Trinucleotide $k$-circular codes I: Theory

Christian J. Michel*, Baptiste Mouillon, Jean-Sébastien Sereni

Theoretical Bioinformatics, ICube, C.N.R.S., University of Strasbourg, 300 Boulevard Sébastien Brant, 67400 Illkirch, France

ARTICLE INFO

Keywords:
- $k$-circular code
- Circular code
- Code
- Algorithm
- Growth function
- Graph

ABSTRACT

A code $X$ is $(\geq k)$-circular if every concatenation of words from $X$ that admits, when read on a circle, more than one partition into words from $X$, must contain at least $k + 1$ words. In other words, the reading frame retrieval is guaranteed for any concatenation of up to $k$ words from $X$. A code that is $(\geq k)$-circular for all integers $k$ is said to be circular. Any code is $(\geq 0)$-circular and it turns out that a code of trinucleotides is circular as soon as it is $(\geq 4)$-circular. A code is $k$-circular if it is $(\geq k)$-circular and not $(\geq k + 1)$-circular. Due to the explosive combinatorics of trinucleotide $k$-circular codes, we developed three classes of algorithms based on: (i) the smallest directed cycles (directed girth) in graphs; (ii) the eigenvalues of matrices; and (iii) the files that incrementally save partial results. These different approaches also allow us to verify the computational results obtained. We determine here the growth functions of trinucleotide $k$-circular codes, $k$ varying between 0 and 4, in the general case and in various particular cases: minimum, minimal, maximum, self-complementary $k$, $(k, k, k)$- and self-complementary $(k, k, k)$-circular.

1. Introduction

The concept of $k$-circular code was recently introduced (Fimmel et al., 2020). It is less restrictive than the circular code concept. Indeed, a circular code retrieves the reading frame for any concatenation of words of the code written on a circle. A code is $(\geq k)$-circular if any concatenation of words of the code written on a circle does not retrieve the reading frame contains at least $k + 1$ words, and it is $k$-circular if in addition some concatenation of $k + 1$ words of the code written on a circle admits several decompositions into words of the code. It follows that a $k$-circular code cannot be $(\geq k + 1)$-circular but must be $(\geq j)$-circular for all $j \leq k$. A code is circular if it is $(\geq k)$-circular for any non-negative integer $k$. It was proved that $k$ is bounded (Fimmel et al., 2020), in the sense that the number of possible values $k$ for which there exists a $k$-circular code is bounded in terms of the length of the words in the code and the size of the alphabet used.

On the genetic alphabet $\mathcal{B}$, there are $2^{64} \approx 10^{19}$ trinucleotide codes (including the empty set) ranging from the 64 trinucleotides of cardinality 1 to the genetic code of cardinality 64. The theory of trinucleotide $k$-circular codes allows us to describe any trinucleotide code of any cardinality among these $2^{64}$ ones according to their circularity property, i.e. their property of reading frame retrieval. Indeed, three classes can be defined as follows:

- trinucleotide codes with no circularity: no sequence generated by such a trinucleotide code can retrieve the reading frame;
- trinucleotide codes with partial circularity: some sequences generated by such a trinucleotide code cannot retrieve the reading frame, but some other sequences can retrieve the reading frame;
- trinucleotide codes with a complete circularity: any sequence generated by such a trinucleotide code can retrieve the reading frame.

In the companion article (Michel and Sereni, 2022), we detail this classification and analyse the ambiguous sequences for each class of trinucleotide $k$-circular codes with $k \in \{0, 1, 2, 3\}$. Furthermore, in order to consider the different cases of ambiguous sequences, we derive a new and general formula to measure the reading frame loss, whatever the trinucleotide $k$-circular code. This formula allows us to study the evolution of any trinucleotide $k$-circular code of (maximal) cardinality 20 to the genetic code, based on the reading frame retrieval property. This approach is applied to analyse the evolution of the trinucleotide circular code $X$ observed in genes to the genetic code. In addition, we analyse balanced $k$-trinucleotide codes, i.e. of cardinality 4, 8, 12, 16 and 20 having the same number of each nucleotide. We develop a general mathematical method to compute the number of balanced trinucleotide codes of each size, which also applies to self-complementary trinucleotide codes. We establish and quantify a relation between this balance property and the self-complementarity property. The combinatorial hierarchy of trinucleotide $k$-circular codes is updated with the growth function results. Finally, the numbers of

* Corresponding author.

E-mail addresses: c.michel@unistra.fr (C.J. Michel), bmouillon@clipper.ens.psl.eu (B. Mouillon), sereni@kam.mff.cuni.cz (J.-S. Sereni).

https://doi.org/10.1016/j.biosystems.2022.104667

Received 15 October 2021; Received in revised form 27 February 2022; Accepted 16 March 2022

Available online 26 March 2022

0303-2647/© 2022 Elsevier B.V. All rights reserved.
amino acids coded by the trinucleotide k-circular codes are given for the cases maximal, minimal, self-complementary k, (k, k, k)- and self-complementary (k, k, k)-circular.

In this theoretical article, we carry out here an extensive combinatorial study of the trinucleotide k-circular codes that constitute an important class of k-circular codes. After having recalled the necessary definitions and notations in Section 2 and the graph theorem associated to a k-circular code in Section 3, we present in Section 4 three algorithms that we developed to determine the growth functions of trinucleotide k-circular codes. In Section 5, several growth functions for different classes of trinucleotide k-circular codes are identified: general case, minimum, minimal, maximum, self-complementary, (k, k, k)-circular and self-complementary (k, k, k)-circular.

2. Definitions and notations

We work with the genetic alphabet \( B := \{A,C,G,T\} \), which has cardinality 4. An element \( N \) of \( B \) is called nucleotide. A word over the genetic alphabet is a sequence of nucleotides. A trinucleotide is a sequence of three nucleotides, that is, an element of \( B^3 \) using the standard word-theory notation. If \( w = N_1 \cdots N_s \) and \( w' = N'_1 \cdots N'_t \) are two sequences of nucleotides of respective lengths \( s \) and \( t \), then the concatenation \( w \cdot w' \) of \( w \) and \( w' \) is the sequence \( N_1 \cdots N_s N'_1 \cdots N'_t \) composed of \( s + t \) nucleotides.

