Trinucleotide k-circular codes I: theory

Christian J Michel, Baptiste Mouillon, Jean-Sébastien Sereni

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
** Corresponding author*

Abstract. A code X is $(\geqslant k)$-circular if any concatenation of at most k words from X, when read on a circle, admits exactly one partition into words from X. A code that is $(\geqslant k)$-circular for all integers k is said to be circular. Any code is $(\geqslant 0)$-circular and it turns out that a code of trinucleotides is circular as soon as it is $(\geqslant 4)$-circular. A code is k-circular if it is $(\geqslant k)$-circular and not $(\geqslant k+1)$-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)-circular and self-complementary (k,k,k)-circular.

1. Introduction

The concept of k-circular code was recently introduced [**3**]. 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 $(\geqslant k)$-*circular* if a concatenation of at most k words of the code written on a circle retrieves the reading frame, 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 $(\geqslant k+1)$-circular but must be $(\geqslant j)$-circular for all $j \leq k$. A code is *circular* if it is $(\geqslant k)$-circular for any non-negative integer k. It was proved that k is bounded [**3**], 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.

E-mail address: c.michel@unistra.fr, bmouillon@clipper.ens.psl.eu, sereni@kam.mff.cuni.cz.
Date: September 1, 2021.
Key words and phrases. k-circular code; circular code; code; algorithm; growth function; graph.

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* $\mathcal{B} \coloneqq \{A, C, G, T\}$, which has cardinality 4. An element N of $\mathcal{B}$ 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 $\mathcal{B}^3$ 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 $w = N_1 N_2 \cdots N_s \in \mathcal{B}^s$ 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 $N_2 N_3 N_1$ 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 $\mathcal{B}^3$ can thus be partitioned into conjugacy classes, where the *conjugacy class* of a trinucleotide $w \in \mathcal{B}^3$ 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 $w \in \mathcal{B}^3$ has size 3 unless w is one of the four *periodic* trinucleotides, namely a trinucleotide in $\mathcal{P} \coloneqq \{AAA, CCC, GGG, TTT\}$, in which case the conjugacy class has size 1.

Definition 2.1. Let $\mathcal{B}$ be the genetic alphabet.

- A *trinucleotide code* is a subset of $\mathcal{B}^3$, 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 $w = x_1 \cdot x_2 \cdots x_t$.

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

Definition 2.2. Let $X \subseteq \mathcal{B}^3$ 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 $c \coloneqq x_1 \cdots x_m$ is an X-decomposition of a circular shift of c.

- Let k be a non-negative integer. The code X is $(\geqslant k)$-*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. Note that every trinucleotide code is trivially $(\geqslant 0)$-circular. The code X is k-*circular* if X is $(\geqslant k)$-circular and not $(\geqslant k+1)$-circular*.
- The code X is *circular* if it is $(\geqslant k)$-circular for all $k \in \mathbf{N}$.

Remark 2.3. Every trinucleotide code X is $(\geqslant 0)$-circular. Further, a trinucleotide code X is $(\geqslant 1)$-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 $w = CATGGC$, 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 $ATG \cdot GCC$. On the other hand, the code X is $(\geqslant 1)$-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 $\mathcal{C}$ be a family of trinucleotide codes. A trinucleotide code $X \in \mathcal{C}$ is *maximum* if every code in $\mathcal{C}$ has size at most $|X|$. A trinucleotide code $X \in \mathcal{C}$ *maximal* if it is inclusion-wise maximal, meaning that no code in $\mathcal{C}$ of size larger than $|X|$ contains X. Similarly, a trinucleotide code $X \in \mathcal{C}$ is *minimum* if every code in $\mathcal{C}$ has size at least $|X|$. A trinucleotide code $X \in \mathcal{C}$ *minimal* if it is inclusion-wise minimal, meaning that no code in $\mathcal{C}$ of size smaller than $|X|$ is contained in X.

The notions formalised in Definition 2.5 always refer to a given family of codes $\mathcal{C}$, 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 $\mathcal{C}$ is the family composed of the three following codes:

$$\{ACG\}, \{ACG, CGA\}, \{AGT, CGA, GTG\}.$$

Then, in $\mathcal{C}$, 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.

*We note here a discrepancy with the notation in some earlier works, where "k-circular" was used to mean what is here written $(\geqslant k)$-circular; we do however need this refined notation in this work.

3. Graphs associated to trinucleotide codes

A new graph approach for studying circular codes (see Definition 3.1) has been recently developed [**2**]. 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[†] associated to a code.

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

- $V(X) \coloneqq \bigcup_{N_1N_2N_3 \in X} \{N_1, N_3, N_1N_2, N_2N_3\}$; and
- $E(X) \coloneqq \{N_1 \to N_2N_3 \,:\, N_1N_2N_3 \in X\} \cup \{N_1N_2 \to N_3 \,:\, N_1N_2N_3 \in X\}$.

The graph $\mathcal{G}(X)$ is the graph *associated* to X.

Figure 1 illustrates Definition 3.1.
The *length* of a directed cycle in a graph $\mathcal{G}$ is the number of arcs of the cycle. We note that every arc of $\mathcal{G}(X)$ joins a nucleotide and a dinucleotide; in particular the graph $\mathcal{G}(X)$ 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 [**3**, 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 $\mathcal{G}(X)$ is $2(k+1)$, that is $\mathcal{G}(X)$ contains a directed cycle of length $2(k+1)$ and no directed cycle of shorter length.*

