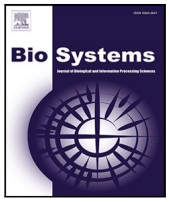
BioSystems 217 (2022) 104668

Contents lists available at ScienceDirect

BioSystems

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



Trinucleotide k-circular codes II: Biology

Christian J. Michel *, 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
Code evolution
Circularity
Self-complementarity
Balance
Genetic code
Reading frame

ABSTRACT

A code X is $(\geqslant k)$-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 $(\geqslant k)$-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. The theoretical aspects of trinucleotide k-circular codes have been developed in a companion article (Michel et al., 2022).

Trinucleotide circular codes always retrieve the reading frame, leaving no ambiguous sequences. On the contrary, trinucleotide k-circular codes, for $k \in \{0,1,2,3\}$ all have ambiguous sequences, for which the reading frame cannot always be retrieved. However, such a trinucleotide k-circular code is still able to retrieve the reading frame for a number of sequences, thereby exhibiting a partial circularity property. We describe this combinatorial property for each class of trinucleotide k-circular codes with $k \in \{0,1,2,3\}$. The circularity, i.e. the reading frame retrieval, is an ordinary property in genes. 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. We apply this approach to analyse the evolution of the trinucleotide circular code X observed in genes to the genetic code.

The $(\geqslant 1)$-circular codes of maximal size 20 necessarily have the same number of each nucleotide, specifically $15 = 3 \cdot 20/4$. This balanceness property can also be achieved by trinucleotide codes of cardinality $4, 8, 12$ and 16. We call such trinucleotide codes balanced. 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 balanceness property and the self-complementarity property.

The combinatorial hierarchy of trinucleotide k-circular codes is updated with the growth function results. The numbers of 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.

1. Introduction

In 1996, a statistical computation of the 64 trinucleotides in each of the three frames of genes of bacteria and eukaryotes ($2 \cdot 3 \cdot 64 = 384$ trinucleotides analysed) identifies, by a simple inspection, 20 trinucleotides which occur preferentially in reading frames compared to the two shifted frames (Arquès and Michel, 1996). Furthermore, this set X of 20 trinucleotides is a maximal C^3 self-complementary circular code (Arquès and Michel, 1996):

$$X = \{AAC, GTT, AAT, ATT, ACC, GGT, ATC, GAT, CAG, CTG, \quad (1.1)$$
$$CTC, GAG, GAA, TTC, GAC, GTC, GCC, GGC, GTA, TAC\}.$$

The circular code X (1.1) is also identified in genes of archaea, plasmids and viruses, in addition to bacteria and eukaryotes, and by two different statistical approaches (Michel, 2015, 2017). The historical context of this result is described in a recent article (Michel, 2020). We also refer the reader to the reviews (Michel, 2008; Fimmel and Strüngmann, 2018) for the biological context and the main combinatorial studies of circular codes. The necessary definitions and theorems will be recalled here.

Motifs from the circular code X (1.1), called *X-motifs*, are significantly enriched in the genes of most organisms, from bacteria to eukaryotes (Michel et al., 2017; Dila et al., 2019a). However, these

* Corresponding author.
E-mail addresses: c.michel@unistra.fr (C.J. Michel), sereni@kam.mff.cuni.cz (J.-S. Sereni).

https://doi.org/10.1016/j.biosystems.2022.104668
Received 15 October 2021; Received in revised form 27 February 2022; Accepted 16 March 2022
Available online 28 March 2022

X-motifs that retrieve the reading frame in genes are discontinuous and separated by motifs that are unable to retrieve the reading frame. The k-circular codes, in particular trinucleotide k-circular codes, have the mathematical property to generate both motifs that retrieve and motifs that do not retrieve the reading frame in genes.

The necessary definitions and notations are gathered in Section 2, and they follow those in the companion article (Michel et al., 2022), which details the theoretical aspects.

In Section 3.1, we develop a method, based on graph theory, to explicitly determine all ambiguous sequences for a given trinucleotide k-circular code. We then apply the method to the classes of trinucleotide k-circular codes in Section 3.2, which allows us to design new rules to retrieve the reading frame in genes.

In Section 4, we show that the circularity property (reading frame retrieval) is actually an ordinary property: all the trinucleotide codes can be classified into three classes, corresponding to "no circularity", "partial circularity" and "complete circularity". As it turns out, every self-complementary trinucleotide k-circular code of size different from 8 and 64 has at least a partial circularity property, and every trinucleotide k-circular code of cardinality different from $1, 8, 27$ and 64 also has at least a partial circularity property.

In Section 5.1, we derive a new and general formula to measure the reading frame loss, whatever the trinucleotide k-circular code. We apply it in Section 5.2 to propose an evolutionary model of the trinucleotide circular code X observed in genes to the genetic code.

In Section 6.1, we study a newly introduced and interesting property: the balanceness of trinucleotide codes, which we relate to the circularity and self-complementarity properties. After having explained why all trinucleotide ($\geqslant$ 1)-circular codes of maximal size 20 are balanced, we develop in Section 6.2 a general method based on linear algebra to compute the number of balanced trinucleotide codes of each size, which also applies to self-complementary trinucleotide codes. We exhibit and quantify the relation with self-complementarity in Section 6.3.

In Section 7, we update the hierarchy of trinucleotide k-circular codes.

In Section 8, we perform an in-depth study of the amino acids coded by the maximal, minimal, self-complementary trinucleotide k- or (k, k, k)-circular codes. For each class, the maximum numbers of amino acids coded are determined and the list of corresponding trinucleotide codes is explicitly given. Interestingly, this maximum number for a class is not always attained by the trinucleotide codes of maximal cardinality (within the class).

2. Definitions and notations

For the reader's convenience we here recall the most relevant notions, in order to have this article self-contained. The theoretical aspects, with computer results, proofs, examples, remarks, illustrations and refinements are found in the companion article (Michel et al., 2022).

We work with the *genetic alphabet* $\mathcal{B} := \{A, C, G, T\}$, 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, using the standard word-theory notation, an element of $\mathcal{B}^3$. 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. 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.

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 $c := x_1 \cdots x_m$ 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 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 ($\geqslant k$)-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. The code X is k*-circular* if X is ($\geqslant k$)-circular and not ($\geqslant k+1$)-circular.[1]
- The code X is *circular* if it is ($\geqslant k$)-circular for all $k \in \mathbf{N}$.

We recall the definition of the graph associated to a trinucleotide code (Fimmel et al., 2016).

Definition 2.3. 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:

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

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

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. As explained in the companion article (Michel et al., 2022), a theorem (Fimmel et al., 2020, Theorem 3.3) implies that a cycle in $\mathcal{G}(X)$, if any, must be have length in $\{2, 4, 6, 8\}$ and, in particular, that a trinucleotide ($\geqslant$4)-circular code must be circular. It follows that all trinucleotide codes over $\mathcal{B}$ can be naturally partitioned into 5 classes using the following definition.

Definition 2.4. 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.

Thus, the possible values of $\mathrm{cir}(X)$ for a trinucleotide code X are $0, 1, 2, 3, 4$, which determine the 5 classes.

3. Ambiguous sequences determined from the graph associated to a trinucleotide k-circular code

We show in this section how all sequences that are ambiguous for a trinucleotide k-circular code – i.e. impossibility to identify the reading frame – can be determined from the associated graph.

[1] 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.

3.1. Method

As it turns out, every ambiguous sequence comes from a "concatenation" of directed cycles in the graph, and corresponds to a directed walk, in the following sense. All notions will be illustrated on examples.

As is usual for graphs without parallel arcs, let us designate a directed walk using only vertices.

Definition 3.1. A *directed walk from* v_0 *to* v_ℓ in a graph $\mathcal{G}$ is a sequence $W := v_0, \dots, v_\ell$ of vertices of $\mathcal{G}$ with $\ell \geq 1$ such that for each $i \in \{1, \dots, \ell\}$, there is an arc in $\mathcal{G}$ from v_{i-1} to v_i. The directed walk W is *closed* if $v_0 = v_\ell$.

Note that the vertices in a directed walk are not required to be all distinct, and neither are the arcs involved.

We are interested in sequences of trinucleotides with more than one circular X-decomposition, for a given trinucleotide k-circular code X, as defined in Definition 2.2.

Definition 3.2. Let X be a trinucleotide code and w a sequence of trinucleotides.

(1) We say that w is a *sequence with ambiguous frame* for X if

- w admits an X-decomposition; and
- the 1-shift w_1 or the 2-shift w_2 of w, possibly both, admits an X-decomposition.

(2) We say that w is a *frameless sequence* for X if

- w does not admit an X-decomposition; and
- both the 1-shift w_1 and the 2-shift w_2 of w admit an X-decomposition.

Note that the length of the sequences defined in Definition 3.2 must be a multiple of 3, as some of their circular shifts must admit an X-decomposition. It is thus useful to define the trinucleotide length of a sequence as follows.

Definition 3.3. If the sequence w is a concatenation of trinucleotides, then its *trinucleotide length* $\ell(w)$ is the number of trinucleotides concatenated, that is, the number of nucleotides in w divided by 3.

Notice also that if w is frameless for a trinucleotide code X, then both its circular 1-shift and its circular 2-shift are sequences with ambiguous frame for X.

Fix a trinucleotide k-circular code X, for some $k \in \{0, 1, 2, 3\}$. Rephrasing arguments already exploited earlier (Fimmel et al., 2016, 2020), a sequence with ambiguous frame for X corresponds to a directed closed walk in $\mathcal{G}(X)$. More precisely, let $W = v_0, \dots, v_\ell$ be a directed closed walk in $\mathcal{G}(X)$ (so $v_0 = v_\ell$), and let w be the sequence of nucleotides obtained by concatenating the nucleotides and dinucleotides (i.e. vertices) $v_0, \dots, v_{\ell-1}$, respecting the order. If v_0 is a nucleotide, then the sequence w_1 also admits an X-decomposition, while if v_0 is a dinucleotide, then w_2 also admits an X-decomposition. Conversely, let $w = N_1 \cdots N_s$ be a sequence of s nucleotides obtained by concatenating trinucleotides from X. If w_1 admits an X-decomposition, then $N_1, N_2N_3, \dots, N_{s-2}, N_{s-1}N_s, N_1$ is a directed closed walk in $\mathcal{G}(X)$. If w_2 admits an X-decomposition, then $N_1N_2, N_3, \dots, N_{s-2}N_{s-1}, N_s, N_1N_2$ is a directed closed walk in $\mathcal{G}(X)$. We thus see that if both w_1 and w_2 admit an X-decomposition, then two different directed closed walks give rise to w: the one starting with N_1 and the one starting with N_1N_2. In summary, every directed closed walk in $\mathcal{G}(X)$ yields a sequence with ambiguous frame, and if two different directed closed walks give rise to the same sequence with ambiguous frame w, then all three circular shifts of w admit an X-decomposition.