Given a sequence \( w = N_1 N_2 \cdots N_s \in B^3 \) and an integer \( j \in \{0,1, \ldots, s-1\} \), the circular \( j \)-shift of \( w \) is the word \( N_{j+1} \cdots N_s N_1 \cdots N_j \). Note that the circular 0-shift of \( w \) is \( w \) itself. For example, if \( s = 3 \) and hence \( w = N_1 N_2 N_3 \) is a trinucleotide, then its circular 0-shift is \( w \) itself, while its circular 1-shift and its circular 2-shift are \( N_2 N_1 N_3 \) and \( N_3 N_1 N_2 \), respectively. A sequence \( w' \) of nucleotides is a circular shift of \( w \) if \( w' \) is the circular \( j \)-shift of \( w \) for some \( j \in \{0,1, \ldots, s-1\} \). The elements in \( B^3 \) can thus be partitioned into conjugacy classes, where the conjugacy class of a trinucleotide \( w \in B^3 \) is the set of all circular shifts of \( w \). For instance, the conjugacy class of the trinucleotide \( ACG \) is \( \{ACG, CGA, GAC\} \). Notice that the conjugacy class of a trinucleotide \( w \in B^3 \) has size 3 unless \( w \) is one of the four periodic trinucleotides, namely a trinucleotide in \( P := \{AAA, CCC, GGG, TTT\} \), in which case the conjugacy class has size 1.

Definition 2.1. Let \( B \) be the genetic alphabet.

- A trinucleotide code is a subset of \( B^3 \), that is, a set of trinucleotides.
- If \( X \) is a trinucleotide code and \( w \) is a sequence of nucleotides, then an \( X \)-decomposition of \( w \) is a tuple \( (x_1, \ldots, x_t) \in X^t \) of trinucleotides from \( X \) such that \( w = x_1 \cdots x_t \).

We now formally define the notion of circularity of a code.

Definition 2.2. Let \( X \subseteq B^3 \) be a trinucleotide code.

- Let \( m \) be a positive integer and let \( (x_1, \ldots, x_m) \in X^m \) be an \( m \)-tuple of trinucleotides from \( X \). A circular \( X \)-decomposition of the concatenation \( c := x_1 \cdots x_m \) is an \( X \)-decomposition of a circular shift of \( c \).
- Let \( k \) be a non-negative integer. The code \( X \) is \((\geq k)\)-circular if every concatenation of trinucleotides from \( X \) that admits more than one circular \( X \)-decomposition contains at least \( k+1 \) trinucleotides. In other words, \( X \) is \((\geq k)\)-circular if for every \( m \in \{1, \ldots, k\} \) and each \( m \)-tuple \( (x_1, \ldots, x_m) \) of trinucleotides from \( X \), the concatenation \( x_1 \cdots x_m \) admits a unique circular \( X \)-decomposition. The code \( X \) is \(k\)-circular if \( X \) is \((\geq k)\)-circular and not \((\geq k+1)\)-circular.
- The code \( X \) is circular if it is \((\geq k)\)-circular for all \( k \in \mathbb{N} \).

Remark 2.3. Every trinucleotide code \( X \) is trivially \((\geq 0)\)-circular.

Further, a trinucleotide code \( X \) is \((\geq 1)\)-circular if and only if \( X \) does not contain a word and one of its circular shifts. This exactly means that \( X \) contains at most one word from each conjugacy class and none of the periodic trinucleotides.

Here is an example to illustrate Definition 2.2.

Example 2.4. The trinucleotide code \( X = \{ATG, CAT, GCC, GGC\} \) is 1-circular. Indeed, the word \( w = CATGCC \), which is the concatenation of 2 trinucleotides from \( X \), namely CAT and GGC, admits a second circular \( X \)-decomposition: that of its circular 1-shift ATG GCC.

On the other hand, the code \( X \) is \((\geq 1)\)-circular since it contains no two trinucleotides in the same conjugacy class and no periodic trinucleotide.

Notions of maximality in a given set of codes are of general and biological interest, and have been studied, for instance, for the trinucleotide codes that are circular. We pursue this study in directions pointed at by the recent introduction of the notion of \( k \)-circularity of a code.

Definition 2.5. Let \( C \) be a family of trinucleotide codes. A trinucleotide code \( X \in C \) is maximum if every code in \( C \) has size at most \( |X| \). A trinucleotide code \( X \in C \) is minimal if it is inclusion-wise maximal, meaning that no code in \( C \) of size larger than \( |X| \) contains \( X \). Similarly, a trinucleotide code \( X \in C \) is minimum if every code in \( C \) has size at least \( |X| \). A trinucleotide code \( X \in C \) is minimal if it is inclusion-wise minimal, meaning that no code in \( C \) of size smaller than \( |X| \) is contained in \( X \).

The notions formalised in Definition 2.5 always refer to a given family of codes \( C \), which will always be clear from the context. We see also that a maximum code is necessarily maximal, but a maximal code need not be maximum — and similarly a minimum code is necessarily minimal but a minimal code need not be minimum.

Example 2.6. Suppose that \( C \) is the family composed of the three following codes:

\[ \{ACG\}, \{ACG, CGA\}, \{AGT, CGA, GTG\} \]

Then, in \( C \), the code \( \{AGT, CGA, GTG\} \) is maximum (and hence maximal), and it is minimal but not minimum, while the code \( \{ACG\} \) is minimum (and hence minimal). The code \( \{ACG, CGA\} \) is not minimal (and hence not minimum either), and it is maximal but not maximum.

We use graph theory to study the circularity of codes. To this end, several useful definitions and facts are gathered in the next section.

3. Graphs associated to trinucleotide codes

A new graph approach for studying circular codes (see Definition 3.1) has been recently developed (Fimmel et al., 2016). As we work only with trinucleotide codes, we restrict all definitions and results to our case of study. The interested reader can consult the article cited for the full results.

Let us define the graph\(^2\) associated to a code.

Definition 3.1. Let \( X \subseteq B^3 \) be a trinucleotide code. We define a graph \( G(X) = (V(X), E(X)) \) with set of vertices \( V(X) \) and set of arcs \( E(X) \) as follows:

- \( V(X) := \bigcup_{N_1 N_2 N_3 \in X} \{N_1, N_3, N_1 N_2, N_2 N_3\} \); and

\(^2\) Since all the graphs we consider are directed graphs, we simply write “graph” instead of “digraph”.

---

\(3\) We note here a discrepancy with the notation in some earlier works, where “k-circular” was used to mean what is here written \((\geq k)\)-circular; we do however need this refined notation in this work.
The graphs associated to the codes

\[ X_1 = \{ AGT, CTA \}, \quad X_2 = X_1 \cup \{ AT \}, \quad X_3 = X_1 \cup \{ ATG \} \text{ and } X_4 = X_1 \cup \{ ATG, TCA, TGG \}. \]

Illustrating also Theorem 3.2, we see that the graph \( G(X_i) \) has infinite directed girth (i.e. contains no directed cycle) and hence \( X_i \) is circular (which is the same as \( (\geq 4) \)-circular); the graph \( G(X_3) \) has directed girth 8 = 2 \( (3 + 1) \) and hence \( X_3 \) is 3-circular; the graph \( G(X_2) \) has directed girth 6 = 2 \( (2 + 1) \) and hence \( X_2 \) is 2-circular; and the graph \( G(X_1) \) has directed girth 4 = 2 \( (1 + 1) \) and hence \( X_1 \) is 1-circular.