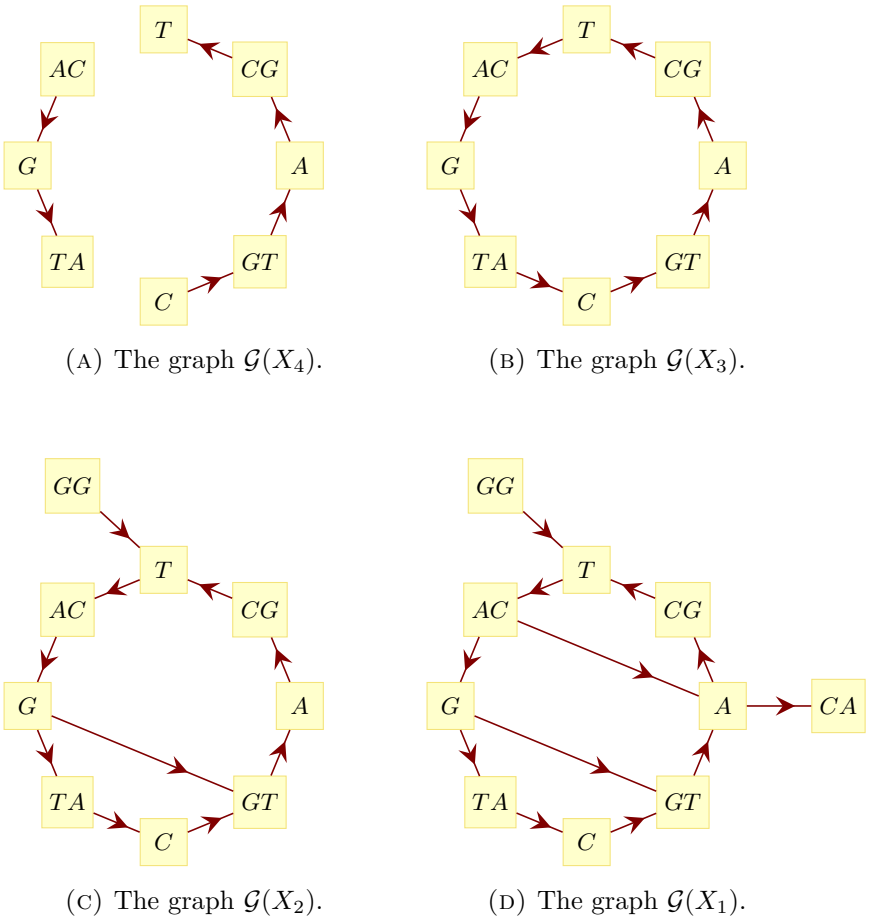
In view of Theorem 3.2, we are interested in the length of the shortest directed cycles in the graph associated to a code: this parameter is called the directed girth.

DEFINITION 3.3. If $\mathcal{G}$ is a directed graph, then the *directed girth* of $\mathcal{G}$ is defined to be infinite if $\mathcal{G}$ contains no directed cycle, and the smallest number of arcs of $\mathcal{G}$ forming a directed cycle otherwise.

As pointed out above, if X is a trinucleotide code then every arc of $\mathcal{G}(X)$ joins a nucleotide and a dinucleotide. Since $\mathcal{B}$ contains exactly four nucleotides, it follows that a cycle in $\mathcal{G}(X)$, if any, must be 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 ($\geqslant$4)-circular code must be circular. Further, $\mathcal{G}(X)$ 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 ($\geqslant$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

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

[†]Since all the graphs we consider are directed graphs, we simply write "graph" instead of "digraph".



(A) The graph $\mathcal{G}(X_4)$.

(B) The graph $\mathcal{G}(X_3)$.

(C) The graph $\mathcal{G}(X_2)$.

(D) The graph $\mathcal{G}(X_1)$.

FIGURE 1. The graphs associated to the codes $X_4 = \{ACG, CGT, GTA\}$, $X_3 = X_4 \cup \{TAC\}$, $X_2 = X_3 \cup \{GGT\}$ and $X_1 = X_2 \cup \{ACA\}$. Illustrating also Theorem 3.2, we see that the graph $\mathcal{G}(X_4)$ has infinite directed girth (i.e. contains no directed cycle) and hence X_4 is circular (which is the same as $(\geqslant 4)$-circular); the graph $\mathcal{G}(X_3)$ has directed girth $8 = 2 \cdot (3+1)$ and hence X_3 is 3-circular; the graph $\mathcal{G}(X_2)$ has directed girth $6 = 2 \cdot (2+1)$ and hence X_2 is 2-circular; and the graph $\mathcal{G}(X_1)$ has directed girth $4 = 2 \cdot (1+1)$ and hence X_1 is 1-circular.

is not $(\geqslant 4)$-circular since the sequence of 4 trinucleotides $TCAAGCCTGGAT$ admits two circular X-decompositions, namely

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

but X is $(\geqslant 3)$-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 $\mathcal{B}$ can be naturally partitioned into 5 classes using the following definition.

DEFINITION 3.4. We define the *circularity* $\mathrm{cir}(X)$ of a non-empty trinucleotide code X to be the largest integer $k \in \{0, 1, 2, 3, 4\}$ such that X is $(\geqslant k)$-circular.

For instance, the circularity of the code X_5 above is 3 (i.e. $\mathrm{cir}(X_5) = 3$), while that of a trinucleotide circular code would be 4.[‡] For sheer convenience (regarding the notion of minimality), we actually define the circularity of the empty code to be 5, that is, $\mathrm{cir}(\varnothing) = 5$. 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 $\{AAC, ACG, GTA, TAC, CGT\}$ is 2-circular, while the code obtained by removing CGT, that is $\{AAC, ACG, GTA, TAC\}$, 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 [**5**].

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 Subsections 4.1 and 4.2 have been parallelized and implemented using the `C` language. The algorithm in Subsection 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 $\mathcal{G}$ 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 $\mathcal{G}$ and e the number of arcs [**4**]. This follows from the fact that for an arbitrary vertex v of $\mathcal{G}$, the length of a shortest directed cycle containing v can be computed in time $O(n+e)$ at worse.

As reported earlier, the graph $\mathcal{G}(X)$ built from a trinucleotide code X on the genetic alphabet $\mathcal{B}$ 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 $\mathcal{B}$. 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 $\mathcal{G}(X)$ 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 $\mathcal{G}(X)$, 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.

[‡]We note here that we could have defined the circularity of a trinucleotide circular code to be infinite; however, since a trinucleotide code that is ($\geqslant$4)-circular must actually be circular, we chose to rather use this boundary of 4.

Taking advantage of these facts, we designed an algorithm based on a parallelized stack. We fix an order on the trinucleotides of $\mathcal{B}^3$, 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 $N_1N_2N_3$ to the current trinucleotide code X creates a directed cycle of length less than the directed girth of $\mathcal{G}(X)$, and if so then we want to know the length of a shortest such cycle. Such a directed cycle must contain the arc $N_1 \to N_2N_3$ or the arc $N_1N_2 \to N_3$, which we can clearly exploit to reduce the number of cases to check. Figure 2 illustrates the situation described below.