It is an elementary fact of graph theory – which can be obtained by a straightforward induction – that if W is a directed closed walk from v_0 to itself, then the subgraph formed by the arcs spanned by W can be decomposed into directed cycles $C_1, \dots, C_t$ such that each of these directed cycles has a vertex in common with (at least) another one of them (and C_1 goes through v_0). In other words, the subgraph spanned by the arcs of these t directed cycles is (strongly) connected, and for each vertex v of the subgraph, the in-degree and the out-degree of v are the same. Therefore, every sequence with ambiguous frame for X is built from a sequence of directed cycles forming a (strongly) connected subgraph of $\mathcal{G}(X)$, and hence one can view the directed cycles of $\mathcal{G}(X)$ as a basis generating all possible sequences with ambiguous frame for X. We however point out that one such sequence $C_1, \dots, C_t$ gives rise to different sequences with ambiguous frame – that are not circular shifts of one another – as once arranged in directed cycles, the vertices may be in a different order than originally, and may even not form a directed walk.

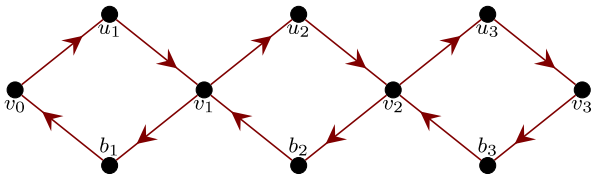


Fig. 1. A graph $\mathcal{G}$ composed of 3 directed cycles of length 4.

For example, let $\mathcal{G}$ be the graph depicted in Fig. 1, which is a subgraph of the graph associated to a trinucleotide 1-circular code. The graph formed by the arcs spanned by the directed closed walk

$$W := v_0, u_1, v_1, u_2, v_2, u_3, v_3, b_3, v_2, b_2, v_1, b_1, v_0$$

is $\mathcal{G}$ itself, which indeed decomposes into three directed cycles C_1, C_2, C_3, where $C_i := v_{i-1} \to u_i \to v_i \to b_i \to v_{i-1}$ for $i \in \{1, 2, 3\}$. However, the sequence of vertices obtained by traversing these three directed cycles is

$$v_0, u_1, v_1, b_1, v_0, v_1, u_2, v_2, b_2, v_1, v_2, u_3, v_3, b_3, v_2,$$

which is different from W and does not correspond to a directed walk.

This fact also puts constraints on the possible lengths of ambiguous sequences. Indeed, recall that the arcs spanned by any directed closed walk in a graph can be partitioned into sets of arcs each spanning a directed cycle of the graph.

As a consequence the length of a sequence gives an important information regarding the reading frame retrieval of a trinucleotide k-circular code. We now make this precise by giving several properties for trinucleotide k-circular codes for each $k \in \{0, 1, 2, 3\}$.

3.2. Application to each class of trinucleotide k-circular codes

3.2.1. Trinucleotide 3-circular codes

If X is a trinucleotide 3-circular code then every directed cycle in $\mathcal{G}(X)$ has length 8. The next two observations follow.

Observation 3.4. *The reading frame of any sequence w with trinucleotide length $\ell(w)$ not divisible by 4 can be retrieved by any trinucleotide 3-circular code, i.e. either $\ell(w) \equiv 0 \pmod 4$ or w is not ambiguous.*

Observation 3.5. *Any sequence w with ambiguous frame for a trinucleotide 3-circular code must have a trinucleotide length $\ell(w)$ multiple of 4, that is $\ell(w) \equiv 0 \pmod 4$.*

As an example, let us consider the following trinucleotide 3-circular code of size 12:

$$X_6 := \{AAT, ACG, ATT, CAG, CGT, CTG, GCC, GGC, GTA, TAC, TCA, TGA\}.$$

The associated graph $\mathcal{G}(X_6)$ is depicted in Fig. 2: $\mathcal{G}(X_6)$ contains a directed cycle of length 8 and no shorter one, so X_6 is 3-circular. As

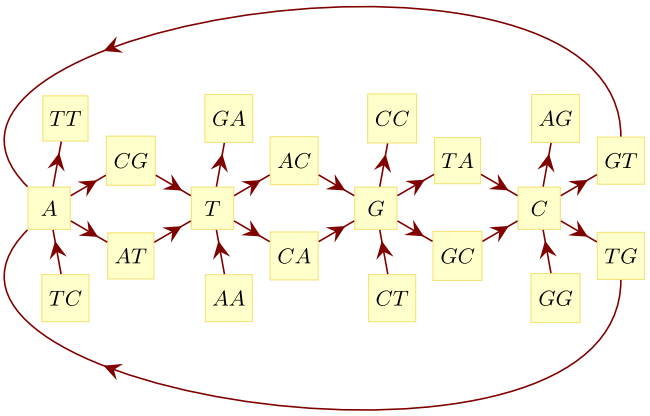


Fig. 2. The graph $\mathcal{G}(X_6)$ associated to the trinucleotide 3-circular code X_6 contains 16 different directed cycles of length 8, two of which having no arc in common.

a matter of fact, $\mathcal{G}(X_6)$ contains precisely $2^4 = 16$ different directed cycles (all of length 8): all of them must go through A, T, G, C (in this circular order), and between any two consecutive nucleotides there are exactly 2 directed paths of length 2 to choose from. In particular, for any directed cycle C in $\mathcal{G}(X_6)$, there exists another directed cycle that is arc-disjoint from C (obtained by always using the path of length 2 not intersecting C between two consecutive nucleotides).

Let us consider the directed closed walk $W := A, CG, T, AC, G, TA, C, GT, A$ in $\mathcal{G}(X_6)$. It gives rise to the sequence $w := ACG \cdot TAC \cdot GTA \cdot CGT$, which is composed of 4 trinucleotides from X_6. Both the circular 1-shift and the circular 2-shift of w admit an X_6-decomposition, namely $w_1 = CGT \cdot ACG \cdot TAC \cdot GTA$ and $w_2 = GTA \cdot CGT \cdot ACG \cdot TAC$. All three circular shifts of w are thus sequences with ambiguous frame. Therefore the sequence of vertices $AC, G, TA, C, GT, A, CG, T, AC$ must also be a directed closed walk in $\mathcal{G}(X)$, which gives rise to the same sequence w. There exist also sequences with exactly 2 circular shifts having ambiguous frame, and hence there exist frameless sequences, as defined in Definition 3.2(2). For instance, the sequence $w' := GAA \cdot TTA \cdot CGG \cdot CCT$ does not admit an X_6-decomposition because $GAA \notin X_6$. However, both $w'_1 = AAT \cdot TAC \cdot GGC \cdot CTG$ and $w'_2 = ATT \cdot ACG \cdot GCC \cdot TGA$ admit an X_6-decomposition, and hence w' is a frameless sequence. Note that, consequently, w'_1 and w'_2 are both sequences with ambiguous frame; the unique directed closed walk in $\mathcal{G}(X_6)$ giving rise to w'_1 being $W'_1 := A, AT, T, AC, G, GC, C, TG, A$ and the unique directed closed walk in $\mathcal{G}(X_6)$ giving rise to w'_2 being $W'_2 := AT, T, AC, G, GC, C, TG, A, AT$. Notice that considering the last vertex of W to be the first of W'_1 (which can be understood as "concatenating" these two directed closed walks), we obtain the directed closed walk

$$A, CG, T, AC, G, TA, C, GT, A, AT, T, AC, G, GC, C, TG, A,$$

of $\mathcal{G}(X_6)$, which decomposes in 2 directed cycles of length 8. It gives rise to the sequence with ambiguous frame

$$w \cdot w'_1 = ACG \cdot TAC \cdot GTA \cdot CGT \cdot AAT \cdot TAC \cdot GGC \cdot CTG,$$

obtained by concatenating w and w'_1. The circular 1-shift of $w \cdot w'_1$, which is

$$CGT \cdot ACG \cdot TAC \cdot GTA \cdot ATT \cdot ACG \cdot GCC \cdot TGA,$$

admits an X_6-decomposition while the circular 2-shift of $w \cdot w'_1$, which is

$$GTA \cdot CGT \cdot ACG \cdot TAATTACGGCCTGAC,$$

does not — and hence this 2-shift is a frameless sequence.

One can also directly observe from the structure of the directed cycles in $\mathcal{G}(X_6)$ that removing a single trinucleotide from X_6 cannot yield a trinucleotide circular code. As every sequence with ambiguous frame must contain the trinucleotide AAT or ACG, removing both AAT and ACG from X_6 yields a trinucleotide circular code, since these removals would destroy all directed cycles in the graph.

3.2.2. Trinucleotide 2-circular codes

Consider now a trinucleotide 2-circular code X. The associated graph must contain a directed cycle of length 6 and no shorter one. It may or may not contain a directed cycle of length 8. In the latter case, our previous considerations imply that the number of trinucleotides from X concatenated to create a sequence with ambiguous frame must be a multiple of 3. In the former case, since directed cycles of different lengths must have a vertex in common (because there are only 4 nucleotides and every directed cycle contains at least two of them in a trinucleotide ($\geqslant$1)-circular code), our previous considerations show that one can build sequences with ambiguous frame by concatenating any number of trinucleotides greater than 2 and different from 5. This exhibits a drastic difference between the class of trinucleotide 2-circular codes and that of trinucleotide 3-circular codes; and actually even within the class of trinucleotide 2-circular codes, depending on whether or not the associated graph contains directed cycles of length 8. We thus obtain the following observations.