The graph \( G(X) \) is the graph associated to \( X \).

Fig. 1 illustrates Definition 3.1.

The length of a directed cycle in a graph \( G \) is the number of arcs of the cycle. We note that every arc of \( G(X) \) joins a nucleotide and a dinucleotide; in particular the graph \( G(X) \) cannot contain a directed cycle of odd length. Directed cycles in the graph associated to a code play an important role, as witnessed by the following theorem (Fimmel et al., 2020, Theorem 3.3), the statement of which we specify to the case of trinucleotide codes.

**Theorem 3.2.** Let \( X \subseteq \mathcal{B}^3 \) be a trinucleotide code and \( k \) a non-negative integer. The code \( X \) is \( k \)-circular if and only if the minimum of the lengths of the directed cycles in \( G(X) \) is \( 2(k + 1) \), that is \( G(X) \) contains a directed cycle of length \( 2(k + 1) \) and no directed cycle of shorter length.

In view of Theorem 3.2, we are interested in the length of the shortest directed cycles in the graph associated to a code: this parameter is called the directed girth.

**Definition 3.3.** If \( G \) is a directed graph, then the directed girth of \( G \) is defined to be infinite if \( G \) contains no directed cycle, and the smallest number of arcs of \( G \) forming a directed cycle otherwise.

As pointed out above, if \( X \) is a trinucleotide code then every arc \( G(X) \) joins a nucleotide and a dinucleotide. Since \( \mathcal{B} \) contains exactly four nucleotides, it follows that a cycle in \( G(X) \), if any, must have length in \( \{ 2, 4, 6, 8 \} \). Therefore, Theorem 3.2 implies in particular that there is no trinucleotide \( k \)-circular code for \( k \geq 4 \); in other words, a trinucleotide \( (\geq 4) \)-circular code must be circular. Further, \( G(X) \) has a cycle of length 2 if and only if \( X \) contains two trinucleotides in a same conjugacy class, or one of the periodic trinucleotides. In this case, \( X \) is \( 0 \)-circular \( (2(k + 1) = 2 \) implies that \( k = 0 \)). The class of all trinucleotide \( (\geq 0) \)-circular codes is precisely the class of all trinucleotide codes.

On the other hand, there exist 3-circular trinucleotide codes. For instance the code

\[ X_5 = \{ AGC, ATT, CAA, CTG, GCC, GAT, TCA, TGG \} \]

is not \( (\geq 4) \)-circular since the sequence of 4 trinucleotides

\[ TCA \cdot AGC \cdot CTG \cdot GAT \quad \text{and} \quad CAA \cdot GCC \cdot TGG \cdot ATT, \]

but \( X \) is \( (3) \)-circular as one can check that no sequence of 3 trinucleotides admits two circular \( X \)-decompositions.

It follows that all non-empty trinucleotide codes over \( B \) can be naturally partitioned into 5 classes using the following definition.

**Definition 3.4.** We define the circularity \( \text{cir}(X) \) of a non-empty trinucleotide code \( X \) to be the largest integer \( k \in \{ 0, 1, 2, 3, 4 \} \) such that \( X \) is \( (\geq k) \)-circular.
For instance, the circularity of the code $X_4$ above is 3 (i.e. $\text{cir}(X_4) = 3$), while that of a trinucleotide circular code would be $4^3$. For sheer convenience (regarding the notion of minimality), we actually define the circularity of the empty code to be 5, that is, $\text{cir}(\emptyset) = 5$. In this way, the empty code forms a special class on its own, and we can focus on non-empty codes.

The notion of $k$-circularity of a code immediately makes interesting the notions of minimality formally introduced in Definition 2.5. These notions of minimality are not interesting for circular codes. Indeed, if $X$ is circular, then any subset of $X$ is also circular. This is no longer true for the circularity of a code. For instance, the trinucleotide code $\{\text{AAC, ACG, GTA, TAC}\}$ is 2-circular, while the code obtained by removing CGT, that is $\{\text{AAC, ACG, GTA, TAC}\}$, is a circular code, and hence has circularity 4. This remark, coupled to the graph representation, leads to an approach for determining the sequences that prevents the reading frame retrieval. This aspect is developed in the companion article (Michel and Sereni, 2022).

4. Development of algorithms to identify trinucleotide $k$-circular codes

Due to explosive combinatorics, we have developed specific algorithms for identifying trinucleotide $k$-circular codes. Algorithms presented in Sections 4.1 and 4.2 have been parallelised and implemented using the C language. The algorithm in Section 4.3 has been implemented using Ocaml.

4.1. Algorithms based on directed cycles in graphs

Theorem 3.2 represents a code as a (directed) graph and links the circularity of the code to the (directed) girth of the graph. Finding the length of a smallest directed cycle in a directed graph $G$ is not as straightforward as in the undirected case, and the worst-case time complexity is $O(n(n+e))$, where $n$ is the number of vertices of $G$ and $e$ the number of arcs (Itai and Rodeh, 1978). This follows from the fact that for an arbitrary vertex $v$ of $G$, the length of a shortest directed cycle containing $v$ can be computed in time $O(n+e)$ at worse.

As reported earlier, the graph $G(X)$ built from a trinucleotide code $X$ on the genetic alphabet $B$ must be bipartite — meaning that it contains no cycle of odd length — and it has a bi-partition with a part containing (at most) 4 vertices — those representing the four nucleotides in $B$. In particular, every directed cycle must contain at least one of these four vertices. In addition, the number of arcs is linear in the number of vertices, both being linear in the size of the code. It thus follows from the preceding paragraph that the length of a shortest directed cycle in $G(X)$ can be computed in time $O(n)$.

Let us give more details about the actual implementation we used. The graphs are built incrementally. We start from $G(X)$, of which we know the directed girth, and we check the effect, on the directed girth, of the addition of a particular word to $X$. Adding this word would add exactly two arcs, and thus we only need to check the possible directed cycles containing at least one of these two arcs.

Taking advantage of these facts, we designed an algorithm based on a parallelised stack. We fix an order on the trinucleotides of $B^3$, and each thread starts with a trinucleotide code of a small fixed size. The generic step is to check whether the addition of the next word $N_3 N_2 N_1$ to the current trinucleotide code $X$ creates a directed cycle of length less than the directed girth of $G(X)$, and if so then we want to know the length of a shortest such cycle. Such a directed cycle must contain the arc $N_1 \rightarrow N_2 N_3$ or the arc $N_1 N_2 \rightarrow N_3$, which we can clearly exploit to reduce the number of cases to check. Fig. 2 illustrates the situation described below.