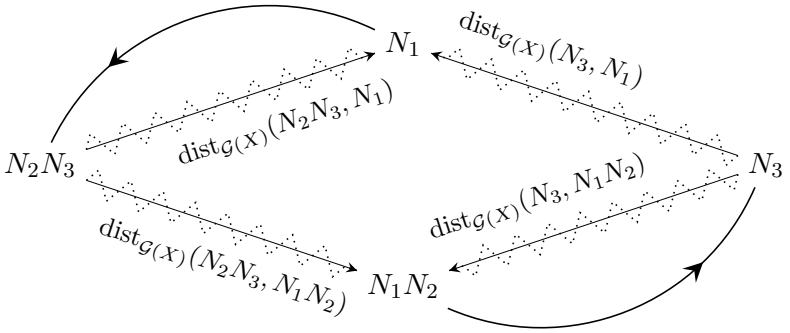


FIGURE 2. When adding the word $N_1N_2N_3$ to the trinucleotide code X, the associated graph is obtained from $\mathcal{G}(X)$ by adding the two arcs $N_1 \to N_2N_3$ and $N_1N_2 \to N_3$. An arrow in the middle of a dotted zigzag path represents a shortest direct path in $\mathcal{G}(X)$ from the source vertex to the destination vertex, if any. A directed cycle containing the arc $N_1 \to N_2N_3$ must contain either a directed path from N_2N_3 to N_1 in $\mathcal{G}(X)$ (whence the computation of the distance from N_2N_3 to N_1 in $\mathcal{G}(X)$), or it also contains the second added arc $N_1N_2 \to N_3$ and then also two directed paths from $\mathcal{G}(X)$: one from N_2N_3 to N_1N_2 and one from N_3 to N_1. It then remains to check for a directed cycle containing the arc $N_1N_2 \to N_3$ but not the arc $N_1 \to N_2N_3$: such a directed cycle must contain a directed path from N_3 to N_1N_2 in $\mathcal{G}(X)$.

Specifically, we proceed by computing four distances between pairs of nodes in $\mathcal{G}(X)$. We first compute the distances from N_2N_3 to N_1, and also from N_2N_3 to N_1N_2. The former lets us know the length of a shortest directed cycle containing $N_1 \to N_2N_3$ and not $N_1N_2 \to N_3$. The latter will be useful to know the length of a shortest directed cycle containing both new arcs.

We next compute the distances from N_3 to N_1N_2 and from N_3 to N_1, from which we can deduce the length of a shorter directed cycle containing $N_1N_2 \to N_3$ (and possibly $N_1 \to N_2N_3$, thanks to the distance from N_2N_3 to N_1N_2, 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 $\mathcal{G}(X)$ and recalling the value of the directed girth — which had been stored at the step of the thread where that particular trinucleotide code had been considered.

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 Subsection 4.1, we designed a straightforward approach using matrices derived from $\mathcal{G}(X)$ 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 $\mathcal{G}(X)$, we build a zero-one square matrix M_X where the lines, and the columns, are in bijection with the arcs of $\mathcal{G}(X)$. 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 $\widetilde{\mathcal{G}}(X)$ of $\mathcal{G}(X)$, defined to have one vertex for each arc of $\mathcal{G}(X)$, and an arc from a vertex u to a vertex v if the corresponding arcs in $\mathcal{G}(X)$, in the same order, form a directed path of length 2. An important observation is that the directed girth of $\mathcal{G}(X)$ is the same as that of its line graph $\widetilde{\mathcal{G}}(X)$.

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_X: the directed girth of $\mathcal{G}(X)$ is the least positive power of M_X containing a non-zero element on the main diagonal. Indeed, for every positive integer ℓ, the entry $M_X^\ell(i,j)$ is exactly the number of directed walks in $\widetilde{\mathcal{G}}(X)$ of length precisely ℓ. A directed closed walk must contain a directed cycle, and thus a directed closed walk of smallest length in $\mathcal{G}(X)$ is indeed a shortest directed cycle of $\mathcal{G}(X)$, where a *directed closed walk* in a graph $\mathcal{G}$ without parallel arcs amounts to a sequence $v_1, \ldots, v_s$ of (non-necessarily distinct) vertices such that $v_{i-1} \to v_i$ is an arc in $\mathcal{G}$ for each $i \in \{2, \ldots, s\}$.

One can avoid doing matrix multiplications by computing the eigenvalues of M_X. Indeed, $\mathcal{G}(X)$ is acyclic if and only if all eigenvalues of M_x are 0, and if that is not the case then the length of a shortest directed cycle in $\mathcal{G}(X)$ is equal to twice the least integer ℓ such that the sum of the ℓ-th power of the eigenvalues of M_X 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 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 2ℓ lines and the first 2ℓ columns when considering a code of size ℓ 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 $\mathcal{G}(X)$ 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 Subsections 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 t classes each of size 3. For each positive integer n, the number $F_{\leqslant 1}(n, t)$ of subsets E' of E of size n such that $|E' \cap C| \leq 1$*

for each class C is

$$F_{\leqslant 1}(n,t) = \binom{t}{n} \cdot 3^n. \tag{5.1}$$

Proposition 5.1 directly gives the number $F_{\geqslant 2}(n,t)$ 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 t classes each of size 3. For each positive integer n, the number $F_{\geqslant 2}(n,t)$ of subsets E' of E of size n such that $|E' \cap C| \geq 2$ for at least one class C is*

$$F_{\geqslant 2}(n,t) = \binom{3t}{n} - \binom{t}{n} \cdot 3^n \tag{5.2}$$

$$= \sum_{c=\lceil n/3 \rceil}^{\min\{t,n-1\}} \binom{t}{c} \sum_{p=0}^{c-1} \binom{c}{p} \binom{c-p}{3c-n-2p} \cdot 3^{3c-n-p}. \tag{5.3}$$

Proof. The number of subsets of E of size n is $\binom{3t}{n}$, 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

$$\sum_{c=1}^{t} \binom{t}{c} \sum_{p_1=0}^{c} \binom{c}{p_1} \binom{c-p_1}{3c-n-2p_1} \cdot 3^{3c-n-p_1}. \tag{5.4}$$

The subset E' satisfies $|E' \cap C| \geq 2$ for at least one class C if and only if $p < c$. In addition, the range of c can be reduced: indeed, c cannot be less than $\lceil n/3 \rceil$ or greater than $n-1$ (coherently, in such cases the rightmost binomial coefficient in (5.4) is 0). We thus deduce that

$$F_{\geqslant 2}(n) = \sum_{c=\lceil n/3 \rceil}^{\min\{t,n-1\}} \binom{t}{c} \sum_{p=0}^{c-1} \binom{c}{p} \binom{c-p}{3c-n-2p} \cdot 3^{3c-n-p}, \tag{5.5}$$

which concludes the proof. □

Propositions 5.1 and 5.2 will be applied to the general case in Subsection 5.2, and to the self-complementary case in Subsection 5.5.