Observation 3.6. *The reading frame of any sequence w with trinucleotide length $\ell(w)$ not divisible by 3 can be retrieved by any trinucleotide 2-circular code without directed cycles of length 8 in the associated graph, i.e. either $\ell(w) \equiv 0 \pmod 3$ or w is not ambiguous (for such a trinucleotide code).*

Observation 3.7. *The reading frame of any sequence w with trinucleotide length $\ell(w) \in \{1, 2, 5\}$ can be retrieved by any trinucleotide 2-circular code.*

Observation 3.8. *Any sequence w with ambiguous frame for a trinucleotide 2-circular code without directed cycles of length 8 in the associated graph must have a trinucleotide length $\ell(w)$ divisible by 3, that is $\ell(w) \equiv 0 \pmod 3$.*

Observation 3.9. *Any sequence w with ambiguous frame for a trinucleotide 2-circular code must have a trinucleotide length $\ell(w)$ greater than 2 and different from 5, that is $\ell(w) \geq 3$ and $\ell(w) \neq 5$.*

3.2.3. Trinucleotide 1-circular codes

A trinucleotide 1-circular code X can potentially admit a sequence with ambiguous frame composed of t trinucleotides from X, for any integer t greater than 1, which as we shall see in Section 3.2.4 is very close to the case of the trinucleotide 0-circular codes. However, if the associated graph contains no directed cycle of length other than 4, then every ambiguous sequence is the concatenation of an even number of trinucleotides — and hence such a code X will retrieve the reading frame in any concatenation of an odd number of trinucleotides from X.

Observation 3.10. *The reading frame of any sequence w with trinucleotide length $\ell(w)$ not divisible by 2 can be retrieved by any trinucleotide 1-circular code without directed cycles of length different from 4 in the associated graph, i.e. either $\ell(w)$ is even or w is not ambiguous (for such a trinucleotide code).*

Observation 3.11. *Any sequence w with ambiguous frame for a trinucleotide 1-circular code without directed cycles of length different from 4 in the associated graph must have a trinucleotide length $\ell(w)$ divisible by 2, that is $\ell(w) \equiv 0 \pmod 2$.*

As another example, consider the code $X_7 := \{AAT, ACG, CAA, CGG, GAA, GGA, TCA\}$, with the associated graph $\mathcal{G}(X_7)$ depicted in Fig. 3. Since $\mathcal{G}(X_7)$ has a directed cycle of length 4 and no shorter one, X_7 is 1-circular. However, as $\mathcal{G}(X_7)$ also contains a cycle of length 6, the trinucleotide code X_7 admits a sequence with ambiguous

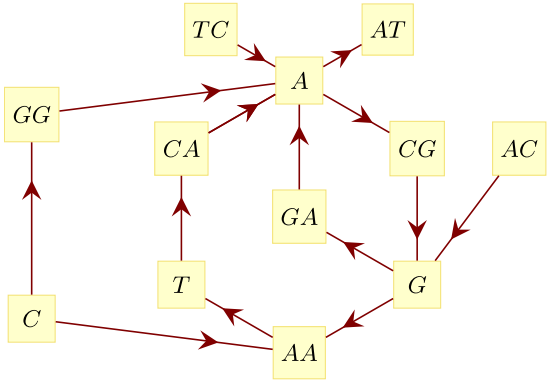


Fig. 3. The graph $\mathcal{G}(X_7)$ associated to the trinucleotide 1-circular code X_7 contains exactly 2 directed cycles: one of length 4 and one of length 6. Therefore there are sequences with ambiguous frames for X_7 composed of any number of trinucleotides greater than 1.

frame composed of t trinucleotides for any integer $t \geq 2$. For instance, $w = ACG \cdot GGA$ and $w' = ACG \cdot GAA \cdot TCA$ are two sequences with ambiguous frames, respectively composed of 2 and 3 trinucleotides. The former is obtained from the directed closed walk A, CG, G, GA, A and the latter from A, CG, G, AA, T, CA, A. We deduce that $w \cdot w$ and $w \cdot w'$ also are ambiguous sequences, which are respectively composed of 4 and 5 trinucleotides.

3.2.4. *Trinucleotide 0-circular codes*

Trinucleotide 0-circular codes can admit sequences with ambiguous frame of any positive trinucleotide length, since they must contain a word and its circular 1-shift.

Observation 3.12. *Sequences w of any positive trinucleotide length $\ell(w)$ with ambiguous frame exist for any trinucleotide 0-circular code.*

4. Circularity (reading frame retrieval): an ordinary property in genes

On the genetic alphabet $\mathcal{B}$, there are $2^{64} \approx 10^{19}$ trinucleotide codes (including the empty set). According to the theoretical work developed earlier (Michel et al., 2022), they can be classified into 3 classes according to their circularity property, i.e. their property of reading frame retrieval:

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

Remark 4.1. A code consisting of a single periodic trinucleotide, $N_1N_1N_1$ where $N_1 \in \mathcal{B}$, has no circularity at all (class (i)), since the only sequences that can be formed from it are constant sequences $N_1 \cdots N_1$. The codes $\{N_1N_1N_1, N_2N_2N_2\}$, $\{N_1N_1N_2, N_1N_2N_1\}$ and $\{N_1N_2N_3, N_2N_3N_1\}$, where $N_1, N_2, N_3 \in \mathcal{B}$ with $N_1 \neq N_2$, all have a partial circularity (class (ii)). In addition, the trinucleotide code formed of a single conjugacy class $\{N_1N_2N_3, N_2N_3N_1, N_3N_1N_2\}$, where not all nucleotides are equal, also has a partial circularity (class (ii)).

The following proposition determines all trinucleotide codes that belong to class (i). Interestingly, and as we shall see later, it implies that class (i) contains only 3 self-complementary trinucleotide codes: the genetic code $\mathcal{B}^3$ itself and two self-complementary trinucleotide codes each of size 8.

Proposition 4.2. *All but 15 of the 2^{64} trinucleotide codes belong to class (ii) or (iii), and thus have a (partial) circularity property. The 15 trinucleotide codes that have no circularity (class (i)) are those consisting of all possible trinucleotides over a subset of $\mathcal{B}$ (which can be of size 1, 2, 3 or 4), that is:*

- *the 4 trinucleotide codes $\{N_1N_1N_1\}$ of size 1 for $N_1 \in \mathcal{B}$,*
- *the 6 trinucleotide codes $\{N_1, N_2\}^3$ of size 8 for $N_1, N_2 \in \mathcal{B}$ with $N_1 \neq N_2$,*
- *the 4 trinucleotide codes $\{N_1, N_2, N_3\}^3$ of size 27 for $N_1, N_2, N_3 \in \mathcal{B}$ with $N_1 \neq N_2 \neq N_3 \neq N_1$, and*
- *the genetic code $\mathcal{B}^3$ of size 64.*

Proof. It is straightforward to realise that all 15 trinucleotide codes given indeed belong to class (i). We now prove that no other trinucleotide code belongs to this class. Let X be a trinucleotide code belonging to class (i). Notice that if S is a conjugacy class, then $|S \setminus X| \neq 2$, for otherwise the sequence formed of the unique trinucleotide in $S \cap X$ would not have ambiguous frame. We establish the statement by proving a series of assertions, the first of which states that X must actually be a union of conjugacy classes.

(A). *For each conjugacy class S, either $S \subseteq X$ or $S \cap X = \emptyset$.*

Proof. Suppose, on the contrary, that $S \cap X \neq S$ for some conjugacy class S intersecting X. Then, there exist nucleotides $\alpha, \beta, \gamma \in \mathcal{B}$ such that $\alpha\beta\gamma, \beta\gamma\alpha \in X$ and $\gamma\alpha\beta \notin X$. It follows that X cannot contain a trinucleotide starting with β and ending with γ, because if w was such a trinucleotide contained in X, then the sequence $\beta\gamma\alpha \cdot w \cdot \alpha\beta\gamma$ would not have ambiguous frame, as both its 1-shift and its 2-shift contain $\gamma\alpha\beta$. Now, the sequence $\alpha\beta\gamma \cdot \beta\gamma\alpha$ implies that $\beta\gamma\beta \in X$ or $\gamma\beta\gamma \in X$. (It follows that $\beta \neq \gamma$.) In the former case, since we proved that $\beta\beta\gamma \notin X$, we deduce that $\gamma\beta\beta \in X$ (as otherwise X would miss exactly 2 members in the conjugacy class of $\beta\gamma\beta$, which is not possible as we noticed). Consequently, the sequence $\gamma\beta\beta \cdot \gamma\beta\beta \cdot \beta\gamma\alpha$ cannot have ambiguous frame, and yet is a concatenation of words in X, a contradiction. The latter case follows by symmetry, up to reversing the trinucleotide code (reading everything from right to left): specifically, we similarly deduce that $\gamma\gamma\beta \in X$ (since $\beta\gamma\gamma \notin X$), and hence the sequence $\alpha\beta\gamma \cdot \gamma\gamma\beta \cdot \gamma\gamma\beta$ cannot have ambiguous frame, a contradiction. ⌟

(B). *If X contains a trinucleotide composed of exactly 2 different nucleotides, called N_1 and N_2, then X contains $\{N_1, N_2\}^3$, that is, all trinucleotides over the alphabet $\{N_1, N_2\}$.*

Proof. Suppose that X contains a trinucleotide composed of 2 occurrences of N_1 and one occurrence of N_2. By (A) we know that X contains the whole conjugacy class of $N_1N_1N_2$. This readily implies that X contains both $N_1N_2N_2$ and $N_1N_1N_1$, as is seen by considering the sequence $N_1N_1N_2 \cdot N_2N_1N_1$ and using (A). As a result, we deduce that $N_2N_2N_2 \in X$ (using the same argument with the roles of N_1 and N_2 swapped), which concludes the proof of the assertion. ⌟

(C). *If X contains a trinucleotide $N_1N_2N_3$ composed of 3 different nucleotides, then X contains $\{N_1, N_2, N_3\}^3$, that is, all trinucleotides over the alphabet $\{N_1, N_2, N_3\}$.*

Proof. The sequence $N_1N_2N_3 \cdot N_2N_3N_1$, which is composed of 2 trinucleotides in X by (A), shows that $N_2N_3N_2$ or $N_3N_2N_3$ belongs to X. Either way, (B) implies that X contains $\{N_2, N_3\}^3$. By (A) the roles

played by N_1, N_2 and N_3 are symmetric, so we similarly infer that X contains $\{N_1, N_2\}^3$ and also $\{N_1, N_3\}^3$. Consequently, it only remains to show that X also contains $\{N_1N_3N_2, N_3N_2N_1, N_2N_1N_3\}$. This holds because the sequence $N_1N_1N_3 \cdot N_2N_2N_1 \cdot N_3N_2N_2$ is a concatenation of trinucleotides from X, and hence must have ambiguous frame. This yields that $N_1N_3N_2 \in X$, and hence the conclusion follows by (A). ⌟