Specifically, we proceed by computing four distances between pairs of nodes in $G(X)$. We first compute the distances from $N_3 N_2 N_1$ to $N_1$, and also from $N_1 N_3$ to $N_1 N_2$. The former lets us know the length of a shortest directed cycle containing $N_1 \rightarrow N_2 N_3$ and not $N_1 N_3 \rightarrow N_1$. The latter will be useful to know the length of a shortest directed cycle containing both arcs.

We next compute the distances from $N_3$ to $N_1 N_2$ and from $N_3$ to $N_1$, from which we can deduce the length of a shorter directed cycle containing $N_1 N_2 \rightarrow N_1$ (and possibly $N_1 \rightarrow N_2 N_3$, thanks to the distance from $N_1 N_2$ to $N_1 N_3$, which was computed before as mentioned in the previous paragraph).

This test allows us to know the effect of adding a word to $X$ without actually making its addition, which saves updating operations. The only updating operations are thus made when addition of the word on the stack is possible, and when we backtrack. In this latter case, a positive number of words have to be removed from $X$, which means removing the corresponding arcs from the graph $G(X)$ and recalling the value of the directed girth — which had been stored at the step of the thread that had computed it (Michel and Sereni, 2022).

Finally, we note that directed graphs are implemented using $n$ adjacency lists, where $n$ is the number of vertices and each adjacency list is represented by a linked list. When backtracking, words are removed in the reverse order from which they had been added, and thus we only have to remove the last element of some of the linked lists to update the graph. When adding a word to the code, we have to add an element at the end of some of the linked lists.

4.2. Algorithms based on adjacency matrices

To have an independent program checking the computer results described in Section 4.1, we designed a straightforward approach using matrices derived from $G(X)$ for a code $X$. An adequate choice of the matrix used allows for a smooth and elegant implementation.

Specifically, given a code $X$ and its associated graph $G(X)$, we build a zero–one square matrix $M_X$ where the lines, and the columns, are in bijection with the arcs of $G(X)$. The entry $M_X(i,j)$ is 1 if the arc corresponding to $j$ starts at the vertex where the arc corresponding to $i$ ends. This matrix $M_X$ can thus be seen as the adjacency matrix of the line graph $G(\overline{G}(X))$ of $G(X)$, defined to have one vertex for each arc of $G(X)$, and an arc from a vertex $u$ to a vertex $v$ if the corresponding arcs in $G(X)$, in the same order, form a directed path of length 2. An important observation is that the directed girth of $G(X)$ is the same as that of its line graph $G(\overline{G}(X))$.

There are then various options to deduce the sought directed girth from the matrix $M_X$. An elementary way to check for the directed girth is to sequentially compute increasing powers of $M^\chi$; the directed girth of $G(X)$ is the least positive power of $M^\chi$ containing a non-zero element on the main diagonal. Indeed, for every positive integer $\epsilon$, the entry $M^\chi(i,j)$ is exactly the number of directed walks in $G(X)$ of length precisely $\epsilon$. A directed closed walk must contain a directed cycle, and thus a directed closed walk of smallest length in $G(X)$ is indeed a shortest directed cycle of $G(X)$, where a directed closed walk in a graph $G$ without parallel arcs amounts to a sequence $v_0, v_1, v_2, \ldots, v_k$ of (non-necessarily distinct) vertices such that $v_{i+1} \rightarrow v_i$ is an arc in $G$ for each $i \in \{0, 1, \ldots, k-1\}$.

Alternatively, one can also proceed by computing the eigenvalues of $M_X$. Indeed, $G(X)$ is acyclic if and only if all eigenvalues of $M^\chi$ are zero, and if that is not the case then the length of a shortest directed cycle in $G(X)$ is equal to twice the least integer $\epsilon$ such that the sum of the $\epsilon$-th power of the eigenvalues of $M^\chi$ is non-zero.

The interest of a matrix representation of the line graph is an efficient and easy-to-implement way to add a new element to a code, or to remove the latest element added to a code. Indeed, to add a new word to a trinucleotide code $X$, it suffices to add two lines and two columns to $M_X$. To delete the latest element added to $X$, it suffices to delete the last two lines and the last two columns of $M_X$. The structure

\[\text{cir}(X_4) = 3\]
\[\text{cir}(\emptyset) = 5\]
of $M_X$ thus makes it particularly suited to a backtracking approach. Concretely, for computing codes of a given size $n$, the algorithm creates a single matrix of size $2n \times 2n$, but only consider the upper left part of adequate size at each given time; that is, the algorithm only deals with the first $2\ell$ lines and the first $2\ell$ columns when considering a code of size $\ell$ during its execution. Thus “deleting” the last two rows and the last two columns is actually just a single integer subtraction, as we simply decrease by 2 the integer bounding the number of rows (and columns) that the algorithm is allowed to consider. “Adding” two lines and two columns to $M_X$, where $|X| = \ell$, means increasing the boundary by 2, and updating the entries in these two lines and two columns.

With the elementary method using matrix powers, only integral values are used. On the other hand, the eigenvalues of the adjacency matrices can be complex numbers. The computations are thus made using floating-point numbers, but the problem is numerically stable, and as a matter of fact we never encountered a single run where the approximation created a discrepancy with the outcome of the other algorithms. The computation of the eigenvalues is performed using the library LAPACK, which uses the library BLAS.

4.3. Incremental algorithm

The strategy this time is to specify the algorithm from the structure of graphs representing trinucleotide $k$-circular codes for some $k \in \{1, 2, 3\}$. Indeed, the graph $G(X)$ associated to such a code $X$ must contain a directed cycle of length $2(k+1)$. This implies that the code has size at least $k+1$. The starting point is then all possible trinucleotide codes of size $k+1$ that give rise to a graph isomorphic to a directed cycle of length $2(k+1)$. All these possibilities give the number of trinucleotide $k$-circular codes of size exactly $k+1$, and are stored in a file. Once all trinucleotide codes with circularity $k$ and size $n$ have been generated and saved, the codes of size $n+1$ are generated by trying, for each saved code of size $n$, to add one extra trinucleotide to the code. The circularity of the new code is checked using the graph representation. If the circularity is still $k$, then we have found a trinucleotide $k$-circular code of size $n+1$. Such a code is saved and the process goes on.

We note that such a procedure might generate several times the same trinucleotide code, and thus once all codes of a given size and circularity have been generated, one needs to suppress those generated more than once. Another drawback is the time spent accessing files, which becomes enormous. (It would be useful here to design a specific lossless compression format, so as to minimise the time spent reading the file.) This method was implemented and executed for all sizes when $k \in \{2,3\}$ and most sizes (but not all) when $k = 1$. It confirmed the outputs obtained by the other methods described in Sections 4.1 and 4.2.