5.2. Growth function of the circularity of trinucleotide codes. As reported earlier, the circularity of a trinucleotide code is between 0 and 4, and every code is ($\geqslant$0)-circular, that is, has circularity at least 0. The next observation follows directly by definition.

OBSERVATION 5.3. *The number $N_{\geqslant 0}(n)$ of trinucleotide ($\geqslant$0)-circular codes X with size n, for $n \in \{1, \ldots, 64\}$, is*

$$N_{\geqslant 0}(n) = \binom{64}{n}.$$

A trinucleotide code is ($\geqslant$1)-circular if and only if it contains no periodic trinucleotide and at most one trinucleotide from each conjugacy class. Proposition 5.1 applied to the partition of $E = \mathcal{B}^3 \setminus \mathcal{P}$ into the $t = 20$ conjugacy classes yields the following observation.

OBSERVATION 5.4. *The number $N_{\geqslant 1}(n)$ of trinucleotide ($\geqslant$1)-circular codes X with size n, for $n \in \{1, \ldots, 20\}$, i.e. $\mathrm{cir}(X) \in \{1, 2, 3, 4\}$, is*

$$N_{\geqslant 1}(n) = F_{\leqslant 1}(n, 20) = \binom{20}{n} \cdot 3^n.$$

The number $N_0(n)$ of trinucleotide 0-circular codes of size $n \in \{1, \ldots, 20\}$ can be expressed in various ways.

PROPOSITION 5.5. *The number $N_0(n)$ of trinucleotide 0-circular codes of size $n \in \{1, \ldots, 20\}$ is precisely*

$$N_0(n) = \binom{64}{n} - \binom{20}{n} \cdot 3^n \tag{5.6}$$

$$= \binom{60}{n-4} + 4\binom{60}{n-3} + 6\binom{60}{n-2} + 4\binom{60}{n-1} + \binom{60}{n} - \binom{20}{n} \cdot 3^n \tag{5.7}$$

$$= \binom{60}{n-4} + 4\binom{60}{n-3} + 6\binom{60}{n-2} + 4\binom{60}{n-1}$$

$$+ \sum_{c=\lceil n/3 \rceil}^{\min\{20, n-1\}} \binom{20}{c} \sum_{p=0}^{c-1} \binom{c}{p} \binom{c-p}{3c-n-2p} \cdot 3^{3c-n-p}. \tag{5.8}$$

PROOF. Proof of (5.6). We have $N_0(n) = N_{\geqslant 0}(n) - N_{\geqslant 1}(n)$, 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 $\mathcal{P}$, or two trinucleotides belonging to the same conjugation class (two trinucleotides that are circular shifts of one another). Thus, Proposition 5.1 applied to $E = \mathcal{B}^3 \setminus \mathcal{P}$ with $t = 20$ implies that the number of trinucleotide 0-circular codes of size n that do not contain a periodic trinucleotide is $\binom{60}{n} - F_{\leqslant 1}(n, 20)$. 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 $\binom{60}{n-4} + 4\binom{60}{n-3} + 6\binom{60}{n-2} + 4\binom{60}{n-1}$ and hence (5.7) follows.

Proof of (5.8). This follows from (5.3) of Proposition 5.2 applied to $E = \mathcal{B}^3 \setminus \mathcal{P}$ with $t = 20$ and the expression for $P(n)$ written in the proof of (5.7). □

We note that the size of a trinucleotide 0-circular code can be as large as 64. The code of size 64 is precisely the genetic code, which thus has circularity 0.

We present the number of all trinucleotide k-circular codes of any given size in Table 1, where we only omitted the trinucleotide codes of size larger than 20, which necessarily have circularity 0.
Observations 5.3 and 5.4 and Proposition 5.5 permit a partial verification of the numbers in Table 1, obtained by computer calculus. The total number of trinucleotide codes on the line corresponding to $|X| = n$ is $N_{\geqslant 0}(n)$, which is given by Observation 5.3. Next, the sum of the entries on this line and a column corresponding to $\mathrm{cir}(X) \in \{1, 2, 3, 4\}$ is $N_{\geqslant 1}(n)$, which is given by Observation 5.4. Last, the entry on the same line and the column corresponding to $\mathrm{cir}(X) = 0$ is equal to $N_0(n)$, which is given by Proposition 5.5.

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.

$\lvert X\rvert$	$\mathrm{cir}(X)$					
	0	1	2	3	4	**Total**
1	4	0	0	0	60	64
2	306	6	0	0	1704	2016
3	10884	348	0	0	30432	41664
4	242931	10275	0	6	382164	635376
5	3857040	198084	984	192	3568212	7624512
6	46718328	2703072	42264	3192	25507512	74974368
7	451679952	27092916	766440	37104	141639780	621216192
8	3599676198	203850216	7772184	298668	614568102	4426165368
9	24234627832	1168509648	49134288	1570536	2086742208	27540584512
10	140563557772	5157137040	204575712	5298048	5542646244	151473214816
11	713842171704	17660170500	578824896	11553600	11503061124	743595781824
12	3217269080286	47179720798	1133758356	16476492	18615667124	3284214703056
13	13013266893264	98620253796	1552755192	15424416	23403485556	13136858812224
14	47670312080376	161186859852	1491008256	9375408	22700634924	47855699958816
15	159296534408592	204675268392	999089112	3573552	16787523072	159518999862720
16	488318375716335	198820855389	460696716	788820	9279022320	488526937079580
17	1379222955497700	143368816140	142169112	83520	3708717048	1379370175283520
18	3601615181125170	72569947818	27843072	2280	1012099740	3601688791018080
19	8719854880393380	23073397716	3104832	0	168726792	8719878125622720
20	19619722295866719	3473671209	148752	0	12964440	19619725782651120
Total	34032813813604773	977188463215	6651690168	64485834	115606988558	34033913325232548

The notion of k-circularity allows for meaningful notions of minimality. The first one deals with the smallest possible size of a trinucleotide k-circular code, as presented next.