(D). *Let N_1, N_2, N_3 be 3 pairwise distinct nucleotides. If each of them is contained in a trinucleotide in X, then X contains $\{N_1, N_2, N_3\}^3$.*

Proof. By (C), it suffices to show that X contains a trinucleotide containing each of N_1, N_2 and N_3. First we observe that (A) implies that X contains a trinucleotide starting with N_i and a trinucleotide ending with N_i for each $i \in \{1,2,3\}$. Concatenating a trinucleotide starting with N_i and a trinucleotide ending with N_j, where $1 \leq i < j \leq 3$, shows that X contains a trinucleotide containing both N_i and N_j, and hence X contains $\{N_i, N_j\}^3$ by (B). Consequently, the sequence $N_1N_1N_2 \cdot N_3N_3N_2 \cdot N_1N_3N_3$ implies that X contains $N_1N_2N_3$ or $N_2N_1N_3$, which in either case yields the conclusion by (C). ⌟

(E). *If for each nucleotide $N \in \mathcal{B}$, there exists a trinucleotide in X containing N, then $X = \mathcal{B}^3$.*

Proof. This directly follows from (D), since it implies that X contains $\{N_1, N_2, N_3\}^3$ for each triple of pairwise distinct nucleotides (N_1, N_2, N_3). ⌟

We are now in a position to conclude: recall that X is a union of conjugacy classes by (A). If X contains only one conjugacy class, then it must be that of a periodic trinucleotide, by (B) and (C). Suppose now that X contains more than one conjugacy class. Then, either the trinucleotides in X use all together at least 3 pairwise distinct nucleotides, in which case the conclusion follows from (D) and (E); or the trinucleotides in X use exactly 2 different nucleotides, in which case the conclusion follows from (B). □

Corollary 4.3. *As the total number of (non-empty) trinucleotide codes is $n = 2^{64} - 1 \approx 10^{19}$, the occurrence probabilities of the 3 circularity classes can be deduced:*

(i) $15/n \approx 10^{-18}$ for the trinucleotide codes with no circularity;

(ii) $(n - 15 - 115606988558)/n \approx 1$ for the trinucleotide codes with a partial circularity;

(iii) $115606988558/n \approx 10^{-8}$ for the trinucleotide codes with a complete circularity (circular).

We detail Corollary 4.3 per cardinality $|X|$, between 1 and 64, of the trinucleotide codes X. Table 1 gives the probabilities (%) of the 3 circularity classes as the function of the cardinality $|X|$ of the trinucleotide codes X (trivially deduced from Table 1 in Michel et al., 2022). Up to the cardinality 4, the probability of complete circularity of a trinucleotide code is greater than the probability of partial circularity.

Among the 15 trinucleotide codes with no circularity (class (i), given by Proposition 4.2), only those of even size (thus 8 or 64) can be self-complementary, and among these exactly those built over a self-complementary subset of $\mathcal{B}$ are self-complementary. Therefore, there are exactly 3 self-complementary trinucleotide codes with no circularity: 2 of size 8 (respectively built over the alphabets $\{A, T\}$ and $\{C, G\}$), and 1 of size 64, the genetic code built over the alphabet $\mathcal{B}$.

Table 2 gives the probabilities (%) of the 3 circularity classes as the function of the cardinality $|X|$ of the self-complementary trinucleotide codes X (trivially deduced from Table 4 in Michel et al., 2022). As in the general case, up to the cardinality 4, the probability of complete circularity of a self-complementary trinucleotide code is greater than the probability of partial circularity.

Interestingly, Fig. 4 demonstrates that, for a given even cardinality $|X| \in \{2, \ldots, 20\}$ of the trinucleotide code X, the probability

Table 1
Probabilities (%) of the 3 circularity classes as the function of the cardinality $|X|$ of the trinucleotide codes X.

$\lvert X\rvert$	Circularity classes of X		
	No	Partial	Complete
1	6.25	0	93.75
2	0	15.48	84.52
3	0	26.96	73.04
4	0	39.85	60.15
5	0	53.20	46.80
6	0	65.98	34.02
7	0	77.20	22.80
8	$\approx 10^{-7}$	≈ 86.12	13.88
9	0	92.42	7.58
10	0	96.34	3.66
11	0	98.45	1.55
12	0	99.43	0.57
13	0	99.82	0.18
14	0	99.95	0.05
15	0	99.99	0.01
16	0	≈ 100	$\approx 10^{-3}$
17	0	≈ 100	$\approx 10^{-4}$
18	0	≈ 100	$\approx 10^{-5}$
19	0	≈ 100	$\approx 10^{-6}$
20	0	≈ 100	$\approx 10^{-7}$
$\{21, \ldots, 26\}$	0	100	0
27	$\approx 10^{-16}$	≈ 100	0
$\{28, \ldots, 63\}$	0	100	0
64	100	0	0

Table 2
Probabilities (%) of the 3 circularity classes as the function of the cardinality $|X|$ of the self-complementary trinucleotide codes X.

$\lvert X\rvert$	Circularity classes of self-complementary X		
	No	Partial	Complete
2	0	12.50	87.50
4	0	32.66	67.34
6	0	56.13	43.87
8	$\approx 10^{-2}$	≈ 76.93	23.06
10	0	90.52	9.48
12	0	96.99	3.01
14	0	99.28	0.72
16	0	99.87	0.13
18	0	99.99	0.01
20	0	≈ 100	$\approx 10^{-3}$
$\{22, \ldots, 62\}$	0	100	0
64	100	0	0

of complete circularity is greater with the self-complementary trinucleotide codes compared to the trinucleotide codes. In other words, self-complementarity has favoured complete circularity, i.e. the trinucleotide circular codes, or reciprocally.

Proposition 4.2 leads to several important consequences.

Corollary 4.4. *Any trinucleotide code that is not a union of conjugacy classes has a circularity property, partial or complete; in particular a "random" trinucleotide code typically has a circularity property.*

Corollary 4.5. *Any trinucleotide code of size not in $\{1, 8, 27, 64\}$ has always a circularity property, partial or complete.*

Corollary 4.6. *Any self-complementary trinucleotide code of size not in $\{8, 64\}$ has always a circularity property, partial or complete.*

Corollaries 4.4–4.6 explain some unexpected and strange distributions of "random" trinucleotide codes in genes that sometimes are close to the distributions of some trinucleotide circular codes. This statistical trinucleotide circular code noise observed by several authors in the past, including the first author as early as 1996 and recently reported again by Gumbel and Wiedemann (Gumbel and Wiedemann, 2021), is now explained by our theoretical work. The method developed in

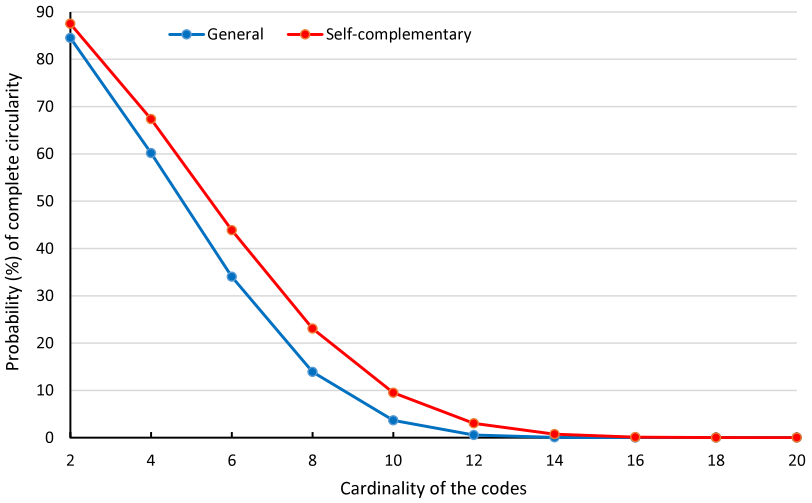


Fig. 4. Probability (%) of complete circularity as the function of the cardinality $|X|$ of self-complementary trinucleotide codes versus trinucleotide codes. The X-axis is the cardinality of the trinucleotide codes (even number between 2 and 20) and the Y-axis is the probability (%) of complete circularity.

Section 5 allows for a new classification of trinucleotide codes with a partial circularity according to the intensity of the loss of reading frame retrieval (see Eq. (5.1)).

5. A new formula to measure the reading frame loss in the trinucleotide k-circular codes

5.1. Method

For the trinucleotide circular codes, the window for retrieving the reading frame is directly determined by the length of a longest directed path p in the associated graph (which is well defined since the associated graph is acyclic). The larger this directed path length is, the larger the number of nucleotides required to always retrieve the reading frame is. Trinucleotide circular codes have thus been partitioned according to this length into 8 classes (as we shall see in Section 7), starting with the more restrictive strong comma-free codes and comma-free codes to the more flexible circular codes in X_8.

With the trinucleotide k-circular codes, which generalise the trinucleotide circular codes, this approach cannot be used anymore: the graph associated to a trinucleotide k-circular code with $k < 4$ is no longer acyclic. As a result, it contains directed paths of arbitrarily large lengths. However, a new measure can be proposed using the graph analysis carried out in Section 3, regarding closed directed walks.

Indeed, as reported earlier, the length of the smallest sequences with ambiguous frame for X for a given trinucleotide k-circular code X only depends on the lengths of the directed cycles in the associated graph $\mathcal{G}(X)$: a directed cycle of length 2ℓ implies a sequence with ambiguous frame composed of ℓ trinucleotides. Since $\mathcal{G}(X)$ has an infinite number of directed closed walks, we rather consider the number of directed cycles of each possible length normalised as follows.

Definition 5.1. The *reading frame loss* function f of a trinucleotide code X is the mapping $f : B^3 \to \mathbf{R}$ given by

$$f(X) := \mathrm{q}_8(\mathcal{G}(X)) + \frac{4}{3}\,\mathrm{q}_6(\mathcal{G}(X)) + 2\,\mathrm{q}_4(\mathcal{G}(X)) + 4\,\mathrm{q}_2(\mathcal{G}(X))$$
$$= \sum_{i=1}^{4} \frac{4}{i} \cdot \mathrm{q}_{2\cdot i}(\mathcal{G}(X)), \tag{5.1}$$

where $\mathrm{q}_i(\mathcal{G})$ is the number of directed cycles of length i in the graph $\mathcal{G}$ for every positive integer i.

Note that $f(X)$ is always a rational number, but not necessarily an integer.

