity modulo 2 generated by dinucleotides [10,11]. 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, [12–15]). Some dinucleotide repeats are highly enriched in enhancers which are genomic elements involved in gene expression [16]. 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 [15]. 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, [17–22]). 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

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

where A stands for adenine, C for cytosine, G for guanine and T for thymine. As commonly used in word theory, $$B^\ell := \{N_1 \ldots N_m \mid N_i \in B, n \in \mathbb{N}\}$$ is the set of all words over B of finite length including the empty word $$\varepsilon$$ while $$B^\ast = B^\ast \setminus \{\varepsilon\}$$.

**Definition 1.**

1. A set $$X \subseteq B^\ast$$ is a code if every word $$w \in X$$ has a single decomposition into words from X, i.e. whenever $$w = w_1 \ldots w_n = w'_1 \ldots w'_m$$ for words $$w_i, w'_i \in X$$, then $$n = m$$ and $$w_i = w'_i$$ follow.
2. For $$\ell \in \mathbb{N}$$ with $$\ell \geq 2$$, an $$\ell$$-letter code is a subset of $$B^\ell$$.
3. Elements of $$B^\ell$$ are called $$\ell$$-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 $$B^\ast$$, we define $$w_1 \cap w_2$$ to be the largest tail segment (suffix) of $$w_1$$ that is an initial segment (prefix) of $$w_2$$. (Note that $$w_1 \cap w_2$$ might be the empty word $$\varepsilon$$.) If $$w_1 \neq w_2$$ and $$w_1 \cap w_2 \neq \varepsilon$$, then $$w_1$$ overlaps $$w_2$$.

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

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

$$AC \mid TG = AC \mid GT \mid GA.$$  

Some codes may have additional properties.

**Definition 2.** A code $$X \subseteq B^\ast$$ is

1. comma-free if every concatenation $$w_1w_2$$ 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 $$X^2 \cap B^\ast X B^\ast = \emptyset$$;
2. circular if for any finite concatenation $$w_1 \ldots w_m$$ of elements from $$(m \in \mathbb{N})$$, there is only one partition into elements from $$X$$ when read on a circle. Any such partition is a circular decomposition of $$w_1 \ldots w_m$$.

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 (Figs. 1 and 2).

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

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

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

$$SW(\omega \varepsilon c); \{A, T, C, G\} \rightarrow \{T, A, G, C\}$$

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 $$\alpha$$: a given trinucleotide $$x = (b_1, b_2, b_3) \in B^3$$ leads to the trinucleotide $$\alpha x = (b_3, b_2, b_1)$$. 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 $$\ell$$-nucleotide, i.e. if $$x = N_1 N_2 \ldots N_{\ell-1} N_\ell \in B^\ell$$ then $$\alpha x = N_\ell N_{\ell-1} \ldots N_2 N_1 \in B^\ell$$.

**Example 2.** $$\alpha c(\{AC, AT\}) = \{AT, GT\}$$ and $$\alpha c(\{CG, GA, AT\}) = \{AC, TG\}$$.

**Definition 3.**

1. A code $$X \subseteq B^\ast$$ is self-complementary if $$X = c(X)$$.
2. A trinucleotide circular code $$X \subseteq B^3$$ is called a C*-code if both the circular shifted codes $$X' := \{N_2 N_3 N_1 \mid N_1, N_2, N_3 \in X\}$$ and $$X'' := \{N_1 N_2 N_3 \mid N_1, N_2, N_3 \in X\}$$ are circular as well.

Let us remark that the C*-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*-self-complementary code.

**Remark 3.** The genetic code $$B^3$$ 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*-code) defined over $$B$$.
Definition 6. For all positive integers $\ell$ and $i$ such that $1 \leq i \leq \ell$, we define the projection on the $i$th coordinate $\pi_i : B^\ell \to B$ by

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

The projections on two coordinates $\pi_i, \pi_j : B^\ell \to B^2$ are defined in a similar way whenever it makes sense.

Theorem 1 (Fimmel et al., [23]). The following statements hold for every integer $\ell \geq 2$.

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

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

1. $V(X) := \{N_1 \cdots N_\ell, N_{j+1} \cdots N_\ell : N_1 N_2 \cdots N_\ell \in X, 1 \leq j \leq \ell - 1\};$ and
2. $E(X) := \{[N_1 \cdots N_j, N_{j+1} \cdots N_\ell] : N_1 N_2 \cdots N_\ell \in X, 1 \leq j \leq \ell - 1\}.$

The graph $G(X)$ is said to be associated with $X$. An example of graph $G(X)$ is given in Fig. 3.

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

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

We investigate the mixed codes $X \subseteq B^2 \cup B^3 \cup B^4$ 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 $\ell$-nucleotides of various values of $\ell$ 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 $\ell$-nucleotide sets (for a fixed $\ell$) which are codes necessarily. The next subsection deals with examples showing that a mixing of circular codes over $B^2$ and $B^3$ 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 $X = X_2 \cup X_3 \subseteq B^2 \cup B^3$ where $X_2 \subseteq B^2$ and $X_3 \subseteq B^3$ 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
Example 4. Consider the two circular codes 
$X_2 = \{GT\}$ and $X_1 = \{AC,G,T\}$. 
The mixed set 
$X = \{GT,AC,T\}$ is a code that is not circular. 
Indeed, the word $TACGT$ admits the two circular decompositions 
$TACGT\rightarrow AC$ and 
$TACGT\rightarrow AC$. The graph $G(X)$ of $X$ (defined naturally for mixed codes; see the next section for a formal definition) 
contains the cycle $G \rightarrow T \rightarrow AC \rightarrow G$ (Fig. 4).