5.3. Sizes of minimum trinucleotide k-circular codes. While the minimum size of a trinucleotide code of circularity 0 or 4 is clearly 1, the situation is more complex when the circularity is 1, 2 or 3, as read from Tables 1 and 2.

TABLE 2. Minimum size of a trinucleotide code with circularity k between 0 and 4.

Circularity $\mathrm{cir}(X)$ of X	0	1	2	3	4
Minimum size of X	1	2	5	4	1
Number of codes of minimum size	4	6	984	6	60

OBSERVATION 5.6. *The* 6 *minimum trinucleotide* 1*-circular codes of size* 2*, follow the structure* $\{\alpha\beta\alpha, \beta\alpha\beta\}$ *for* α *and* β *two different nucleotides in* $\mathcal{B}$*. There are precisely* $\binom{4}{2} = 6$ *choices for the set* $\{\alpha, \beta\}$*. All these* 6 *codes are different but equivalent (in the sense that any of them is obtained from either of them by a suitable permutation of the nucleotides), and in particular all generate the same graph: a directed cycle of length* 4*.*

For the reader's convenience, we now list the 6 minimum trinucleotide 1-circular codes and the 6 minimum trinucleotide 3-circular codes.

LIST 5.7 (The 6 minimum trinucleotide 1-circular codes of size 2).

$$\{ACA, CAC\}, \{AGA, GAG\}, \{ATA, TAT\},$$
$$\{CGC, GCG\}, \{CTC, TCT\}, \{GTG, TGT\}.$$

LIST 5.8 (The 6 minimum trinucleotide 3-circular codes of size 4).

$$\{ACG, CGT, GTA, TAC\}, \{ACT, CTG, GAC, TGA\},$$
$$\{AGC, CTA, GCT, TAG\}, \{AGT, CAG, GTC, TCA\},$$
$$\{ATC, CGA, GAT, TCG\}, \{ATG, CAT, GCA, TGC\}.$$

5.4. Growth function of minimal trinucleotide k-circular codes. We now turn to the notion of inclusion-wise minimality of a trinucleotide code with a given circularity.

DEFINITION 5.9. For each $k \in \{1, 2, 3\}$, a trinucleotide k-circular code X is *minimal* if each code X' strictly contained in X is $(\geqslant k+1)$-circular.

In other words, Definition 5.9 states that a trinucleotide k-circular code X is minimal if and only if for each word $w \in X$, the code $X \setminus \{w\}$ is $(\geqslant k+1)$-circular.
Table 3 presents the number of trinucleotide k-circular codes that are minimal in the sense of Definition 5.9, for all relevant values of k, i.e. $k \in \{1, 2, 3\}$, and all possible code sizes. A striking fact occurs: one would have thought that for fixed k, the growth function seen as a function of the code size n, would first increase with n until a certain point, from which the function would be always 0. However, trinucleotide 3-circular codes show that this is not the case, since there are no minimal such codes of size 5, and yet there do exist minimal such codes of size 4 and of size 6.

TABLE 3. Growth function of the minimal trinucleotide k-circular codes X with cardinality $|X|$ between 1 and 20 and circularity k between 1 and 3.

$\lvert X\rvert$	$\mathrm{cir}(X)$			
	1	2	3	**Total**
1	0	0	0	0
2	6	0	0	6
3	24	0	0	24
4	840	0	6	846
5	0	984	0	984
6	0	6600	636	7236
7	0	0	2976	2976
8	0	0	4248	4248
≥ 9	0	0	0	0
Total	870	7584	7866	16320

EXAMPLE 5.10. The trinucleotide code $\{ACG, CAC, GCA\}$ is 1-circular since it contains the directed cycle $C \to AC \to G \to CA \to C$ and no shorter directed cycle. Yet this code contains none of the trinucleotide 1-circular of size less than 3 (see List 5.7), and therefore it is minimal.

5.5. Growth function of self-complementary trinucleotide k-circular codes. The biological notion of complementarity leads to study families of self-complementary trinucleotide codes.

DEFINITION 5.11. The *complementary nucleotide* $\overline{N}$ of a nucleotide $N \in \mathcal{B}$ is given by $\overline{A} \coloneqq T$, $\overline{T} \coloneqq A$, $\overline{C} \coloneqq G$ and $\overline{G} \coloneqq C$. A trinucleotide code $Y \subseteq \mathcal{B}^3$ is *self-complementary* if for every trinucleotide $N_1N_2N_3$ in Y, the *complementary trinucleotide* $\overline{N_1N_2N_3} \coloneqq \overline{N_3}\,\overline{N_2}\,\overline{N_1}$ belongs to Y, that is, if

$$Y = \overline{Y} \coloneqq \left\{ \overline{N_3}\,\overline{N_2}\,\overline{N_1} \ : \ N_1N_2N_3 \in Y \right\}.$$

REMARK 5.12. If $w \in \mathcal{B}^3$ is a trinucleotide, then the complementary trinucleotide of $\overline{w}$ is w itself — in other words, the complementary operation is an involution. The trinucleotide code $Y = \{ACT, AGT, CCG, CGG\}$ is self-complementary, as $\overline{ACT} = AGT$ and $\overline{CCG} = CGG$.

REMARK 5.13. 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 $N_{\geqslant 0}^{sc}(m)$ of self-complementary trinucleotide $(\geqslant 0)$-circular codes Y with size $2m$, for $m \in \{1, \ldots, 32\}$, is*

$$N_{\geqslant 0}^{sc}(m) = \binom{32}{m}.$$

A self-complementary trinucleotide is $(\geqslant 1)$-circular if and only if it contains no periodic trinucleotide $\mathcal{P}$ and at most one trinucleotide from each conjugacy class. The 60 trinucleotides not in $\mathcal{P}$ 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 $\overline{w}$ are the complementary trinucleotides of the circular shifts of w. More precisely, for $j \in \{1, 2\}$, the circular j-shift of $\overline{w}$ is the complementary of 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 $\mathscr{C}_1$ and $\mathscr{C}_2$ of size 10, each subset being formed of the conjugacy classes of the 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 $\mathscr{C}_1$. Thus, Proposition 5.1 with $t = 10$ leads to the following statement.