Proposition 5.2. *For every trinucleotide code X, we have $0 \le f(X) \le 301056$. Moreover, $f(X) = 0$ if and only if X is a trinucleotide circular code, and $f(X) = 301056$ if and only if X is the genetic code X_g, where*

$$\mathrm{q}_2(X_g) = 64, \qquad \mathrm{q}_4(X_g) = 1440, \qquad \mathrm{q}_6(X_g) = 26880,$$
$$\mathrm{q}_8(X_g) = 262080.$$

Proof. Let X be a trinucleotide circular code and $\mathcal{G}(X)$ its associated graph. Since each directed cycle in $\mathcal{G}(X)$ must have even length not exceeding 8, we deduce that $f(X) = 0$ if and only if the associated graph $\mathcal{G}(X)$ is acyclic, which holds if and only if X is circular.

Moreover, every trinucleotide code X is contained in X_g, which implies that $\mathrm{q}_{2i}(X) \le \mathrm{q}_{2i}(X_g)$ for each $i \in \{1,2,3,4\}$, and hence $f(X) \le f(X_g)$. Now, because every arc of $\mathcal{G}(X_g)$ is contained in directed 2-cycle, if $X \ne X_g$ then necessarily $\mathrm{q}_2(X) < \mathrm{q}_2(X_g)$, and hence $f(X) < f(X_g)$.

Finally, for each $i \in \{1,2,3,4\}$, every directed cycle in $\mathcal{G}(X_g)$ corresponds to the choice of i circularly ordered nucleotides $N_1, \dots, N_i$ and i ordered dinucleotides $d_1, \dots, d_i$, because X_g contains all possible trinucleotides. There are $i!$ possible ways of ordering any of the $\binom{16}{i}$ possible subsets of i dinucleotides. Similarly, there are precisely $(i-1)!$ ways of circularly ordering any of the $\binom{4}{i}$ possible subsets of i nucleotides. Therefore,

$$\mathrm{q}_{2i}(X_g) = \binom{4}{i}(i-1)! \cdot \binom{16}{i} i!,$$

which concludes the proof. □

A similar analysis can be developed for the *amino acid code* X_{AA}, composed of all trinucleotides except the `stop` codons, namely TAA, TAG and TGA, i.e. $X_{\text{AA}} = X_g \setminus \{TAA, TAG, TGA\}$.

Proposition 5.3. *The reading frame loss function f of the amino acid code X_{AA} is $f(X_{\text{AA}}) = \frac{600332}{3} \approx 200110$, where*

$$\mathrm{q}_2(X_{\text{AA}}) = 58, \qquad \mathrm{q}_4(X_{\text{AA}}) = 1171, \qquad \mathrm{q}_6(X_{\text{AA}}) = 19628,$$
$$\mathrm{q}_8(X_{\text{AA}}) = 171366.$$

Proof. We proceed similarly as in the proof of Proposition 5.2, but we now have to take into account the fact that the following arcs are not present in $\mathcal{G}(X_{\text{AA}})$:

$$T \to AA, \qquad T \to AG, \qquad T \to GA$$
$$A \leftarrow TA, \qquad A \leftarrow TG, \qquad G \leftarrow TA.$$

Consequently, to compute $\mathrm{q}_2(X_{\text{AA}})$ we have 4 choices for the nucleotide N in a directed cycle of length 2, and then respectively

16, 15, 14, or 13 choices for the dinucleotide regarding whether N is C, G, A or T. This yields a total of 58 different directed cycles of length 2.

For $q_4(X_{AA})$, there are $\binom{4}{2} = 6$ choices for the set $S = \{N_1, N_2\}$ of two nucleotides in a directed cycle of length 4 — the order does not matter here. Discriminating regarding each possible choice of N_1 and N_2, the number of possibilities for the two dinucleotides are

$15 \cdot 14$ if $S = \{A, C\}$, 14^2 if $S = \{A, G\}$,
$15 \cdot 11$ if $S = \{A, T\}$,
15^2 if $S = \{C, G\}$, $15 \cdot 13$ if $S = \{C, T\}$,
$15 \cdot 12$ if $S = \{G, T\}$,

for a total of 1171.

For $q_6(X_{AA})$, there are 4 choices for the set $S = \{N_1, N_2, N_3\}$ of three nucleotides in a directed cycle of length 6, and each set can occur in two different orderings along the cycle. The number of choices for the three dinucleotides are

$2 \cdot 14^3$ if $S = \{A, C, G\}$,
$11 \cdot 14 \cdot 15 + (11 \cdot 13 \cdot 14 + 2 \cdot 14^2)$ if $S = \{A, C, T\}$,
$11 \cdot 14^2 + (11 \cdot 13 \cdot 14 + 14^2)$ if $S = \{A, G, T\}$,
$(12 \cdot 14^2 + 14 \cdot 15) + 12 \cdot 15 \cdot 14$ if $S = \{C, G, T\}$,

for a total of 19628.

Finally, for $q_8(X_{AA})$, all 4 nucleotides appear on a directed cycle of length 8, in 6 different possible orders. We observe that the two orders A, C, G, T and A, G, C, T yield the same number of choices for the 4 dinucleotides, and similarly for the two orders A, G, T, C and T, C, G, A, and for the two orders A, C, T, G and T, G, C, A. These three numbers respectively are

$11 \cdot 14^2 \cdot 13$, $11 \cdot 13^3 + 14 \cdot 13^2 + 14^2 \cdot 13$,
$11 \cdot 13 \cdot 14 \cdot 13 + 14^2 \cdot 13$,

for a total of $2 \cdot 85683 = 171366$. □

The function f can be considered as a measure of the reading frame loss: for a trinucleotide code X, the smaller the value of the function $f(X)$ is, the lower the reading frame loss is.

We remark that the approach taken here (and in Section 3) generalises to arbitrary finite word lengths (dinucleotide, tetranucleotide) and to arbitrary finite alphabets.

We point out that the two aforementioned measures (the length of a longest directed path p and the *reading frame loss* function f) are enough to analyse the reading frame retrieval property in all classes of trinucleotide codes.

5.2. Application: evolution of the trinucleotide circular code X to the genetic code

The study proposed in Section 5.1 allows us to propose for the first time a model of evolution from a trinucleotide circular code to the genetic code, and more precisely to study the ability to retrieve the reading frame for trinucleotide codes of cardinality greater than 20 thanks to the reading frame loss function f (Definition 5.1). Fig. 5 proposes an evolution from the trinucleotide circular code X defined in (1.1) to the genetic code X_g.

Keeping the self-complementarity property of X, we subsequently add to X all possible pairs of complementary codons. More precisely, at first there are exactly $32 - 10 = 22$ pairs of complementary codons not in X. Consequently, if we want to add to X a certain number n of pairs of complementary codons, then there are exactly $\binom{22}{n}$ possible choices. For instance, for the cardinality 22, we have $\binom{22}{1} = 22$ different trinucleotide codes, for the cardinality 24 we have $\binom{22}{2} = 231$ different trinucleotide codes and so forth.

For each possible cardinality of the trinucleotide codes (that are self-complementary extensions of X), Fig. 5 gives the minimum, the mean and the maximum of the reading frame loss function f over all the possible trinucleotide codes. In particular, the mean $\bar{f}$ is thus calculated as follows for each even cardinality $2 \cdot (10 + n) \in \{22, \ldots, 64\}$:

$$\bar{f} := \frac{1}{\binom{22}{n}} \sum_{X'} f(X'), \tag{5.2}$$

where the sum runs over all the self-complementary trinucleotide codes X' of cardinality $2 \cdot (10 + n)$ that contain X.

Moreover, for a given cardinality, several sequences can achieve the minimum for the reading frame loss function (see Appendix). From a certain point, the extensions that contain the periodic trinucleotides AAA and TTT do not achieve the minimum of the reading frame loss function (see Appendix).

6. Three new properties in the evolution of the genetic code: circularity, complementarity and balance

The evolution of primitive codes to the genetic code according to a process of reading frame retrieval, may involve three properties which are investigated in this section: the two classical properties of circularity and self-complementarity, and a new identified property of trinucleotide code balance.

6.1. Balanced trinucleotide codes

Any trinucleotide circular code X of maximal size 20 must be *balanced*, in the sense that the 4 nucleotides must appear the same number of times (15) in the code. Formally, the number of occurrences of a given nucleotide in the sequence of nucleotides formed by the concatenation of all 20 trinucleotides of X, which has thus size $3 \cdot 20 = 60$, contains precisely 15 occurrences of each nucleotide.

This balance property actually holds for all trinucleotide ($\geqslant 1$)-circular codes of cardinality 20. Indeed, by definition such a code X contain exactly one trinucleotide in each of the 20 conjugacy classes S, and hence X can be seen as a mapping $g : \mathcal{C} \to \mathcal{B}^3 \setminus \mathcal{P}$ where $\mathcal{C}$ is the set composed of the 20 conjugacy classes and $\mathcal{P} := \{AAA, CCC, GGG, TTT\}$ is the set of the 4 periodic trinucleotides. The 60 non-periodic trinucleotides $N_1N_2N_3$ contain in total exactly 45 occurrences of each nucleotide. The number of occurrences of a nucleotide N in all the trinucleotides $N_1N_2N_3$ of X is

$$\mathrm{Nb}_N(X) := \sum_{S \in \mathcal{C}} \mathrm{Nb}_N(g(S)), \tag{6.1}$$

where $\mathrm{Nb}_N(N_1N_2N_3)$ stands for the number of occurrences of N in $N_1N_2N_3$. As each conjugacy class is composed of the circular shifts of a trinucleotide, all trinucleotides in a given conjugacy class S have the same number of occurrences of each given nucleotide N, which we write $\mathrm{Nb}_N(S)$. Consequently, (6.1) does not actually depend on g (that is, on X), and

$$\mathrm{Nb}_N(X) = \sum_{S \in \mathcal{C}} \mathrm{Nb}_N(S) = \frac{180}{4 \cdot 3} = 15 = \frac{3 \cdot |X|}{4},$$

where the second equality follows from the definition of the conjugacy classes.

Definition 6.1. A trinucleotide code X is *balanced* if for each nucleotide $N \in \mathcal{B}$ the number of occurrences of N in the trinucleotides of X is $\frac{3 \cdot |X|}{4}$.

The cardinality of a balanced trinucleotide code must be a multiple of 4.

In this section we analyse more deeply the balance property of the trinucleotide k-circular codes, by considering trinucleotide codes of cardinalities smaller than 20, and hence cardinality in $\{4, 8, 12, 16\}$.