5. Results

5.1. A general formula to count the trinucleotide $0$-circular codes according to various partitions

We here establish a general formula to count the number of trinucleotide $0$-circular codes according to different partitions of the trinucleotides: for example, the partition can be given by the conjugacy classes, the self-complementarity or the mirror relation.

The following statement is straightforward.

Proposition 5.1. Let $E$ be a set of trinucleotides partitioned into $i$ classes each of size 3. For each positive integer $n$, the number $F_{s1}(n, i)$ of subsets $E'$ of $E$ of size $n$ such that $|E' \cap C| \leq 1$ for each class $C$ is

$$F_{s1}(n, i) = \binom{i}{n} \cdot 3^n.$$  (5.1)

Proposition 5.1 directly gives the number $F_{s2}(n, i)$ of subsets $E'$ of size $n$ such that $|E' \cap C| \geq 2$ for at least one class $C$, as indicated below. We however provide a second formula, which is combinatorially equivalent but is more amenable to further generalisations to other partition types (e.g. the mirror relation).

Proposition 5.2. Let $E$ be a set of trinucleotides partitioned into $i$ classes each of size 3. For each positive integer $n$, the number $F_{s2}(n, i)$ of subsets $E'$ of $E$ of size $n$ such that $|E' \cap C| \geq 2$ for at least one class $C$ is

$$F_{s2}(n, i) = \frac{3^i}{n} - \binom{i}{n} \cdot 3^n = \sum_{c = \lfloor n/3 \rfloor}^\min\{i, n-1\} \left(\binom{i}{c} \sum_{p=0}^{c-1} \binom{c-p}{3c-n-2p} \cdot 3^{2c-n-p}\right).$$  (5.2)

Proof. The number of subsets of $E$ of size $n$ is $\binom{3^i}{n}$, and hence (5.2) follows from Proposition 5.1.

To establish (5.3), consider a subset $E'$ of $E$ of size $n$. Let $c$ be the number of classes $C$ intersected by $E'$. For each $i \in \{1, 2, 3\}$, let $p_i$ be the number of classes $C$ such that $|E' \cap C| = i$. It follows that $p_1 + p_2 + p_3 = c$ and $p_1 + 2p_2 + 3p_3 = n$. In particular, $p_2 = 3c - n - 2p_1$ and $p_3 = c - p_1 - p_2 = n + p_1 - 2c$.

The number of possibilities for $E'$ can thus be written

$$\sum_{c=1}^i \binom{i}{c} \sum_{p_1=0}^c \binom{c-p_1}{3c-n-2p_1} \cdot 3^{2c-n-p_1}.$$  (5.4)

The subset $E'$ satisfies $|E' \cap C| \geq 2$ for at least one class $C$ if and only if $p_1 < c$. In addition, the range of $c$ can be reduced: indeed, $c$ cannot be less than $\lfloor n/3 \rfloor$ or greater than $n - 1$ (coherently, in such
Proposition 5.5. Proof of (5.7). A trinucleotide follows from Observations 5.3 and 5.4.

Proof of (5.6). We have

\[ N_{30}(n) = N_{30}(n) - N_{30}(n), \]  

and hence (5.6) follows from Observations 5.3 and 5.4.

Proposition 5.5. The number \( N_{30}(n) \) of trinucleotide 0-circular codes of size \( n \in \{1, \ldots, 64\} \) is

\[ N_{30}(n) = \left( \frac{64}{n} \right) - \left( \frac{20}{n} \right) \cdot 3^n. \]  

(5.6)

Proof. Proof of (5.6). We have \( N_{30}(n) = N_{30}(n) - N_{30}(n) \), and hence (5.6) follows from Observations 5.3 and 5.4.

Proof of (5.7). A trinucleotide 0-circular code must contain a trinucleotide in \( P \), or two trinucleotides belonging to the same conjugacy class (two trinucleotides that are circular shifts of one another). Thus, Proposition 5.1 applied to \( E = B^3 \setminus P \) with \( t = 20 \) implies that the number of trinucleotide 0-circular codes of size \( n \) that do not contain a periodic trinucleotide is \( \binom{64}{n} - F_{30}(n, 20) \). On the other hand, every trinucleotide code (of size \( n \)) containing (at least) one of the four periodic trinucleotides is necessarily 0-circular. Consequently, their number \( P(n) \) is \( \binom{64}{n} - 4 \cdot \binom{64}{n} + 6 \cdot \binom{64}{n} + 4 \cdot \binom{64}{n} \) and hence (5.7) follows. Proof of (5.8). This follows from (5.3) of Proposition 5.2 applied to \( E = B^3 \setminus P \) with \( t = 20 \) and the expression for \( P(n) \) written in the proof of (5.7).
Table 1
Growth function of the trinucleotide 𝑘-circular codes 𝑋 with cardinality |𝑋| between 1 and 20 and circularity 𝑘 between 0 and 4.

<table>
<thead>
<tr>
<th></th>
<th>𝑐𝑟(𝑋)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th></th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>306</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1704</td>
<td>2016</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>10884</td>
<td>348</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30432</td>
<td>41664</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>242931</td>
<td>10275</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>382164</td>
<td>635376</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>3857040</td>
<td>198084</td>
<td>984</td>
<td>0</td>
<td>192</td>
<td>3568212</td>
<td>7624512</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>46718328</td>
<td>2703072</td>
<td>42246</td>
<td>3192</td>
<td>0</td>
<td>250570512</td>
<td>74974368</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>451679952</td>
<td>2792916</td>
<td>766440</td>
<td>37104</td>
<td>0</td>
<td>141639780</td>
<td>621216192</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>3599676198</td>
<td>203850216</td>
<td>7772184</td>
<td>298668</td>
<td>0</td>
<td>614568102</td>
<td>4426165368</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>24234627832</td>
<td>1168509648</td>
<td>49134288</td>
<td>1570536</td>
<td>0</td>
<td>2086742208</td>
<td>27504855412</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>140563557772</td>
<td>5157137040</td>
<td>20575712</td>
<td>5298048</td>
<td>0</td>
<td>5542646244</td>
<td>15174321486</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>713842171704</td>
<td>1766070500</td>
<td>578248996</td>
<td>11553600</td>
<td>0</td>
<td>11503061124</td>
<td>743595781824</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>312726908286</td>
<td>47179726978</td>
<td>113358356</td>
<td>16476492</td>
<td>0</td>
<td>18651667124</td>
<td>328421703056</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>13013266895294</td>
<td>98620253796</td>
<td>1552755192</td>
<td>15424416</td>
<td>0</td>
<td>23606843556</td>
<td>13138658812224</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>47670312080676</td>
<td>161186829852</td>
<td>1491008256</td>
<td>9375408</td>
<td>0</td>
<td>2270634924</td>
<td>4785569988816</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>1592965344085992</td>
<td>204675268392</td>
<td>999089112</td>
<td>373552</td>
<td>0</td>
<td>16787523072</td>
<td>1595189999622720</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>488318377516335</td>
<td>198820855389</td>
<td>460697617</td>
<td>788820</td>
<td>0</td>
<td>97290022320</td>
<td>488526937097580</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>137922955497700</td>
<td>143368816140</td>
<td>142169112</td>
<td>83520</td>
<td>0</td>
<td>3708717048</td>
<td>1379370175265320</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>360151518125170</td>
<td>7256947818</td>
<td>27843072</td>
<td>2280</td>
<td>0</td>
<td>1012099740</td>
<td>360168979101880</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>871958488093380</td>
<td>25073997716</td>
<td>3104832</td>
<td>0</td>
<td>0</td>
<td>168726792</td>
<td>8719878125222720</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>19619722958676179</td>
<td>3473671209</td>
<td>148752</td>
<td>0</td>
<td>0</td>
<td>12964440</td>
<td>196197257852641120</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>34032813816304773</td>
<td>977188463215</td>
<td>66516906168</td>
<td>64485834</td>
<td>0</td>
<td>115606988558</td>
<td>3403931332523254</td>
</tr>
</tbody>
</table>