OBSERVATION 5.15. *The number $N_{\geqslant 1}^{sc}(m)$ of self-complementary trinucleotide $(\geqslant 1)$-circular codes Y with size $2m$, for $m \in \{1, \ldots, 10\}$, i.e. $\mathrm{cir}(Y) \in \{1, 2, 3, 4\}$, is*

$$N_{\geqslant 1}^{sc}(m) = F_{\leqslant 1}(m, 10) = \binom{10}{m} \cdot 3^m.$$

The number $N_0^{\mathrm{sc}}(m)$ 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 $N_0^{sc}(m)$ of self-complementary trinucleotide 0-circular codes of size $2m$, where $m \in \{1, \ldots, 10\}$, is*

(5.9)

$$N_0^{sc}(m) = \binom{32}{m} - \binom{10}{m} \cdot 3^m$$

$$= \binom{30}{m-2} + 2\binom{30}{m-1} + \binom{30}{m} - \binom{10}{m} \cdot 3^m \tag{5.10}$$

$$= \binom{31}{m-1} + \binom{30}{m-1} + \sum_{c=\lceil m/3 \rceil}^{\min\{10, m-1\}} \binom{10}{c} \sum_{p=0}^{c-1} \binom{c}{p} \binom{c-p}{3c-m-2p} \cdot 3^{3c-m-p}. \tag{5.11}$$

PROOF. Proof of (5.9). We have $N_0^{\mathrm{sc}}(m) = N_{\geqslant 0}^{\mathrm{sc}}(m) - N_{\geqslant 1}^{\mathrm{sc}}(m)$, and hence (5.9) follows from Observations 5.14 and 5.15.

Proof of (5.10). Proposition 5.1 applied to the conjugacy classes in $\mathcal{C}_1$ with $t = 10$ implies that the number of self-complementary trinucleotide 0-circular codes of size $2m$ that do not contain a periodic trinucleotide is $\binom{30}{m} - F_{\leqslant 1}(m, 10)$. In addition, the periodic trinucleotides contained in a self-complementary trinucleotide code are completely determined by its intersection with $\{AAA, CCC\}$. Self-complementary trinucleotide codes containing a periodic trinucleotide

(i.e. with a non-empty intersection with the set $\mathcal{P}$ of periodic trinucleotides) are necessarily 0-circular, and hence their number $P^{\mathrm{sc}}(m)$ is $\binom{30}{m-2} + 2\binom{30}{m-1}$. Thus, (5.10) follows.

Proof of (5.11). This follows from (5.3) in Proposition 5.2 applied to the conjugacy classes in $\mathcal{C}_1$ with $t = 10$ and the expression for $P^{\mathrm{sc}}(m)$ written in the proof of 5.10. □

Some of the codes counted contain a whole conjugacy class. To avoid counting such cases, one can proceed as follows. First, each periodic trinucleotide in $\mathcal{P}$ forms its whole conjugacy class, so the codes we count should not contain an element in $\mathcal{P}$. Second, at least one conjugacy class should contain exactly two trinucleotides from the trinucleotide code, to ensure circularity 0. We thus conclude the following.

Observation 5.17. *For each $m \in \{1, \ldots, 10\}$, the number $\widetilde{N}_0^{sc}(m)$ of self-complementary trinucleotide 0-circular codes of size $2m$ that do no contain a whole conjugacy class is*

$$\widetilde{N}_0^{sc}(m) = \sum_{d=1}^{\lfloor m/2 \rfloor} \binom{10}{d} \binom{10-d}{m-2d} \cdot 3^{m-d}.$$

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:

$$29,985,966 = N_0^{\mathrm{sc}}(10) - P^{\mathrm{sc}}(10) = 64,453,191 - 34,467,225.$$

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

$$\widetilde{N}_0^{\mathrm{sc}}(10) = 21,581,316.$$

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:

$$8,404,650 = 29,985,966 - \widetilde{N}_0^{\mathrm{sc}}(10) = 29,985,966 - 21,581,316.$$

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.
Observations 5.14 and 5.15 and Proposition 5.16 permit a partial verification of the numbers in Table 4, obtained by computer calculus. The total number of trinucleotide codes on the line corresponding to $|Y| = 2m$ is $N_{\geqslant 0}^{\mathrm{sc}}(m)$, which is given by Observation 5.14. Next, the sum of the entries on this line and a column corresponding to $\mathrm{cir}(Y) \in \{1, 2, 3, 4\}$ is $N_{\geqslant 1}^{\mathrm{sc}}(m)$, which is given by Observation 5.15. Last, the entry on the same line and the column corresponding to $\mathrm{cir}(Y) = 0$ is equal to $N_0^{\mathrm{sc}}(m)$, which is given by Proposition 5.16.
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).

$$\{AAA, TTT\}, \{CCC, GGG\}.$$

TABLE 4. 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.

$\lvert Y\rvert$	$\mathrm{cir}(Y)$					
	0	1	2	3	4	**Total**
2	2	2	0	0	28	32
4	91	67	0	4	334	496
6	1720	992	8	64	2176	4960
8	18950	8180	160	376	8294	35960
10	140140	40344	888	904	19100	201376
12	753102	123014	1844	968	27264	906192
14	3103416	235948	1704	464	24324	3365856
16	10223055	281145	780	80	13240	10518300
18	27851970	192622	176	0	4032	28048800
20	64453191	58505	16	0	528	64512240
Total	106545637	940819	5576	2860	99320	107594212

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 trinucleotides 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).

$$\{ACG, CGT, GTA, TAC\}, \{AGC, GCT, CTA, TAG\},$$
$$\{ATC, GAT, CGA, TCG\}, \{ATG, CAT, GCA, TGC\}.$$

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*

$$\{\alpha\alpha\beta, \bar{\beta}\bar{\alpha}\bar{\alpha}, \alpha\beta\bar{\beta}, \beta\bar{\beta}\bar{\alpha}, \bar{\alpha}\alpha\beta, \bar{\beta}\bar{\alpha}\alpha\}$$

for α *and* β *two different and non-complementary nucleotides in* $\mathcal{B}$*. Fixing for instance* $\alpha = A$ *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-complementarity trinucleotide* 2*-circular*

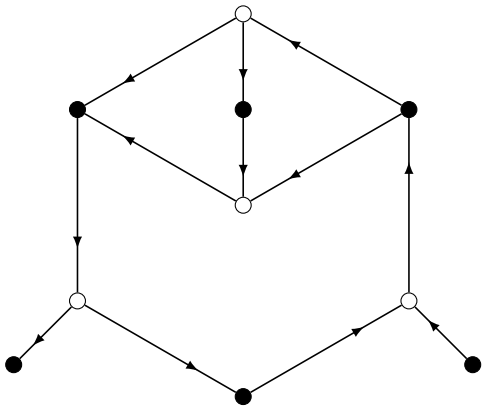


FIGURE 3. The unique graph generated by the 8 minimum self-complementary trinucleotide 2-circular codes of size 6. The white vertices correspond to those associated to nucleotides while the black vertices are those associated to dinucleotides.