Example 5. Consider the two circular codes 
$X_2 = \{AC,AT,G\}$ and $X_1 = \{GA,AT\}$. 
The mixed set 
$X = \{AC,AT,G,A,GA,TA\}$ is not a code as the word $GAATAC$ has two different decompositions into words in $X$, namely $GA|AT|AC$ and $GA|AT|AC$. However, the graph $G(X)$ of $X$ (naturally defined; see the next section for a formal definition) is acyclic (Fig. 5).

Example 6. Consider the two circular codes 
$X_2 = \{AC\}$ and $X_1 = \{AC,CGT,CTG,GAG,GT\}$. 
The mixed set 
$X = \{AC,AC,CGT,CTG,GAG,GT\}$ is not a code as the word $ACAGTGCTGAC$ has two different decompositions into words in $X$, namely $AC|ACG|TCT|ACG$ and $ACAGT|CTG|ACG$. Moreover, the graph $G(X)$ of $X$ (naturally defined; see the next section for a formal definition) contains the cycle $AC \rightarrow A \rightarrow C \rightarrow GT \rightarrow T \rightarrow CT \rightarrow G \rightarrow AC$ (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 $X_i \subseteq B_i$, $X_i \subseteq B_i$ and $X_i \subseteq B_i$ be circular codes. We want to know whether the mixed set $X := X_1 \cup X_2$ is a circular code. The associated graph $G(X)$ of $X$ is $G(X) = G(X_1) \cup G(X_2)$, respectively $G(X) = G(X_1) \cup G(X_2)$. 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 A.1.

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

(1) The mixed set $X$ is a code if and only if there exists no directed path in $G(X)$ between a pair of vertices with labels in $X_i$.
(2) The mixed set $\hat{X}$ is a code if and only if there exists no directed path in $G(X)$ between a pair of vertices with labels in $X_3$.
(3) The mixed sets $X$ and $\hat{X}$ are circular codes if and only if $X$ and $\hat{X}$ are codes and $G(X)$ and $G(\hat{X})$ are acyclic.

**Theorem 2** directly implies the following statement.

**Corollary 1.** For $i \in \{1,2\}$, let $X_i$ be a subset of $B^i$ and set $X := X_1 \cup X_2$. 
Let $G(X)$ be the digraph [24] obtained from $G(X)$ by identifying all vertices of $G(X)$ corresponding to the dinucleotides in $X_i$ (keeping loops and multiple edges should they arise). 
The mixed set $X$ is a circular code if and only if the digraph $G(X)$ is acyclic. A similar statement holds for mixed subsets $X \subseteq B^1 \cup B^2$.

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 $X \subseteq B^1 \cup B^2$ of dinucleotides and trinucleotides. We point out that, as it has been noticed already [25], 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 $X \subseteq B^1 \cup B^2$ 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 $D \subseteq B^2$ be a maximal dinucleotide circular code. Then $D$ can be embedded into a mixed code $X \subseteq B^2 \cup B^3$ such that

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

**Proof.** Let $D \subseteq B^2$ 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 $X_3 \subseteq B^3$ by adding letters to the dinucleotides of $D$ in the first position ensuring at the same time that $\pi_{12}(X_3) \cap \pi_{23}(X_3) = \emptyset$. 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 [26] and by Fimmel et al. [25,27], any maximal dinucleotide circular code is of the following form:

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

writing $B = \{ N_1, N_2, N_3, N_4 \}$. 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,

$$X_3 = \{ N_1 N_1 N_2, N_1 N_1 N_3, N_1 N_1 N_4, N_2 N_1 N_2, N_3 N_1 N_2, N_4 N_1 N_2, N_2 N_1 N_3, N_2 N_1 N_4, N_3 N_1 N_3, N_3 N_1 N_4, N_4 N_1 N_3, N_4 N_1 N_4, N_1 N_2 N_3, N_1 N_2 N_4, N_2 N_2 N_3, N_2 N_2 N_4, N_3 N_2 N_3, N_3 N_2 N_4, N_4 N_2 N_3, N_4 N_2 N_4, N_1 N_3 N_4, N_2 N_3 N_4, N_3 N_3 N_4, N_4 N_3 N_4 \}.$$

Consequently, $|D \cup X_3| = 26$ and $\pi_{23}(X_3) = D$ while $\pi_{12}(X_3) = \{ N_i N_j : 1 \leq i \leq j \leq 4 \}$, 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 $G(X)$ 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 $G(X)$ labelled by a dinucleotide has both positive in-degree and positive out-degree. Several consequences readily follow. First, $G(X)$ cannot contain a path starting at a vertex in $D$, so Theorem 2(1) implies that $X$ is a code. Second, $G(X_3)$ contains no path of length greater than 2 and no cycle of length 2. The graph $G(X_3)$ is thus acyclic and $X_3$ is a comma-free circular code. Further, no cycle in $G(X)$ can contain a vertex labelled by a dinucleotide. Since $D$ itself is circular, and hence $G(D)$ is acyclic, we infer that $G(X)$ is acyclic and hence $X$ is a circular code by Theorem 2.