Remark 5.12. Every self-complementary trinucleotide code has even size, because no trinucleotide is its own complementary trinucleotide. This property does not hold anymore for codes with words of even length, e.g. dinucleotide codes and tetranucleotide codes.

The next observation follows directly by definition.

Observation 5.14. The number \( N_{SC}(m) \) of self-complementary trinucleotide \((≥0)\)-circular codes \( Y \) with size \( 2m \), for \( m ∈ \{1, \ldots, 32\} \), is

\[
N_{SC}(m) = \left( \frac{32}{m} \right)^3.
\]

A self-complementary trinucleotide code is \((≥1)\)-circular if and only if it contains no periodic trinucleotide \( P \) and at most one trinucleotide from each conjugacy class. The 60 trinucleotides not in \( P \) are partitioned into 20 conjugacy classes of size 3. Note that the complementary trinucleotide of a trinucleotide \( w \) cannot be a circular shift of \( w \). Further, the circular shifts of \( w \) are the self-complementary trinucleotides of the circular shifts of \( w \). More precisely, for \( j ∈ \{1, 2\} \), the circular \( j \)-shift of \( w \) is the complementary trinucleotide of the \( (3 − j) \)-circular shift of \( w \). As a result, the set \( \mathcal{C} \) of the 20 conjugacy classes can be partitioned into 2 subsets \( \mathcal{C}_1 \) and \( \mathcal{C}_2 \) of size 10, each subset being formed of the conjugacy classes of the self-complementary trinucleotides of the trinucleotides in the other subset. This means that any self-complementary trinucleotide code that does not contain a periodic trinucleotide is entirely determined by its intersections with the 10 conjugacy classes in \( \mathcal{C}_1 \). Thus, Proposition 5.1 with \( t = 10 \) leads to the following statement.

Observation 5.15. The number \( N_{SC}(m) \) of self-complementary trinucleotide \((≥1)\)-circular codes \( Y \) with size \( 2m \), for \( m ∈ \{1, \ldots, 10\} \), i.e. \( cir(Y) ∈ \{1, 2, 3, 4\} \), is

\[
N_{SC}(m) = F_{≤1}(m, 10) = \left( \frac{10}{m} \right)^3 m^3.
\]

The number \( N_{SC}(m) \) of self-complementary trinucleotide 0-circular codes of size \( 2m \in \{2, \ldots, 20\} \) can be expressed in various ways.

Proposition 5.16. The number \( N_{SC}(m) \) of self-complementary trinucleotide 0-circular codes of size \( 2m \), where \( m ∈ \{1, \ldots, 10\} \), is

\[
N_{SC}(m) = \left( \frac{32}{m} \right) - \left( \frac{10}{m} \right)^3 m^3.
\]
the trinucleotide code, to ensure circularity one conjugacy class should contain exactly two trinucleotides from codes we count should not contain an element in \( \{0\} \) avoid counting such cases, one can proceed as follows. First, each written in the proof of (5.10).

Proof of (5.11). This follows from (5.3) in Proposition 5.2 applied to the conjugacy classes in \( \mathcal{C} \).

By (5.5) and Observation 5.17 applied with Example 5.18.

Example 5.18. By (5.5) and Observation 5.17 applied with \( m = 10 \), we know that among the 64,453,191 self-complementary trinucleotide 0-circular codes of size 20, there are exactly 29,985,966 of them that contain no periodic trinucleotide: 29,985,966 = \( N_{0}^{0}(10) - P_{0}^{0}(10) = 64,453,191 - 34,467,225. \)

By Observation 5.17, the number of self-complementary trinucleotide 0-circular codes of size 20 that do not contain a whole conjugacy class is \( \hat{N}_{0}^{0}(10) = 21,581,316. \)

It follows that exactly 8,404,650 self-complementary trinucleotide 0-circular codes of size 20 contain a whole conjugacy class but no periodic trinucleotide: 8,404,650 = 29,985,966 – \( \hat{N}_{0}^{0}(10) = 29,985,966 - 21,581,316. \)

Table 4 gives the growth function of the self-complementary trinucleotide \( k \)-circular codes \( Y \) with even cardinality \(|Y|\) between 2 and 20 and circularity \( k \) between 0 and 4.

There are exactly 2 self-complementary trinucleotide 0-circular codes of size 2, which are thus minimum (see List 5.19). The situation is similar for self-complementary trinucleotide 1-circular codes (see List 5.20).

List 5.19 (The 2 Minimum Self-Complementary Trinucleotide 0-Circular Codes of Size 2).

\[ \{AAA, TTT\}, \{CCC, GGG\} \]

List 5.20 (The 2 Minimum Self-Complementary Trinucleotide 1-Circular Codes of Size 2).

\[ \{ATA, TAT\}, \{CGC, GCG\} \]

No self-complementary trinucleotide code of size less than 4 is 3-circular and there are exactly 4 self-complementary trinucleotide 3-circular codes of size 4, the list of which is found in List 5.21.

List 5.21 (The 4 Minimum Self-Complementary Trinucleotide 3-Circular Codes of Size 4).

\[ \{ACG, CGT, GTC, TAG\}, \{AGC, GCT, CTA, TAC\}, \{ATC, GAT, CGA, TGC\}, \{ATG, CAT, GCA, TGC\} \]

No self-complementary trinucleotide code of size less than 6 is 2-circular and there are exactly 8 self-complementary trinucleotides 2-circular codes of size 6, the list of which is found in List 5.24.