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 Figure 3.*

LIST 5.23 (The 8 permutations that preserve the self-complementary property of trinucleotide codes).

$$\begin{array}{llll} (A, & C, & G, & T), \\ (A, & G, & C, & T), \\ (C, & A, & T, & G), \\ (C, & T, & A, & G), \end{array} \qquad \begin{array}{llll} (T, & C, & G, & A), \\ (T, & G, & C, & A), \\ (G, & T, & A, & C), \\ (G, & A, & T, & C). \end{array}$$

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).

$$\{AAC, GTT, ACG, CGT, GTA, TAC\}, \{AAG, CTT, AGC, GCT, CTA, TAG\},$$
$$\{ACC, GGT, ACG, CGT, GTA, TAC\}, \{AGC, GCT, AGG, CCT, CTA, TAG\},$$
$$\{ATC, GAT, CGA, TCG, GAA, TTC\}, \{ATC, GAT, CGA, TCG, GGA, TCC\},$$
$$\{ATG, CAT, CAA, TTG, GCA, TGC\}, \{ATG, CAT, CCA, TGG, GCA, TGC\}.$$

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 $(\geqslant k+1)$-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 $Y \setminus \{w, \overline{w}\}$ is $(\geqslant k+1)$-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.

TABLE 5. Growth function of the minimal self-complementary trinucleotide k-circular codes Y with even cardinality $|Y|$ between 1 and 20 and circularity k between 1 and 3.

$\lvert Y\rvert$	$\mathrm{cir}(Y)$			
	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

EXAMPLE 5.26. The trinucleotide code $Y_1 \coloneqq \{ACA, TGT, CAG, CTG, GTA, TAC\}$ is self-complementary and also 3-circular since it contains exactly two directed cycles, both of length 8. Their intersection is

$$A \to CA \to G \to TA \to C \to TG \to T.$$

This intersection contains an arc from every pair of complementary trinucleotides of Y_1, which is enough to prove 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).

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

LIST 5.28 (The 14 minimal self-complementary trinucleotide 1-circular codes of size 4).

$$\{AAT, ATT, CGA, TCG\}, \{AAT, ATT, GCA, TGC\}, \{ACA, TGT, CAC, GTG\},$$
$$\{ACG, CGT, GCA, TGC\}, \{ACG, CGT, TAA, TTA\}, \{AGA, TCT, CTC, GAG\},$$
$$\{AGC, GCT, CGA, TCG\}, \{AGC, GCT, TAA, TTA\}, \{ATC, GAT, CCG, CGG\},$$
$$\{ATC, GAT, CTA, TAG\}, \{ATG, CAT, GCC, GGC\}, \{ATG, CAT, GTA, TAC\},$$
$$\{CCG, CGG, GTA, TAC\}, \{CTA, TAG, GCC, GGC\}.$$

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 $X \subseteq \mathcal{B}^3$ 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

$$X_1 \coloneqq \{N_2N_3N_1 \,:\, N_1N_2N_3 \in X\}\,, \quad \text{and}$$
$$X_2 \coloneqq \{N_3N_1N_2 \,:\, N_1N_2N_3 \in X\}\,.$$

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

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

EXAMPLE 5.31. The C^3 self-complementary trinucleotide code (X) of maximal size 20 identified in genes [**1**] 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 $X_2 = \{AAT, AGT, CTA, TTA\}$. 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 $\mathrm{cir}(X) = 0$, then $\mathrm{cir}(X_1) = 0 = \mathrm{cir}(X_2)$, and therefore every trinucleotide 0-circular code has shifted circularity $(0,0,0)$. This fact of course does not generalise to larger values of $\mathrm{cir}(X)$.

Table 6 gives the growth function of trinucleotide (k,k,k)-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 $4! = 24$. We may thus define each family by giving the general shape of the codes it contains, as follows.

TABLE 6. Growth function of trinucleotide (k,k,k)-codes X with cardinality $|X|$ between 1 and 20 and k between 1 and 4.

$\lvert X\rvert$	shifted circularity of X				
	$(1,1,1)$	$(2,2,2)$	$(3,3,3)$	$(4,4,4)$	**Total**
1	0	0	0	60	60
2	0	0	0	1692	1692
3	0	0	0	29736	29736
4	0	0	0	362340	362340
5	288	0	0	3208140	3208428
6	24624	72	0	20979360	21004056
7	819696	2184	0	101278980	102100860
8	15046488	32472	0	358986546	374065506
9	173052684	293688	0	934952112	1108298484
10	1321102596	1403784	96	1810992816	3133499292
11	6905470284	3193416	0	2659948812	9568612512
12	25274438019	3529416	0	3016531848	28294499283
13	66114692304	2117352	0	2671142076	68787951732
14	125886576816	800640	0	1851870852	127739248308
15	176584791216	224832	0	998646600	177583662648
16	182565809382	55620	0	411632826	182977497828
17	136685642724	12312	0	125522712	136811177748
18	70713412164	1944	0	26719056	70740133164
19	22760177964	144	0	3548208	22763726316
20	3449390967	0	0	221544	3449612511
Total	818450448216	11667876	96	14996576316	833458692504

OBSERVATION 5.34. *If X is one of the* 96 *trinucleotide* $(3,3,3)$*-circular codes of size* 10*, then there exists a bijection $\pi\colon \{\alpha,\beta,\gamma,\delta\} \to \mathcal{B}$ such that $X = \pi(F)$ where F is one of the following four codes:*