It remains to show that $X$ is a $C^3$-code. Let $X' := \{ M_2 M_2 M_3 : M_2, M_3 \in X \}$ and $X'' := \{ M_2 M_3 M_2 : M_2, M_3 \in X \}$ be the two circular shifts of $X_3$. We show that $X' \subseteq X$ is a circular code, the argument being analogue for $X''$. It follows from the definition of $X_3$ that

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

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

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

where $N_{i_k} \in B$ for each $k \in \{1, \ldots, 7\}$. The form of $X'$ then implies that

$$4 \geq i_6 > i_5 > i_4 > i_3 > i_2 > i_1 \geq 1,$$

leading to $i_4 = i_5 = 3, i_3 = i_1 = 2$ and $i_2 = 1$. 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, $X'$ 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 circular code $X \subseteq B^2 \cup B^3$ is 26.

In fact, we will show in Section 4 that there are exactly 32 self-complementary maximum mixed circular codes $X \subseteq B^2 \cup B^3$ 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 [25].

---

### 3.2. Construction of maximum mixed circular codes of trinucleotides and tetraneutelodous

As in Theorem 3, we can also construct maximum (and hence maximal) mixed circular codes $X \subseteq B^2 \cup B^3$ of trinucleotides and tetraneutelodous. These codes have a cardinality of 80 for the following reason. Let $X \subseteq B^2 \cup B^3$ be a mixed circular code. Then $X_3 := X \cap B^3$ and $X_4 := X \cap B^4$ are also circular codes. Therefore, $X_3$ has cardinality at most 20 and $X_4$ at most 60 as there are 256 tetraneutelodous among which the 16 tetraneutelodous of the form $N_1 N_2 N_3 N_2$ for some $N_1, N_2 \in B$ cannot be part of a circular code. Any other tetraneutelodous 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 $X \subseteq B^2 \cup B^3$ of trinucleotides and tetraneutelodous of cardinality 80.

**Proof.** We shall construct a maximum mixed circular code $X \subseteq B^2 \cup B^3$ of cardinality 80 by taking the union of

$$X_3 := \{ AAC, AAC, AAG, AAT, CAC, CAT, CAG, CTT, GAC, GAG \}$$

and

$$X_4 := \{ AAC, AAG, AAT, ACAT, AGAT, AGGC, AGCT, AGGT, ATCC, CAG, CAA, CAC, CAG, CCG, CCG, CCT, GCT, GGT, GGT, GGT, CAT, TAC, TAG, TGAT, TCCT, TTG, TGT, GAC, GAG, GAT, GTG, GCT, GCT, GGT, GTG, TGG, TGT, TTG, TGT \}.$$

This construction keeps the property that $\pi_{12}(X_3) \cap \pi_{23}(X_3) = \emptyset$. Since $X_3 \subseteq \pi_{23}(X_3)$, there is no directed path in $G(X)$ 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 $B \cup B^3$ is acyclic. It remains to check the component induced by $B^3$, which by the construction is a transitive tournament,2 and hence acyclic too. □

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 $X \subseteq B^2 \cup B^3$ of dinucleotides and tetraneutelodous 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 tetraneutelodous circular code is 60.

We finally approach the construction of mixed circular codes containing dinucleotides, trinucleotides and tetraneutelodous 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

---

1 A conjugacy class of an $\ell$-nucleotide $N_1 \cdots N_{\ell} \in B^\ell$ is defined as the set $\{ N_1 \cdots N_{\ell} \cdots N_{\ell-k+1} : 1 \leq k \leq \ell \}$; clearly an $\ell$-nucleotide circular code can contain at most one $\ell$-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 [24] for the notion of tournaments and to the work by Fimmel et al. [23].

---

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 A.2.

**Theorem 5.** Let \( M_1 < M_2 < M_3 < M_4 \) be an ordering of the alphabet \( B \).

For a natural number \( n > 1 \), the set \( X^{(n)} \subset B^G \) defined by

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

is a maximal mixed circular code.

**Corollary 3.** There are maximal mixed circular codes in \( B^2 \cup B^3 \cup B^4 \) of dinucleotides, trinucleotides and tetranucleotides of cardinality \( 71 \).

**Proof.** A maximal mixed circular code \( X \subset B^2 \cup B^3 \cup B^4 \) is constructed by removing from the maximal mixed circular code \( X' \subset B^2 \cup B^3 \) 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 \( X \subset B^2 \cup B^3 \cup B^4 \) of cardinality \( 71 \): \( X := X_2 \cup X_3 \cup X_4 \) where

\[
X_2 := \{ AC, AG, AT, CG, CT, CT \}
\]

\[
X_3 := \{ AAC, AAG, AAT, CAC, CAG, CAT, CCT, GAC, GAG, GAT, GCG, GCT, GTG, TAC, TAG, TAT, TCG, TCT, TGT \}
\]

and

\[
X_4 := \{ AAC, AAG, AAT, CAC, CAAC, CAAG, CAAT, CCAC, CCAG, CCAT, CGCC, CCCT, GAAC, GAG, GAAT, GCAC, GCAG, GCG, GCCG, CGCC, GCGG, GCT, GCTG, GTGC, GGT, TAC, TAG, TAT, TGTC, TCT, TTC, TTG, TTGT \}
\]

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 \( \binom{4}{2} = 6 \) 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 \( B \). Then the following set \( X \subset B^G \) where

\[
X := \{ N_m \cdots N_1 : N_i \in B, m \in \mathbb{N} \\
\text{and } N_m \geq N_{m-1} \geq \cdots \geq N_2 \text{ but } N_2 < N_1 \}
\]