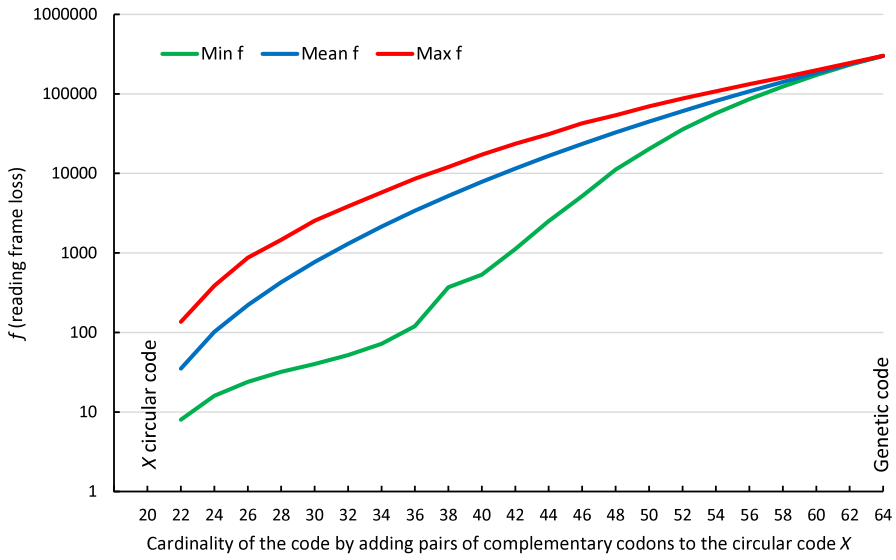


Fig. 5. Evolution of the trinucleotide circular code X (1.1) to the genetic code. The three curves represent the minimum Min, the mean $\bar{f}$ (5.2) and the maximum Max of the reading frame loss function f (5.1). The X-axis is the cardinality of the trinucleotide codes (even number between 22 and 64) and the Y-axis is the reading frame loss function f in a logarithmic scale.

Furthermore, we establish a new theoretical relation between the balance property and the classical biological property of self-complementarity. Indeed, if X is a self-complementary trinucleotide code then $\mathrm{Nb}_A(X) = \mathrm{Nb}_T(X)$ and $\mathrm{Nb}_C(X) = \mathrm{Nb}_G(X)$. It thus seems natural to study balanceness with respect to self-complementarity.

6.2. Method

We can use the algorithmic approaches developed in our companion article (Michel et al., 2022) to enumerate the trinucleotide k-circular codes that are balanced. However, we are interested also in trinucleotide 0-circular codes, for which the computations would take several weeks on a standard PC. Thus, we here develop a much quicker approach based on linear algebra, which allows us to count – and enumerate if desired – the number of trinucleotide ($\geqslant$0)-circular codes of any cardinality and with any prescribed number of occurrences of each nucleotide in total. We point out that the counting is essentially instantaneous for trinucleotide codes of any possible size (less than a second).

We here present the general method on an example, to avoid unnecessary abstraction. Assume that we want to enumerate the number of trinucleotide ($\geqslant$0)-circular codes X of cardinality n without a periodic trinucleotide and such that $\mathrm{Nb}_N(X) = n_N$ for each nucleotide $N \in \mathcal{B}$; so $n_A + n_C + n_G + n_T = 3n$. We partition the possibilities for X according to the number of trinucleotides contained in each of the conjugacy classes. To this end, let $\mathcal{C}_1, \ldots, \mathcal{C}_{20}$ be the conjugacy classes of the non-periodic trinucleotides. We associate to X the vector $\mathbf{v}_X := (|X \cap \mathcal{C}_1|, \ldots, |X \cap \mathcal{C}_{20}|)^t$, which is thus a vector of integers all between 0 and 3.

We associate to each conjugacy class $\mathcal{C}$ the vector $\mathbf{v}_{\mathcal{C}} := (\mathrm{Nb}_N(C))^t_{N\in\{A,C,G\}}$. For instance, the conjugacy class $\{AAC, ACA, CAA\}$ yields the vector $(2,1,0)^t$.

We can now write $\widetilde{\mathbf{M}} \cdot \mathbf{v}_X = \mathbf{b}$ where $\mathbf{b} := (n, n_A, n_C, n_G)^t$ and $\widetilde{\mathbf{M}}$ is the matrix with columns $(1, \mathbf{v}_{\mathcal{C}_i})^t$ for $i \in \{1, \ldots, 20\}$. If the conjugacy classes are enumerated in lexicographically increasing order regarding their lexicographic minimal element (so we have $\mathcal{C}_1 = \{AAC, ACA, CAA\}$ and $\mathcal{C}_{20} = \{GTT, TTG, TGT\}$), then we have $\widetilde{\mathbf{M}}$ given in Box I.

Conversely, every vector $\mathbf{v} = (v_1, \ldots, v_{20})^t$ satisfying $\widetilde{\mathbf{M}} \cdot \mathbf{v} = \mathbf{b}$ with v_i a non-negative integer at most 3 for each $i \in \{1, \ldots, 20\}$, corresponds to several different sought trinucleotide codes X. More precisely, each vector $\mathbf{v}$ is associated with exactly $\prod_{i=0}^{20} \binom{3}{v_i}$ different trinucleotide codes X; and, by definition, different such vectors cannot be associated with the same trinucleotide code. Let $\mathcal{S}$ be this set of specific solutions to our matrix equation.

Finding the set $\mathcal{S}$ is standard, and is immediate using any computer algebra system that can solve linear systems. From a theoretical point of view, we can first compute a basis of the kernel of the matrix $\widetilde{\mathbf{M}}$, and then any particular solution $\mathbf{s}$ to the matrix equation. The solutions to our matrix equation are then exactly the linear combinations of elements of the basis to which we add $\mathbf{s}$. We thus efficiently obtain a general form for the vectors that are solution. In any case, it is then straightforward to extract the vectors that belong to $\mathcal{S}$, which allows us to compute the aforementioned product for each of them, and thus the total number of sought trinucleotide codes.

The method can also be slightly adapted to allow for periodic trinucleotides, or tailored to self-complementary trinucleotide codes by using only 10 different conjugacy classes, for instance. If periodic trinucleotides are allowed then the matrix becomes as given in Box II and the last 4 variables $v_{21}, \ldots, v_{24}$ of $\mathbf{v}$ must each be either 0 or 1.

As for self-complementary trinucleotide codes, we obtain a balanced trinucleotide code X as soon as $\mathrm{Nb}_A(X) = \mathrm{Nb}_C(X)$ since the other equalities will follow by self-complementarity. Therefore, the vector $\mathbf{b}$ becomes $(m, n_A, n_C)^t$ where $m := n/2$ is half the size of the self-complementary trinucleotide code. The following matrices can be used when periodic trinucleotides are forbidden or allowed, respectively:

$$\widetilde{\mathbf{M}}^{\mathrm{sc}} := \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 3 & 1 & 1 & 1 & 2 & 2 & 1 & 0 \\ 1 & 1 & 0 & 2 & 2 & 2 & 1 & 1 & 2 & 3 \end{pmatrix}$$

or

$$\mathbf{M}^{\mathrm{sc}} := \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 3 & 1 & 1 & 1 & 2 & 2 & 1 & 0 & 3 & 0 \\ 1 & 1 & 0 & 2 & 2 & 2 & 1 & 1 & 2 & 3 & 0 & 3 \end{pmatrix}.$$

For example, the third column corresponds to the conjugacy class $\mathcal{C}_3 = \{AAT, ATA, TAA\}$: if w is a trinucleotide in $\mathcal{C}_3$, then its complementary trinucleotide $\overline{w}$ belongs to $\{ATT, TAT, TTA\}$. Therefore, w and $\overline{w}$ contain together 3 occurrences of the nucleotide A and 0 occurrence of the nucleotide C, hence the associated vector $(3,0)^t$.

The general process can be optimised by gathering conjugacy classes with the same associated vector, e.g. $\mathcal{C}_5 = \{ACG, CGA, GAC\}$ and $\mathcal{C}_7 = \{AGC, GCA, CAG\}$ are both associated with the vector $(1,1,1)^t$. We can thus blend the corresponding two variables into a single variable v'_5,

$$\widetilde{\mathbf{M}} := \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 2 & 2 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 2 & 1 & 0 & 1 & 0 & 1 & 0 & 2 & 1 & 1 & 0 & 2 & 1 \end{pmatrix}.$$

Box I.

$$\mathbf{M} := \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 3 & 0 & 0 & 0 \\ 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 1 & 0 & 0 & 2 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 2 & 2 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 2 & 1 & 0 & 1 & 0 & 1 & 0 & 2 & 1 & 1 & 0 & 2 & 1 & 0 & 0 & 0 & 3 \end{pmatrix}.$$

Box II.

which is then allowed to vary between 0 and 6; the corresponding contribution for the number of codes is $\binom{6}{v_5'}$. To illustrate this, we mention that the self-complementary case can be fully computed (with or without periodic trinucleotides) using only four variables and a single matrix, namely

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & 3 & 0 \\ 1 & 2 & 0 & 3 \end{pmatrix},$$

where the variables v_1 and v_2 range in $\{0, \dots, 12\}$, while the variables v_3 and v_4 range in $\{0, \dots, 4\}$ if the periodic trinucleotides are allowed and $\{0, \dots, 3\}$ otherwise.