Observation 5.22. The 8 minimum self-complementary trinucleotide 2-circular codes of size 6, follow the structure

\[ \{a\bar{a}b, \bar{a}b\bar{a}, a\bar{b}a, \bar{b}a\bar{a}, \bar{a}\bar{b}a, \bar{a}a\bar{b}, \bar{b}a\bar{b}, \bar{b}\bar{a}a\} \]

for \( a \) and \( \beta \) two different and non-complementary nucleotides in \( B \). Fixing for instance \( a = A \) and \( \beta = C \), each of the 8 permutations that preserves the self-complementarity of the code can be applied, yielding all the 8 different minimum self-complementary trinucleotide 2-circular codes of size 6. More precisely, these are the permutations swapping either (possibly both) pairs of complementary nucleotides, the two possible permutations swapping \( A \) with one of \( C, G \), and \( T \) with the other one, the permutation \( (A, C, T, G) \) and its inverse for a total of seven codes in addition to the first one (see List 5.23). Consequently, the graphs associated to these 8 codes are all pairwise isomorphic: the unique graph obtained is depicted in Fig. 3.

List 5.23 (The 8 Permutations That Preserve the Self-Complementary Property of Trinucleotide Codes).

by removing both a trinucleotide and its complementary trinucleotide

cleotide

Definition 5.25. For each complementary trinucleotide code with a given circularity.

Table 5

| \( |Y| \) | \( \text{cir}(Y) \) | 1 | 2 | 3 | Total |
|-----|-----|-----|-----|-----|------|
| 2   | 2   | 0   | 0   | 2   |
| 4   | 14  | 0   | 4   | 18  |
| 6   | 64  | 8   | 8   | 80  |
| 8   | 117 | 56  | 56  | 229 |
| 10  | 0   | 64  | 0   | 64  |
| ≥12 | 0   | 0   | 0   | 0   |
| Total| 197 | 128 | 68  | 393 |

For the reader’s convenience, we explicitly list the 8 minimum self-complementary trinucleotide 2-circular codes of size 6.

List 5.24 (The 8 Minimum Self-Complementary Trinucleotide 2-Circular Codes of Size 6).

\{ [ACA, TGT, ATG, CAT, GAC, GTC], [ACA, TGT, CAG, CTG, GTA, TAC], [ACG, CGT, CAG, CTG, TCA, TGA], [ACT, AGT, CAT, GAC, GTC, GTG], [ACT, AGT, CGA, TCG, CTC, GAG], [AGA, TCT, ATC, GAT, CAT, CTG], [AGA, TCT, CTG, CAT, TAG, GAC, GTC], [AGC, GCT, CTC, GAG, TCA, TGA] \}.

5.6. Growth function of minimal self-complementary trinucleotide \( k \)-circular codes

We now turn to the notion of inclusion-wise minimality of a self-complementary trinucleotide code with a given circularity.

Definition 5.25. For each \( k \in \{1, 2, 3\} \), a self-complementary trinucleotide \( k \)-circular code \( Y \) is minimal if each code \( Y' \) obtained from \( Y \) by removing both a trinucleotide and its complementary trinucleotide is \( (\geq k + 1) \)-circular.

In other words, Definition 5.25 states that a self-complementary trinucleotide \( k \)-circular code \( Y \) is minimal if and only if for each word \( w \in Y \), the code \( Y \setminus \{w, \bar{w}\} \) is \( (\geq k + 1) \)-circular.

Table 5 presents the number of self-complementary trinucleotide \( k \)-circular codes that are minimal in the sense of Definition 5.25, for all relevant values of \( k \), i.e. \( k \in \{1, 2, 3\} \), and all possible code sizes.

Example 5.26. The trinucleotide code \( Y_1 := \{ ACA, TGT, CAG, CTG, GTA, TAC \} \) is self-complementary and also 3-circular since it contains exactly two directed cycles, both of length 8. Their intersection is \( A \rightarrow CA \rightarrow G \rightarrow TA \rightarrow C \rightarrow TG \rightarrow T \).

This intersection contains an arc from every pair of complementary trinucleotides of \( Y_1 \), which is enough to show that \( Y_1 \) is one of the 8 minimal self-complementary trinucleotide 3-circular codes, the list of which is found in List 5.27.

List 5.27 (The 8 Minimal Self-Complementary Trinucleotide 3-Circular Codes of Size 6).

\{ [ACA, TGT, ATG, CAT, GAC, GTC], [ACA, TGT, CAG, CTG, GTA, TAC], [ACG, CGT, CAG, CTG, TCA, TGA], [ACT, AGT, CAT, GAC, GTC, GTG], [ACT, AGT, CGA, TCG, CTC, GAG], [AGA, TCT, ATC, GAT, CAT, CTG], [AGA, TCT, CTG, CAT, TAG, GAC, GTC], [AGC, GCT, CTC, GAG, TCA, TGA] \}.

5.7. Growth function of trinucleotide \( (k, k, k) \)-circular codes

Using the notion of circular shifts, a trinucleotide code naturally gives rise to two other codes: the set of \( j \)-circular shifts of the trinucleotides in \( X \) for \( j \in \{1, 2\} \).

Definition 5.29. If \( X \subseteq B^3 \) is a trinucleotide code, then for \( j \in \{1, 2\} \) we define \( X_j \) to be the code composed of the \( j \)-circular shifts of all trinucleotides in \( X \), that is

\( X_1 := \{ N_1 N_2 N_1 : N_1 N_2 N_3 \in X \} \) \quad and \quad \( X_2 := \{ N_1 N_2 N_3 : N_1 N_3 N_2 \in X \} \).

Given a trinucleotide code of circularity \( k \), we are interested in the circularity of the two circular shifts of \( X \), namely \( \text{cir}(X_1) \) and \( \text{cir}(X_2) \).

Definition 5.30. We define the shifted circularity of a trinucleotide code \( X \) to be the triplet \( (\text{cir}(X_1), \text{cir}(X_1), \text{cir}(X_2)) \), and we write that \( X \) is \( (\text{cir}(X_1), \text{cir}(X_1), \text{cir}(X_2)) \)-circular.

Example 5.31. The \( C^3 \) self-complementary trinucleotide code \( X \) of maximal size 20 identified in genes (Arquès and Michel, 1996) has shifted circularity \((4, 4, 4)\), since it is \( C^3 \).

Example 5.32. Let \( X \) be the trinucleotide code \( \{ ATA, GTA, TAC, TAT \} \), which is 1-circular. Then \( X_1 = \{ TAA, TAG, ACT, ATT \} \) and \( X_2 = \{ ATT, AGT, CAT, TTA \} \). We see that both \( X_1 \) and \( X_2 \) are circular, and hence the shifted circularity of \( X \) is \((1, 4, 4)\). Note that the shifted circularity of \( X_1 \) is \((4, 4, 1)\) and that of \( X_2 \) is \((4, 1, 4)\).