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 \( B^2 \cup B^3 \) 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**

| \(|X|\) | 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 | 128 | 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 | 121664 |
| 14 | 0 | 472 | 1376 | 35172 | 6168 | 32104 | 4472 | 18032 | 98796 |
| 15 | 0 | 316 | 560 | 40996 | 4208 | 39728 | 2872 | 33722 | 121729 |
| 16 | 0 | 176 | 328 | 26520 | 3408 | 23176 | 2872 | 23168 | 88648 |
| 17 | 0 | 72 | 24 | 21356 | 1592 | 19112 | 968 | 27120 | 80284 |
| 18 | 0 | 40 | 32 | 13360 | 1088 | 20504 | 968 | 17360 | 53352 |
| 19 | 0 | 8 | 32 | 7404 | 336 | 13404 | 152 | 12672 | 33976 |
| 20 | 0 | 4 | 32 | 4432 | 192 | 8456 | 152 | 7600 | 20836 |
| 21 | 0 | 0 | 32 | 1648 | 12 | 3736 | 8 | 3264 | 8688 |
| 22 | 0 | 0 | 32 | 936 | 16 | 2144 | 8 | 1832 | 4936 |
| 23 | 0 | 0 | 0 | 224 | 24 | 560 | 0 | 400 | 1184 |
| 24 | 0 | 0 | 0 | 120 | 0 | 296 | 0 | 208 | 624 |
| 25 | 0 | 0 | 0 | 0 | 16 | 0 | 32 | 0 | 16 | 64 |
| 26 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 0 | 8 | 32 |

| Total | 164 | 11788 | 24768 | 314700 | 44016 | 271900 | 28144 | 190232 | 885712 |

---


---
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 $N_1N_2 = (c(N_1)c(N_2))$ and a self-complementary trinucleotide circular code of cardinality 2 since self-complementary trinucleotide codes 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 $(AT, TA, AT)$ and $(GC, GC, GC)$. Consequently, there are at most $4 \cdot 28 = 112$ self-complementary mixed circular codes of cardinality 3. Each such code $X$ has the form $\{N_1N_2N_3, c(N_1)c(N_2)c(N_3)\}$ where $N_1, N_2, N_3 \in B$ 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 $X = \{N'c(N')N, (c(N))'c(N'), N(c(N))\}$ where $N' \in B$ and $N' \notin B \setminus \{N\}$. 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 $Q(X)$ must have a cycle that contains the edge $N \rightarrow c(N)$. Necessarily, the vertex preceding $N$ on the cycle is labelled by a dinucleotide $d$, and $d$ 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 $X = \{N'c(N')N, (c(N))'c(N'), N(c(N))\}$, which finishes the proof.

Proposition 3. The number of self-complementary mixed circular codes in $B^2 \cup B^3$ 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 $X = \{N_1N_2N_3, c(N_1)c(N_2)c(N_3), M_1M_2, M_1M_4\}$, where $N_1, N_2, N_3, M_1, M_2, M_3, M_4 \in B$. Moreover, there are 28 self-complementary trinucleotide circular codes of cardinality 2 (see the proof of Proposition 2) and 8 self-complementary mixed 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 $28 \cdot 2 \cdot 8 = 224$ mixed self-complementary circular codes of cardinality 4.

Now observe that $X$ is not circular if and only if $Q(X)$ has a cycle containing a path of the form $d_1 \rightarrow L_1 \rightarrow L_2 \rightarrow d_2$ where $d_1, d_2 \in B^2$ are dinucleotides and $L_1, L_2 \in \{M_1M_2, M_1M_4\}$. 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 $N, N' \notin B$. Moreover, since $c(N)L_1 \notin X$, we infer that $L_2c(N')$ 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 $\{M_1M_2, M_1M_4\}$ 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 $L_1 \in B$, $L_2 \in B \setminus \{L_1, c(L_1)\}$ and $N \in B$. Because the codes containing a trinucleotide of the form $Mc(M)M$ for $M \in B$ 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 $B^2 \cup B^3$ that have a maximal cardinality of 26. However, none of them contains the maximal $C^1$ self-complementary circular code observed in genes (see Proposition 5). The proof of the following result is contained in Appendix A.3 and the list appears in Appendix A.5.

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

We now consider the maximal $C^1$-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 $B^2 \cup B^3$ that contains the maximal $C^1$-self-complementary trinucleotide circular code $X$ of cardinality 20 observed in genes, defined by (1). Then $Y \cap B^2 \subseteq \{(AT), (GC), (AT, GC)\}$. In other words, $Y \subseteq \{AT, GC, AAC, GTC, GAT, TAA, ATT, ACC, GGT, GAC, GTC, CAG, CTG, GTA, TAC, ATC, GAT, GCC, GGC, GTC, GAG\}$.

Proof. Let $Y$ be as stated in the proposition. We systematically exclude all dinucleotides except for AT and GC. Let $X$ be the set of all dinucleotides that are not self-complementary, ignoring the trivial ones (NN for $N \in B$). 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 $GT$ and $TTC$ belong to $Y$, the digraph $Q(Y)$ contains a directed path from $GA$ to $TC$, and hence the pair $(GA, TC)$ cannot belong to $Y$. Similarly, the fact that $GGT$ and $TAC$ belong to $Y$ forbids the pair $(GA, CT)$. For the latter two cases, if $(AG, CT) \notin Y$, then as $GTA$ and $GTT$ also belong to $Y$ the digraph $Q(Y)$ would contain a directed cycle, namely $GT \rightarrow A \rightarrow G \rightarrow GT$.