6.3. Application

6.3.1. Case with periodic trinucleotides

We use the notation $N_{\geqslant 0}(n) = \binom{64}{n}$ and $N_{\geqslant 0}^{\mathrm{sc}}(m) = \binom{32}{m}$ already introduced in the companion article (Michel et al., 2022) for the number of trinucleotide ($\geqslant$0)-circular codes of length n and the number of self-complementary trinucleotide ($\geqslant$ 0)-circular codes of length $2m = n$, respectively. We furthermore let $N_{\geqslant 0}^{\mathrm{b}}(n)$ be the number of balanced trinucleotide codes of cardinality $n \in \{0, \dots, 64\}$, so $N_{\geqslant 0}^{\mathrm{b}}(n) = 0$ if $n \not\equiv 0 \pmod 4$. We also define $N_{\geqslant 0}^{\mathrm{sc,b}}(m)$ to be the number of balanced self-complementary trinucleotide codes of cardinality $2m = n$, and hence $N_{\geqslant 0}^{\mathrm{sc,b}}(m) \leq N_{\geqslant 0}^{\mathrm{b}}(2\,m)$ for $m \in \{0, \dots, 32\}$. Let X be a trinucleotide ($\geqslant$0)-circular code. As $\mathcal{B}^3$ is itself a balanced trinucleotide ($\geqslant$0)-circular code, we deduce that X is balanced if and only if $\mathcal{B}^3 \setminus X$ is balanced. Consequently, $N_{\geqslant 0}^{\mathrm{b}}(n) = N_{\geqslant 0}^{\mathrm{b}}(64-n)$ and, similarly, $N_{\geqslant 0}^{\mathrm{sc,b}}(m) = N_{\geqslant 0}^{\mathrm{sc,b}}(32-m)$. Table 3 gives the numbers $N_{\geqslant 0}^{b}(n)$ and $N_{\geqslant 0}^{\mathrm{sc,b}}(n/2)$ for $n \in \{4, 8, \dots, 60\}$, computed with the method presented in Section 6.2, specifically with the matrices $\mathbf{M}$ and $\mathbf{M}^{\mathrm{sc}}$. Combining with the numbers $N_{\geqslant 0}(n)$ and $N_{\geqslant 0}^{\mathrm{sc}}(n/2)$ from the companion article (Michel et al., 2022), the probability $p_{\geqslant 0}^{\mathrm{b}}(n)$ that a uniform random trinucleotide code of a given cardinality be balanced, can be determined as follows:

$$p_{\geqslant 0}^{\mathrm{b}}(n) := \frac{N_{\geqslant 0}^{\mathrm{b}}(n)}{N_{\geqslant 0}(n)}. \tag{6.2}$$

Similarly, the probability $p_{\geqslant 0}^{\mathrm{sc,b}}(m)$ that a uniform random self-complementary trinucleotide code of a given cardinality be balanced, can be determined as follows:

$$p_{\geqslant 0}^{\mathrm{sc,b}}(m) := \frac{N_{\geqslant 0}^{\mathrm{sc,b}}(m)}{N_{\geqslant 0}^{\mathrm{sc}}(m)}. \tag{6.3}$$

In order to evaluate the two properties, balanceness and self-complementarity, we define the ratio $r_{\geqslant 0}(n)$ between these two probabilities, that is

$$r_{\geqslant 0}(n) := \frac{p_{\geqslant 0}^{\mathrm{sc,b}}(n/2)}{p_{\geqslant 0}^{\mathrm{b}}(n)}. \tag{6.4}$$

6.3.2. Case without periodic trinucleotides

We use a similar approach, with analogous mathematical symbols labelled with a tilde for differentiation. In particular, $\widetilde{N}_{\geqslant 0}(n) = \binom{60}{n}$ and $\widetilde{N}_{\geqslant 0}^{\mathrm{sc}}(m) = \binom{30}{m}$. The numbers $\widetilde{N}_{\geqslant 0}^{\mathrm{b}}(n)$ and $\widetilde{N}_{\geqslant 0}^{\mathrm{sc,b}}(m)$ are obtained using the matrices $\widetilde{\mathbf{M}}$ and $\widetilde{\mathbf{M}}^{\mathrm{sc}}$ from Section 6.2. Table 4 presents the values obtained. The two probabilities $\tilde{p}_{\geqslant 0}^{\mathrm{b}}(n)$ and $\tilde{p}_{\geqslant 0}^{\mathrm{sc,b}}(m)$, and the ratio $\tilde{r}_{\geqslant 0}(n)$ are computed similarly as in Section 6.3.1. Let $X \subseteq \mathcal{B}^3 \setminus \mathcal{P}$ be a trinucleotide ($\geqslant$0)-circular code without a periodic trinucleotide. As $\mathcal{B}^3 \setminus \mathcal{P}$ is itself a balanced trinucleotide ($\geqslant$0)-circular code, we deduce that X is balanced if and only if $(\mathcal{B}^3 \setminus \mathcal{P}) \setminus X$ is balanced. Consequently, $\widetilde{N}_{\geqslant 0}^{\mathrm{b}}(n) = \widetilde{N}_{\geqslant 0}^{\mathrm{b}}(60-n)$ and, similarly, $\widetilde{N}_{\geqslant 0}^{\mathrm{sc,b}}(m) = \widetilde{N}_{\geqslant 0}^{\mathrm{sc,b}}(30-m)$.

6.3.3. Graphical representations of the numerical values in Tables 3 and 4

Fig. 6 gives a graphical representation of the numerical values in Tables 3 and 4. It shows that the probability of balanceness for both the trinucleotide codes and the self-complementary trinucleotide codes decreases as the cardinality of the code increases. Furthermore, the shapes of the curves are similar for the periodic and non-periodic cases. From an evolutionary point of view, the property of balanceness would have been stronger in primitive life, with trinucleotide codes of small cardinalities. Then, it becomes weaker after mutations in the trinucleotide codes, process leading to an increase of their cardinalities. However, the statistical behaviour of the decrease of the balanceness differs between the trinucleotide codes and the self-complementary trinucleotide codes, for both the periodic and the non-periodic cases. Indeed, Fig. 7 shows that the ratio $r_{\geqslant 0}(n)$ increases with the cardinality n for $n \leq 32$. For the cardinality 20, which is the maximal cardinality for trinucleotide ($\geqslant$1)-circular codes, the value is 47.89. Between cardinalities 4 and 20, this ratio is multiplied by 3.43. Similarly, for the case without periodic trinucleotide, the ratio $\tilde{r}_{\geqslant 0}(n)$ increases with the cardinality n of the trinucleotide code for $n \leq 30$. For the cardinality 20, the value is 39.98. Between cardinalities 4 and 20, this ratio is multiplied by 3.19. The difference between the values 3.43 (periodic case) and 3.19 (non-periodic case) might hint at a particular property of the periodic trinucleotides with respect to self-complementarity and balanceness, which could be further investigated in the future.

In summary, these results demonstrate a relation between the properties of self-complementarity and balanceness. Precisely, according to

Table 3
Case with periodic trinucleotides: numbers $N^{sc,b}_{\geqslant 0}(m)$ and $N^{b}_{\geqslant 0}(n)$, and probabilities $p^{sc,b}_{\geqslant 0}(m)$ (6.3) and $p^{b}_{\geqslant 0}(n)$ (6.2) of balanceness for self-complementary trinucleotide codes versus trinucleotide codes with cardinality $n = 2m$ in $\{4, 8, \ldots, 60\}$. The ratio $r_{\geqslant 0}(n)$ (6.4) gives a quantitative measure of balanceness between these two classes of trinucleotide codes. As mentioned in the companion article (Michel et al., 2022), the numbers $N^{sc}_{\geqslant 0}(m)$ and $N_{\geqslant 0}(n)$ are equal to $\binom{32}{m}$ and to $\binom{64}{n}$, respectively.

Card.	Self-complementary codes			Codes			Ratio
$n = 2m$	$N^{sc}_{\geqslant 0}(m)$	$N^{sc,b}_{\geqslant 0}(m)$	$p^{sc,b}_{\geqslant 0}(m)$(%)	$N_{\geqslant 0}(n)$	$N^{b}_{\geqslant 0}(n)$	$p^{b}_{\geqslant 0}(n)$(%)	$r_{\geqslant 0}(n)$
{4, 60}	496	160	32.26	635 376	14 688	2.31	13.95
{8, 56}	35 960	8456	23.52	4 426 165 368	42 048 456	0.95	24.75
{12, 52}	906 192	181 376	20.02	3 284 214 703 056	19 253 443 632	0.59	34.14
{16, 48}	10 518 300	1 903 490	18.10	488 526 937 079 580	2 111 538 087 534	0.43	41.87
{20, 44}	64 512 240	10 925 696	16.94	19 619 725 782 651 120	69 390 367 780 296	0.35	47.89
{24, 40}	225 792 840	36 649 008	16.23	250 649 105 469 666 120	779 622 128 266 488	0.31	52.18
{28, 36}	471 435 600	74 716 224	15.85	1 118 770 292 985 239 888	3 237 736 351 419 828	0.29	54.76
{32}	601 080 390	94 532 308	15.73	1 832 624 140 942 590 534	5 181 557 395 735 824	0.28	55.62

Table 4
Case without periodic trinucleotides: numbers $\widetilde{N}^{sc,b}_{\geqslant 0}(m)$ and $\widetilde{N}^{b}_{\geqslant 0}(n)$, and probabilities $\tilde{p}^{sc,b}_{\geqslant 0}(m)$ (Section 6.3.2) and $\tilde{p}^{b}_{\geqslant 0}(n)$ (Section 6.3.2) of balanceness for self-complementary trinucleotide codes versus trinucleotide codes with no periodic trinucleotide and with cardinality $n = 2m$ in $\{4, 8, \ldots, 56\}$. The ratio $\tilde{r}_{\geqslant 0}(n)$ (Section 6.3.2) gives a quantitative measure of balanceness between these two classes of trinucleotide codes. As mentioned in the companion article (Michel et al., 2022), the numbers $\widetilde{N}^{sc}_{\geqslant 0}(m)$ and $\widetilde{N}_{\geqslant 0}(n)$ are equal to $\binom{30}{m}$ and to $\binom{60}{n}$, respectively.