Definition 5.30 broadly generalises the notion of \( C^3 \) for a trinucleotide circular code. We are particularly interested in the generalisation formed by trinucleotide \( (k, k, k) \)-circular codes for \( k \in \{1, \ldots, 4\} \).

Remark 5.33. A trinucleotide code \( X \) is 0-circular if and only if it contains a trinucleotide \( w \) and one of its circular shifts (which is \( w \) itself if \( w \) is one of the periodic trinucleotides). It follows that if \( \text{cir}(X) = 0 \), then \( \text{cir}(X_1) = 0 = \text{cir}(X_2) \), and therefore every trinucleotide 0-circular code has shifted circularity \((0, 0, 0)\). This fact of course does not generalise to larger values of \( \text{cir}(X) \).

Table 6 gives the growth function of trinucleotide \( (k, k, k) \)-circular codes \( X \) with cardinality \( |X| \) between 1 and 20 and \( k \) between 1 and 4.
The peculiarity of the case of trinucleotide (3, 3, 3)-circular codes begs for study. It is much striking that such codes exist only for size 10, as shown in Table 6. As it turns out, these 96 codes all have a very particular structure. Although we do not have at the moment a complete mathematical argument to establish that no other code is (3, 3, 3)-circular, we are currently working on establishing this fact.

An analysis of these 96 codes shows that they can be divided into four families of different codes: inside each family, any code is obtained from any other code by a suitable permutation of the nucleotides. In addition, inside each family no two nucleotides are “symmetric”, in the sense that all four nucleotides play different roles. Consequently, each family has size $4^4 = 24$. We may thus define each family by giving the general shape of the codes it contains, as follows.

**Observation 5.34.** If $X$ is one of the 96 trinucleotide (3, 3, 3)-circular codes of size 10, then there exists a bijection $x : \langle \alpha, \beta, \gamma \rangle \to B$ such that $X = x(F)$ where $F$ is one of the following four codes:

1. $(a\alpha, a\beta, a\gamma, \beta\beta, \gamma\gamma, \gamma\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma)$
2. $(a\alpha, a\beta, a\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma)$
3. $(a\alpha, a\beta, a\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma)$
4. $(a\alpha, a\beta, a\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma, \beta\beta, \gamma\gamma)$

Furthermore, the graph associated to any of these 96 codes is isomorphic to one of the graph depicted in Fig. 4.

The last part of Observation 5.34 is interesting: it tells us that, despite having non-equivalent codes among the 96 ones, they all share the same associated graph. It thus seems that the graph is able to capture intrinsic properties related to circularity while smoothing out some of the differences irrelevant to that matter.

### 5.8. Growth function of self-complementary trinucleotide $(k, k, k)$-circular codes

**Definition 5.35.** A trinucleotide code $Y$ is self-complementary $(k, k, k)$-circular if $Y$ is both self-complementary and $(k, k, k)$-circular.

We stress the important fact that, contrary to the general setting, Definition 5.35 is not symmetric: indeed, neither the 1-circular shift $Y_1$ nor the 2-circular shift $Y_2$ of a self-complementary code $Y$ is self-complementary itself (unless $Y \subseteq P$). Indeed, $Y_1$ and $Y_2$ are complementary of each other.

---

**Example 5.36.** Let $Y$ be the trinucleotide code $\{ATC, GAT, CCG, CCG, GCA, TGC\}$, which is self-complementary and 1-circular. First, the 1-circular shift $Y_1$ of $Y$ is the trinucleotide code $\{TCA, ATG, GCC, CCG, GAG, CTC\}$, which is also 1-circular but is not self-complementary. Second, the 2-circular shift $Y_2$ of $Y$ is $\{CAT, TGA, GCC, CCG, GCA, CTC\}$, which is also 1-circular (and not self-complementary). Hence $Y$ is a self-complementary trinucleotide (1, 1, 1)-circular code.

**Table 7**

Growth function of self-complementary trinucleotide $(k, k, k)$-circular codes $Y$ with even cardinality $|Y|$ between 2 and 20 and $k$ between 1 and 4.

| $|Y|$ | Shifted circularity of $Y$ |
|-----|--------------------------|
|     | $(1, 1, 1)$ | $(2, 2, 2)$ | $(3, 3, 3)$ | $(4, 4, 4)$ | Total |
| 2   | 0 | 0 | 0 | 0 | 28 | 28 |
| 4   | 0 | 0 | 0 | 0 | 330 | 330 |
| 6   | 0 | 0 | 0 | 0 | 2064 | 2132 |
| 8   | 0 | 0 | 0 | 0 | 7102 | 8866 |
| 10  | 0 | 0 | 0 | 0 | 13956 | 13956 |
| 12  | 0 | 0 | 0 | 0 | 16764 | 16764 |
| 14  | 0 | 0 | 0 | 0 | 12876 | 12876 |
| 16  | 0 | 0 | 0 | 0 | 62552 | 62552 |
| 18  | 0 | 0 | 0 | 0 | 17522 | 17522 |
| 20  | 0 | 0 | 0 | 0 | 2168 | 2168 |
| 20  | 0 | 0 | 0 | 0 | 2168 | 2168 |
| Total | 799144 | 340 | 0 | 61340 | 860824 |

---

6. Conclusion

We developed three classes of algorithms to compute the trinucleotide $k$-circular codes based on: (i) the smallest directed cycles (directed girth) in graphs; (ii) the eigenvalues of matrices; and (iii) the files that incrementally save partial results. They allowed us to determine quickly and safely the growth functions of the trinucleotide $k$-circular codes in the general case and in five important particular cases: minimum, minimal, self-complementary, $(k, k, k)$-circular and self-complementary $(k, k, k)$-circular. The general shape and the graph structure of some codes are described, in particular for the 96 trinucleotide (3, 3, 3)-circular codes of size 10.

In all their generality, the algorithms developed here allow us to study tetranucleotide codes (i.e. each word of the code is composed of 4 nucleotides). We already obtained partial results with the growth function of self-complementary tetranucleotide circular codes, most notably, the maximum number and its size. There are precisely 3,089,394,792 maximum self-complementary tetranucleotide circular codes of size 60.
Fig. 4. The two graphs generated by the 96 trinucleotide (3,3,3)-circulants of order 10. The white vertices correspond to those associated to nucleotides while the black vertices are those associated to dinucleotides.

Biological analyses inspired from this work are presented in the companion article (Michel and Sereni, 2022).

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.

Data availability

No data was used for the research described in the article.

References