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 $N'c(N')N, c(N)c(N')N'$ for some $N' \notin B$, since the obtained code would not be circular. Indeed, the associated digraph would contain the cycle $N'c(N') \rightarrow N \rightarrow c(N) \rightarrow N'c(N')$.

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 $Y \cap B^2$ can indeed be equal to $\{AT, GC\}$, as the obtained code is then circular (and self-complementary). The digraph $Q(Y)$ associated with $Y$ in this case is given in Fig. 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 $B^2 \cup B^3$. See Appendix A.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:

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

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 [30]. The absence of a code, as proposed by Noller [31] and Krupkin et al. [32], 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 [33]). 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 Glu, 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 aminoaeryl-tRNA synthetase and a peptidyl-transferase ribozyme [34]. Circularity could have been necessary for determining the position of nucleotide motifs in the primitive construction of the genetic code [35].

For the second stage, many schemes have suggested that a simpler code based on dinucleotides preceded the modern trinucleotide genetic code (see, for instance, [17,19,22]), 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 [36]. 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 [37]. 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” [38]. 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.

Fig. 8 proposes an evolutionary hypothesis of self-complementary mixed circular codes 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 ℓ (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 $SMC_ℓ$ 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 $GC$. 

Fig. 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*-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 (AF, 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.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

A.1. Proof of Theorem 2

Proof. First note that for every edge e = wu′ ∈ E(X), the concatenation wu′ of the labels of the end-vertices of e yields an element of X, by the definition of Q(X).

(1) Assume that there exists a directed path w1, . . . , wr ∈ Q(X) such that both w1 and wr belong to X. Then the word w := w1 · · · wr has an ambiguous decomposition. Indeed, if r is odd then both w1[w2w3| . . . |wr−1wr] and w1[w2| . . . |wr−1wr−1] are decompositions of w into elements of X. Further, if r is even then both w1[w2| . . . |wr−1wr] and w1[w2w3| . . . |wr−2wr−1] are decompositions of w into elements of X. Conversely, assume that the mixed set X is 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 w1| . . . |wr ∈ X∗ and w′ 1| . . . |wr ∈ X∗ be two such decompositions of w. Since w is minimal, w1 = w′ 1 and w′ 1 = w1, and hence we may assume without loss of generality that w1 ∈ X, and w1 ∈ X. Similary, either w1 is in X, and w′ 1 is in X, or w1 in X and w′ 1 in X. Both cases being similar to analyse, let us suppose that the latter one occurs. We show that w1 overlaps w′ 1 for each i ∈ {1, . . . , r}, and that w′ 1 overlaps w1+i for each i ∈ {1, . . . , r−1}. We set x1 := w1 ∩ w′ 1 and x2 := w′ 1 ∩ w1+i, and also show that there is a path in Q(X) from w1 to x1 to x2+i. We proceed by a (finite) induction on the index i. (i) If i = 1 then w1 = w1N for some N ∈ B, so w1 overlaps w′ 1. Further, w1 ∩ w1 = N, and hence w1 overlaps w1+i. By definition, w1 = x1 and so Q(X) contains a (trivial) path from w1 to x1. (ii) Fix i ∈ {2, . . . , r} and assume, by induction, that w1+i overlaps w′ 1+i, that w1+i overlaps w′ 1+i, and that Q(X) contains a path from x1 to x2+i. If w1 does not overlap w′ 1, then w1 is contained in w′ 1+i. Since each of them has length either 2 or 3, and moreover w1+i overlaps w′ 1+i, it follows that w1 is a suffix of w′ 1+i. Consequently, the word w1−1−i+1 overlaps w′ 1−1+w′ 1+i, and so admits two decompositions into elements of X, which is forbidden by the minimality of w. As a result, w1 overlaps w′ 1−1+i, and for the same reason if i < r then w′ 1−i overlaps w1+i−1. It follows that w1+i−1 = x1−1x2+i−2 and w1+i = x2+i−1. Therefore, Q(X) contains an edge from x1+i−2 to x1+i and an edge from x2+i−2 to x2+i−1, which can be used to extend the path from x1 to x2+i−2 into a path from x1 to x2+i−1.

Applying the above with i = r, we deduce that Q(X) contains a path from w1 to x2+i = w′ 1, which are two dinucleotides contained in X12+i.

(2) The proof of the statement of the theorem for X is analogous to (1).

(3) Assume that Q(X) contains a directed cycle w1, . . . , wr. If r is odd then the cyclic word w1 · · · wr−1wr has two circular decompositions into words on X, namely

w1[w2w3| . . . |wr−1wr] and w1[w2| . . . |wr−1wr−1].

The case where r is even is similar, the word w1 · · · wr−1 admitting the two circular decompositions w1[w2w3| . . . |wr−2wr−1] and w1[w2| . . . |wr−2wr−1].

Conversely, assume that X is a code that contains a word with two circular decompositions w1[ . . . |w′ 1| . . . |w′ r] of X. Assume without loss of generality, that |w1| ≤ |w′ 1|. Then we may assume that for every index i, the word w overlap w′ i and w′ 1 overlaps w1+i. 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 x1 := w1 ∩ w′ 1 and x2 := w′ 1 ∩ w1+i, it follows that x1 → · · · → x2 → x1 is a directed cycle in Q(X), which ends the proof. □

A.2. Proof of Theorem 5

Proof. The following property, coined Property O for future references, directly follows from the definition of X(0).