(1) $\{\alpha\alpha\beta, \alpha\beta\gamma, \alpha\gamma\delta, \beta\delta\beta, \beta\delta\gamma, \gamma\alpha\gamma, \gamma\beta\beta, \gamma\gamma\delta, \delta\alpha\alpha, \delta\beta\alpha\}$*;*
(2) $\{\alpha\alpha\beta, \alpha\beta\delta, \alpha\gamma\gamma, \beta\alpha\gamma, \beta\beta\delta, \beta\gamma\beta, \gamma\delta\beta, \gamma\delta\gamma, \delta\alpha\alpha, \delta\gamma\alpha\}$*;*
(3) $\{\alpha\alpha\beta, \alpha\gamma\beta, \beta\delta\alpha, \beta\delta\delta, \gamma\alpha\delta, \gamma\beta\gamma, \gamma\gamma\alpha, \delta\alpha\alpha, \delta\beta\gamma, \delta\gamma\delta\}$*;*
(4) $\{\alpha\alpha\beta, \alpha\delta\beta, \beta\gamma\alpha, \beta\gamma\gamma, \gamma\alpha\gamma, \gamma\beta\delta, \gamma\delta\alpha, \delta\alpha\alpha, \delta\beta\delta, \delta\delta\gamma\}$*.*

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

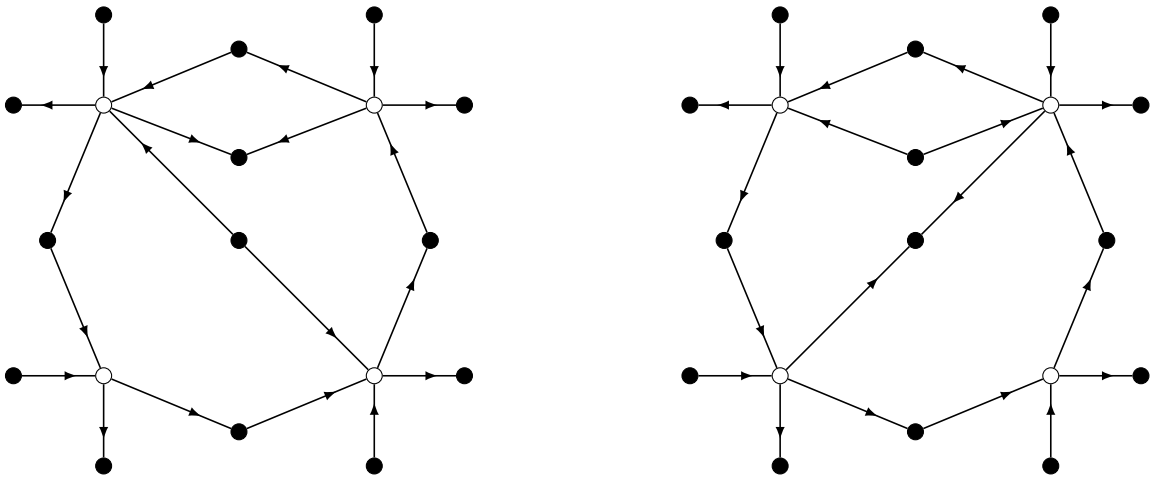


FIGURE 4. The two graphs generated by the 96 trinucleotide $(3,3,3)$-circular codes of size 10. The white vertices correspond to those associated to nucleotides while the black vertices are those associated to dinucleotides.

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 $Y \subseteq \mathcal{P}$). Indeed, Y_1 and Y_2 are complementary of each other.

EXAMPLE 5.36. Let Y be the trinucleotide code $\{ATC, GAT, CCG, CGG, GCA, TGC\}$, which is self-complementary and 1-circular. First, the 1-circular shift Y_1 of Y is the trinucleotide code $\{TCA, ATG, CGC, GGC, CAG, GCT\}$, which is also 1-circular but is not self-complementary. Second, the 2-circular shift Y_2 of Y is $\{CAT, TGA, GCC, GCG, AGC, CTG\}$, which is also 1-circular (and not self-complementary). Hence Y is a self-complementary trinucleotide $(1,1,1)$-circular code.

Table 7 gives the 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.
As one sees in Table 7, there is no self-complementary trinucleotide $(3,3,3)$-circular code. In addition, all self-complementary trinucleotides $(2,2,2)$-circular codes have size at least 10 and at most 16. There are exactly 4 self-complementary trinucleotides 2-circular codes of size 16 (see List 5.37).

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	28	28
4	0	0	0	330	330
6	68	0	0	2064	2132
8	1764	0	0	7102	8866
10	17408	96	0	13956	31460
12	80915	184	0	16764	97863
14	195388	56	0	12876	208320
16	259624	4	0	6252	265880
18	186296	0	0	1752	188048
20	57681	0	0	216	57897
Total	799144	340	0	61340	860824

LIST 5.37 (The 4 minimum self-complementary trinucleotide $(2, 2, 2)$-circular codes of size 16).

$\{AAC, GTT, AAG, CTT, AAT, ATT, CAC, GTG, CAG, CTG, CTC, GAG, GAC, GTC, TCA, TGA\}$,
$\{ACA, TGT, ACC, GGT, ACT, AGT, AGA, TCT, CAG, CTG, GCC, GGC, GGA, TCC, TCA, TGA\}$,
$\{ACA, TGT, ACT, AGT, AGA, TCT, AGG, CCT, CCA, TGG, CCG, CGG, GAC, GTC, TCA, TGA\}$,
$\{ACT, AGT, CAA, TTG, CAC, GTG, CAG, CTG, CTC, GAG, GAA, TTC, GAC, GTC, TAA, TTA\}$.

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.
Biological analyses inspired from this work are presented in the companion article [**5**].

References

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

[2] E. Fimmel, C. J. Michel, and L. Strüngmann, *n-nucleotide circular codes in graph theory*, Philosophical Transactions of the Royal Society A **374, 20150058** (2016), 1–19.

[3] E. Fimmel, C. J. Michel, F. Pirot, J.-S. Sereni, M. Starman, and L. Strüngmann, *The relation between k-circularity and circularity of codes*, Bulletin of Mathematical Biology **82, 105** (2020), 1–34.

[4] A. Itai and M. Rodeh, *Finding a minimum circuit in a graph*, SIAM J. Comput. **7** (1978), no. 4, 413–423.

[5] C. J. Michel and J.-S. Sereni, *Trinucleotide k-circular codes II: biology* (2021), available at MiSe21.pdf. Submitted for publication.