(O): Let w and w′ be two words in X(0). If w overlaps w′, then either w′ is a suffix of w, or w ∩ w′ = N ∈ B.

In particular, Property O implies that a word in X(0) cannot be a prefix of another word in X(0). This readily implies that X(0) is a code. Indeed, suppose on the contrary that there is a concatenation of words w = w1 · · · wr from X(0) that has a second decomposition w = w′ 1 · · · w′ r with w′ 1 ∈ X(0). Up to considering such a word w of minimal length, we may assume that w′ 1 = w′ 2 and |w1| < |w′ 1|. However, this implies that w′ 1 is a prefix of w′ 2, a contradiction. Thus X(0) 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 w1 · · · wr and w′ 1 · · · w′ r into words in X(n), with ε ≥ 2. The fact that X(n) is a code implies that if w overlaps w′ 1, then w′ 1 cannot be a suffix of w. Consequently, Property O implies that w ∩ w′ 1 is a one-letter word. In particular, w overlaps w′ 1, which overlaps w1, from which it follows that w′ 1 has length 2. We therefore infer that |w1| = 2 for each i ∈ {1, . . . , k} and |w1| = 2 for each i ∈ {1, . . . , r′}. 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 ≥ 2, the statement following from Theorem 3 when n = 2. Assume that X(0) is maximal for all k ≤ n and set X := X(n+1) ⊆ B(n+1). Assume furthermore that X is not maximal, i.e. there is a word w ∈ B(n+1) such that X ∩ {w} is a circular code. Since X ∩ B(n) = X(n) is a maximal mixed circular code by the induction hypothesis, w = w1 · · · wr ∈ B(n). The construction of X and the fact that w ∈ X imply that w can be decomposed into words from X ∩ B(n) and a word of the form v = N1 · · · Nr such that N1 ≥ · · · ≥ Nr and ε ≥ 0, with v = ε if ε = 0. Let w = w1 · · · wr be such a decomposition, where t can be zero — implying that w = v. If v = ε, then w and w1 · · · wr are two decompositions of w into words from X ∪ {v}, which contradicts that X ∪ {w} is a code. We deduce that v = ε. We now distinguish two cases.

(i) If t = 0, then w = v = N1 · · · Nr with N1 ≥ · · · ≥ Nr. Clearly, we cannot have N1 = · · · = Nr as this would contradict the circularity of X ∪ {w}. Therefore, N1 < Nr+1. As a result, N1+1N1 ∈ X, which contradicts the circularity of X ∪ {w}.
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 $N_2 = C$.

(ii) If $t \neq 0$, then $v = N_v \cdots N_1$ with $\ell \leq n - 1$. Consequently, the word $wM_1M_2$ has two different decompositions over $X \cup \{w\}$, namely 

$wM_1M_2 = w_1 \cdots w_{\ell} M_1 M_2$ since $M_1 M_2 \in X$ 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.

A.3. Proof of Proposition 4

Proof. A circular code over $B^\ell$ intersects each conjugacy class of length $\ell$ in at most one $\ell$-nucleotide. We proceed by finding all self-complementary mixed circular codes contained in $B^2 \cup B^3$ 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 $B^2 \cup B^3$. It follows that $X_3 := X \cap B^3$ is one of the 8 maximum (of cardinality 6) self-complementary dinucleotide circular codes:

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

We note that the directed graphs associated with all these 8 codes are isomorphic to the directed graph depicted in Fig. 9. In particular, $X_2$ yields a total order $< X$ on $B$ such that $X < Y$ implies that $c(Y) < c(X)$. Let us write $B = \{N_1, N_2, N_3, N_4\}$ with $N_i < N_j$ if $i < j$ (in particular, $c(N_1) = N_0$). We set $D := \{N_1, N_2\}$ and $F := c(D)$. Notice that $D < F$, that is, $\alpha < \beta$ for each $(\alpha, \beta) \in D \times F$, and hence $D \cap F = D$ 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 $X_2 \cap D \cup F X_2$ for $w \in X \cap X_2$, and thus $c(w) \in X \cap X_2 \cap D$ and thus $c(X)$ would contain a directed path (of length 3) between two vertices labelled by elements of $X_3$.

We shall prove that there are exactly four valid ways to extend $X_3$, which in total will give the 32 different codes $X$. Recall that $X$ has to contain exactly one element in each conjugacy class. 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 $c(X)$ has no directed cycle: if a dinucleotide $w \in B^2$ has an in-neighbour $N_i$ and an out-neighbour $N_j$ then $N_i < N_j$. It is obviously a necessary condition, for otherwise $c(X)$ would contain the directed cycle $N_j \to w \to N_i \to N_j$ and is also sufficient as one readily checks. This property also helps checking if $c(X)$ has no directed path between two dinucleotides in $X_2$.

Let $A$ be the set of self-complementary dinucleotides in $B^2$, i.e. $A = \{AT, TA, CG, GC\}$. The self-complementarity of $X$ implies that it contains no element in $A D \cup F A$, because $c(X)$ contains no directed cycle of length 3.

The words in DBF 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

$$N_1 N_1 N_2, N_1 N_2 N_3, N_1 N_2 N_4, N_1 N_3 N_4, N_1 N_4 N_2, N_2 N_1 N_3, N_2 N_1 N_4,$$
$$N_2 N_3 N_1, N_2 N_4 N_1, N_3 N_2 N_4.$$

In the conjugacy class of $N_1 N_4 N_3$, the code $X$ must contain $N_1 N_4 N_3$ as otherwise it would contain $N_1 N_2 N_4$ thus creating the cycle $N_2 \rightarrow N_3 \rightarrow N_1 \rightarrow N_2$. This implies that $X$ contains $N_1 N_2 N_4$. Similarly, $X$ must contain $N_1 N_3 N_1$. This implies that $N_1 N_4 N_3 \notin X$ and hence $N_1 N_3 N_1 \in X$. It thus remains 2 conjugacy classes in DBF for which we have two choices, namely that of $N_1 N_3 N_1$ and that of $N_2 N_1 N_4$.

There are 4 classes that have not been mentioned yet: those of elements in $D^3 \cup F^2$. Similarly as above, those 4 classes are composed of 2 pairs of conjugacy classes, and only 2 elements in each class can be valid choices. In total, the 6 remaining classes are that of $N_1 N_2 N_3$, of $N_1 N_4 N_2$, of $N_2 N_1 N_3$ 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 $N_2 N_1 N_3$ and $N_2 N_1 N_4$ does not belong to $X$. Similarly, one of $N_1 N_1 N_4$ and $N_1 N_1 N_2$ does not belong to $X$. Using that $X$ is self-complementary, we deduce that only four choices can be valid:

$$N_1 N_1 N_2, N_1 N_2 N_4, N_1 N_3 N_3, N_1 N_4 N_2, N_2 N_1 N_3, N_2 N_1 N_4,$$
$$N_3 N_1 N_2, N_4 N_2 N_4, N_4 N_3 N_2, N_4 N_4 N_2, N_2 N_1 N_3, N_2 N_1 N_4,$$
$$N_3 N_1 N_2, N_4 N_2 N_4, N_4 N_3 N_2, N_4 N_4 N_2, N_2 N_1 N_3, N_2 N_1 N_4.$$

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

A.4. Proof of Proposition 7

Proof. There are 4 maximum self-complementary trinucleotide comma-free codes of 16 trinucleotides (Tables 3a and 3b in [39]). There are 4 maximum dinucleotide comma-free codes of 5 dinucleotides [40]:

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

All of them consist of 2 complementary pairs of dinucleotides and 1 self-complementary dinucleotide, e.g. $\{AC, GT, AG, CT, AT\} = \{AC, c(AC), AG, c(AG), AT = c(AT)\}$. 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 dinucleotide 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 $N_1 N_2 = c(N_1 N_2)$, 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 dinucleotide 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 altogether four possibilities:

\[ X^1 = \{ AC, GT, AG, CT \}, \quad X^2 = \{ AC, GT, GA, TC \}, \]
\[ X^3 = \{ AG, CT, CA, TG \}, \quad X^4 = \{ GA, TC, CA, TG \}. \]

We notice that the directed graphs associated with these 4 codes are all isomorphic to the graph depicted in Fig. 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 \( X \) contains mixed self-complementary dinucleotides. Indeed, suppose that \( x \) is a trinucleotide code containing the self-complementary dinucleotide \( x_1 \). Then, because \( x_1 \) is self-complementary, it must be contained in \( x \). Therefore, we observe that the trinucleotide code \( X \) contains \( x \). Since \( x \) is a trinucleotide code, we can neither intersect the set \( X \cup T \).

Specifically, we prove that if \( X \) contains \( x \), then \( X \) contains \( x_1 \). Hence, we obtain the following trinucleotide codes in Case 2. For instance, for \( X^1 \), one has \( N_1 = A, N_2 = C, N_3 = G \) and \( N_4 = T \).

\[ X^1 = \{ AG, CT, AT, CG \}, \quad X^2 = \{ GA, TC, TA, GC \}. \]

Let us point out that, similarly as in Case 1, the directed graphs associated with these 4 codes are isomorphic, this time to the directed graph depicted in Fig. 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 \( \pi_1(X^1) \cap \pi_2(X^1) = \emptyset \). Furthermore we observe that the trinucleotide code, being self-complementarity, must be contained in \( X^1 \). Indeed, suppose that the trinucleotide code contains \( x \). Then, because \( x \) is self-complementary, it must be contained in \( X^1 \). So, \( X \) contains \( x \). We infer that \( X = \pi_1(X^1) \). Further we observe that the trinucleotide code \( X \) contains \( x \). Hence, \( X \) contains \( x \), and hence also that of its complement \( X \).

\[ \pi_1(X) = \emptyset \quad \pi_1(X) = c(\pi_1(X)) \]

[29]. By mixing the dinucleotide and the trinucleotide codes, we obtain a mixed self-complementary comma-free code since \( \pi_1(X) \cap \pi_2(X) = \emptyset = \pi_1(X) \cap \pi_2(X) \), hence no directed path in the directed graph associated with \( X \) can be extended using an edge from the directed graph associated with \( X \). There is therefore exactly one way to extend \( X \) 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. □

A.5. List of the 32 maximum self-complementary mixed circular codes of dinucleotides and trinucleotides with cardinality 26

\[ \{ AC, GT, AG, CT, AT, CG, AAC, GTT, AAG, CTT, AAT, ATT, CAC, GTG, ACG, CQT, ACT, AGT, CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG \}, \]
\[ \{ AG, CT, AT, CG, AAC, GTT, AAG, CTT, AAT, ATT, CAC, GTG, ACG, CQT, ACT, AGT, CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG \}, \]
\[ \{ AC, GT, AG, CT, AT, CG, AAC, GTT, AAG, CTT, AAT, ATT, CAC, GTG, ACG, CQT, ACT, AGT, CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG \}, \]
\[ \{ AC, GT, AG, CT, AT, CG, AAC, GTT, AAG, CTT, AAT, ATT, CAC, GTG, ACG, CQT, ACT, AGT, CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG \}, \]
\[ \{ AC, GT, AG, CT, AT, CG, AAC, GTT, AAG, CTT, AAT, ATT, CAC, GTG, ACG, CQT, ACT, AGT, CAG, CTG, AGG, CCT, ATG, CAT, CCG, CGG \}, \]
\[ \text{AC, GT, AG, CT, AT, GC, AAC, GTT, AAG, CTT, AAT, AAT, ACC, GGT, GAC, ACT, AGT, AGC, GCT, ATC, GAT, GCC, GCC, CTC, GAG, ACC, GTG, ATC, GTC, ACT, AGT, AGC, GCT, GTC, GAT, GCC, GCC} \]

[AC, GT, AG, CT, AT, GC, AAC, GTT, AAG, CTT, AAT, AAT, ACC, GGT, GAC, ACT, AGT, AGC, GCT, ATC, GAT, GCC, GCC, CTC, GAG, ACC, GTG, ATC, GTC, ACT, AGT, AGC, GCT, GTC, GAT, GCC, GCC]

\[ \text{ACC, GTG, ATC, GAC, GCC, GCC, CTC, GAG} \]

[AC, GT, AG, CT, AT, GC, AAC, GTT, AAG, CTT, AAT, AAT, ACC, GGT, GAC, ACT, AGT, AGC, GCT, ATC, GAT, GCC, GCC, CTC, GAG, ACC, GTG, ATC, GTC, ACT, AGT, AGC, GCT, GTC, GAT, GCC, GCC]

\[ \text{AGC, GCT, ATC, GAT, GCC, GCC} \]

[AC, GT, AG, CT, AT, GC, AAC, GTT, AAG, CTT, AAT, AAT, ACC, GGT, GAC, ACT, AGT, AGC, GCT, ATC, GAT, GCC, GCC, CTC, GAG, ACC, GTG, ATC, GTC, ACT, AGT, AGC, GCT, GTC, GAT, GCC, GCC]

\[ \text{ACC, GTG, ATC, GAC, GCC, GCC, CTC, GAG} \]

[AC, GT, AG, CT, AT, GC, AAC, GTT, AAG, CTT, AAT, AAT, ACC, GGT, GAC, ACT, AGT, AGC, GCT, ATC, GAT, GCC, GCC, CTC, GAG, ACC, GTG, ATC, GTC, ACT, AGT, AGC, GCT, GTC, GAT, GCC, GCC]

\[ \text{AGC, GCT, ATC, GAT, GCC, GCC} \]

[AC, GT, AG, CT, AT, GC, AAC, GTT, AAG, CTT, AAT, AAT, ACC, GGT, GAC, ACT, AGT, AGC, GCT, ATC, GAT, GCC, GCC, CTC, GAG, ACC, GTG, ATC, GTC, ACT, AGT, AGC, GCT, GTC, GAT, GCC, GCC]

\[ \text{ACC, GTG, ATC, GAC, GCC, GCC, CTC, GAG} \]

[AC, GT, AG, CT, AT, GC, AAC, GTT, AAG, CTT, AAT, AAT, ACC, GGT, GAC, ACT, AGT, AGC, GCT, ATC, GAT, GCC, GCC, CTC, GAG, ACC, GTG, ATC, GTC, ACT, AGT, AGC, GCT, GTC, GAT, GCC, GCC]

\[ \text{AGC, GCT, ATC, GAT, GCC, GCC} \]

\[ \text{CA, TG, AG, CT, TA, CG, CAA, TTG, GAA, TTC, TAA, TTA, CCA, TGG, CGA, TCG, CT, TAA, TTA, CCA, TGG, CGA, TCG, CT, TAA, TTA, CCA, TGG, CGA, TCG, CT, TAA, TTA, CCA, TGG, CGA, TCG, CT, TAA} \]

\[ \text{CAG, CTG, GGA, TCC, TCA, TGA, CCG, CCG} \]

\[ \text{CA, TG, AG, CT, TA, CG, CAA, TTG, GAA, TTC, TAA, TTA, CCA, TGG, CGA, TCG, CT, TAA, TTA, CCA, TGG, CGA, TCG, CT, TAA} \]

\[ \text{CAG, CTG, GGA, TCC, TCA, TGA, CCG, CCG} \]

\[ \text{CA, TG, AG, CT, TA, CG, CAA, TTG, GAA, TTC, TAA, TTA, CCA, TGG, CGA, TCG, CT, TAA} \]

\[ \text{CAG, CTG, GGA, TCC, TCA, TGA, CCG, CCG} \]

\[ \text{CA, TG, AG, CT, TA, CG, CAA, TTG, GAA, TTC, TAA, TTA, CCA, TGG, CGA, TCG, CT, TAA} \]

\[ \text{CAG, CTG, GGA, TCC, TCA, TGA, CCG, CCG} \]

References