Card.	Self-complementary codes			Codes			Ratio
$n = 2m$	$\widetilde{N}^{sc}_{\geqslant 0}(m)$	$\widetilde{N}^{sc,b}_{\geqslant 0}(m)$	$\tilde{p}^{sc,b}_{\geqslant 0}(m)$(%)	$\widetilde{N}_{\geqslant 0}(n)$	$\widetilde{N}^{b}_{\geqslant 0}(n)$	$\tilde{p}^{b}_{\geqslant 0}(n)$(%)	$\tilde{r}_{\geqslant 0}(n)$
{4, 56}	435	153	35.17	487 635	13 689	2.81	12.53
{8, 52}	27 405	6981	25.47	2 558 620 845	30 286 845	1.18	21.52
{12, 48}	593 775	128 501	21.64	1 399 358 844 975	10 384 658 505	0.74	29.16
{16, 44}	5 852 925	1 147 389	19.60	149 608 375 854 525	829 956 638 277	0.55	35.34
{20, 40}	30 045 015	5 531 229	18.41	4 191 844 505 805 495	19 301 198 755 293	0.46	39.98
{24, 36}	86 493 225	15 331 173	17.73	36 052 387 482 172 425	1 483 39 543 503 821	0.41	43.08
{28, 32}	145 422 675	25 318 293	17.41	103 719 945 525 634 515	404 636 393 455 353	0.39	44.63

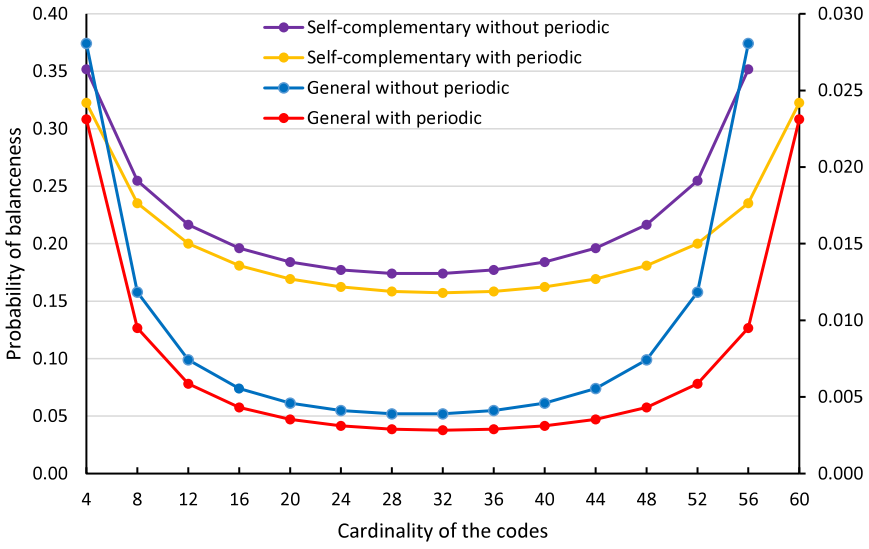


Fig. 6. Probabilities $p^{sc,b}_{\geqslant 0}(m)$ (6.3) and $p^{b}_{\geqslant 0}(n)$ (6.2), respectively $\tilde{p}^{sc,b}_{\geqslant 0}(m)$ (Section 6.3.2) and $\tilde{p}^{b}_{\geqslant 0}(n)$ (Section 6.3.2), of balanceness for self-complementary trinucleotide codes versus trinucleotide codes, respectively with and without periodic trinucleotides and cardinality $n = 2m$ in $\{4, 8, \ldots, 60\}$ and in $\{4, 8, \ldots, 56\}$. The probabilities $p^{sc,b}_{\geqslant 0}(m)$ and $\tilde{p}^{sc,b}_{\geqslant 0}(m)$ are represented on the left y-axis. The probabilities $p^{b}_{\geqslant 0}(m)$ and $\tilde{p}^{b}_{\geqslant 0}(m)$ are represented on the right y-axis. The red and yellow curves are symmetric with respect to the cardinality 32 and the blue and violet ones are symmetric with respect to the cardinality 30.

these quantitative evaluations, the self-complementarity of the trinucleotide codes decreases the balanceness loss occurring when their cardinalities increase during evolution.

7. Hierarchy of the trinucleotide k-circular codes

In Section 6 and Figure 4 of an earlier work (Fimmel et al., 2020) was proposed an evolutionary hypothesis of the genetic code based on a growing combinatorial hierarchy of trinucleotide codes with circularity k, where $k \in \{0, 1, 2, 3, 4\}$. Fig. 8 updates Figure 4 from this work (Fimmel et al., 2020) as the minimum and maximum sizes of trinucleotide k-circular codes and their numbers are determined here for $k \in \{1, 2, 3\}$ (see Table 1 in the companion article (Michel et al., 2022)). As the minimum sizes of trinucleotide 3- and 2-circular codes are 4 and 5 trinucleotides, respectively, codes in the primitive soup for constructing the modern standard genetic code with less than 4 trinucleotides could not be 3- or 2-circular. Furthermore, as the maximum sizes of trinucleotide 3- and 2-circular codes are 18 and 20 trinucleotides, respectively, the trinucleotide 3-circular codes would be more primitive than the trinucleotide 2-circular codes. These two observations agree with the evolutionary model of the genetic code proposed earlier (Fimmel et al., 2020).

Evolution would have started with the trinucleotide ($\geqslant 4$)-circular (circular) codes in X_p with an increasing complexity according to the

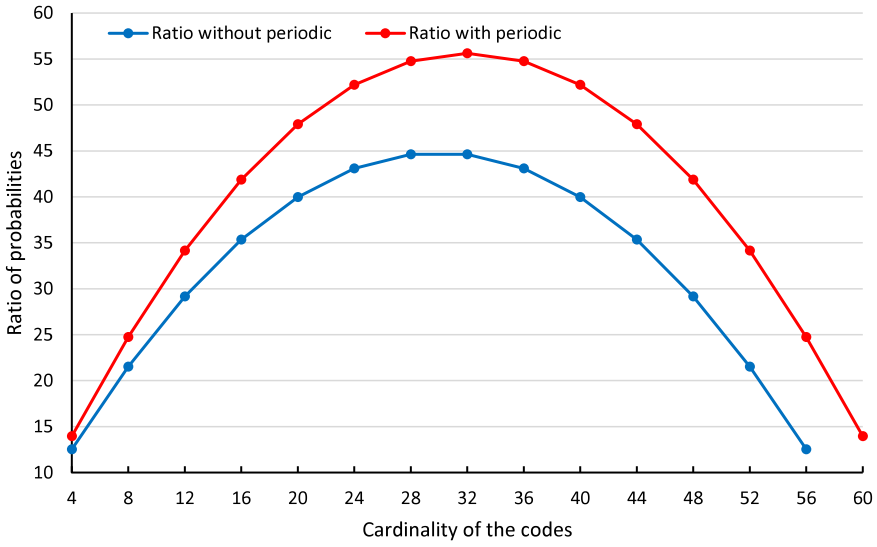


Fig. 7. Ratios $r_{\geqslant 0}(n)$ (6.4) and $\tilde{r}_{\geqslant 0}(n)$ (Section 6.3.2) giving a quantitative measure of balanceness between the self-complementary trinucleotide codes and the trinucleotide codes, respectively with and without periodic trinucleotides and cardinality n in $\{4, 8, \ldots, 60\}$ and in $\{4, 8, \ldots, 56\}$. The red curve is symmetric with respect to the cardinality 32 and the blue one is symmetric with respect to the cardinality 30.

maximal path length p (from 1 to 8) in their associated graph. As the maximal path length p is related to the window nucleotide length of reading frame retrieval, the circular codes in X_1 (strong comma-free) or in X_2 (comma-free) are more constrained than those in X_8. The maximal C^3-self-complementary trinucleotide circular code X observed in genes (1.1) belongs to the class X_8. Then evolution continued with the three classes of k-circular codes, where $k \in \{1, 2, 3\}$, which are less constrained than the classes of circular codes as they have a partial circularity (see Section 4). Only the maximal trinucleotide 1-circular codes of 20 trinucleotides can code 20 amino acids: 52 out of $3,473,671,209$ trinucleotide 1-circular codes have this property (see Appendix II in Fimmel et al., 2020). Evolution from the trinucleotide 1-circular codes to the genetic code of cardinality 64 can be achieved by the trinucleotide 0-circular codes which have a partial circularity and a cardinality that can be greater than 20 and up to 64.

We have recently proposed a model of coevolution of genetic code and translation system in four stages (Dila et al., 2019b). Based on different statistical analyses of universal X motifs (motifs from the circular code X defined in (1.1) that are conserved in the rRNA of more than 90% of the organisms) in the proto-ribosome and the accretion model of ribosome evolution (Hsiao et al., 2009; Petrov et al., 2015), we have suggested that the trinucleotide 4-circular codes (comma-free codes and circular codes) represented ancestors of the modern genetic code and were used to map the first trinucleotides to amino acids. The first encoding system was based on a comma-free code, such as $\{GCC, GGC\}$, which would have allowed encoding of the amino acids and the reading frame within a single code. With the addition of new amino acids, comma-free codes were no longer viable and were replaced by the circular codes. For example, there exists a non maximal circular code that is a subset of X, containing exactly 10 trinucleotides and capable of coding 8 (Michel et al., 2017) out of the 10 hypothesised "early amino acids" (Koonin, 2017). The maximal trinucleotide circular code X (size 20) can code only 12 amino acids (Arquès and Michel, 1996). There are exactly $12,964,440$ maximal trinucleotide circular codes (Arquès and Michel, 1996; Michel and Pirillo, 2010). No maximal trinucleotide circular code among these $12,964,440$ codes 20 or 19 amino acids with the standard genetic code SGC, only 10 maximal circular codes code 18 amino acids with the SGC (Michel and Pirillo, 2013, Introduction). So the circular code property was not sufficient when more amino acids were needed in the standard and variant genetic codes. With the evolution of the genetic code, the replacement of 4-circular by $1, 2, 3$-circular codes and finally by 0-circular codes, i.e. codes in general terms, has allowed for a greater flexibility in the coding of amino acids in the different genetic codes, but, as a result, less ability to retrieve the reading frame. The weakening of the property to retrieve the reading frame has led to the emergence of complex start codon recognition systems with a specific start codon that initiates translation, and sophisticated ratchet mechanisms for maintaining the reading frame during translation elongation.

8. Amino acids coded by the trinucleotide k-circular codes

We have computed the number of amino acids coded by the trinucleotide k-circular codes where $k \in \{2, 3\}$ according to the standard genetic code. As we shall see in the forthcoming subsections, from an amino acid coding point of view the maximum number of amino acids coded by the 3-circular codes is 16. The maximum number of amino acids coded by the 2-circular codes is 17. Furthermore, the number 429 ($183 + 183 + 58 + 5$; see Section 8.1.2) of 2-circular codes coding 17 amino acids is much larger than the number 8 ($6+2$; see Section 8.1.1) of 3-circular codes coding 16 amino acids. These observations might suggest that the 2-circular codes appeared in the course of evolution after the 3-circular codes.

The maximum number of amino acids coded by the self-complementary trinucleotide 2- and 3-circular codes is identical for both classes and equal to 14. However, the number 26 ($4 + 17 + 5$; see Section 8.2.2) of self-complementary trinucleotide 2-circular codes coding 14 amino acids is larger than the number 3 ($1+2$; see Section 8.2.1) of the self-complementary trinucleotide 3-circular codes, which might suggest again that the 2-circular codes would have appeared after the 3-circular codes.

Finally, the trinucleotide $(3, 3, 3)$-circular codes exist only for length 10. We do verify below that 8 of these 96 codes code 10 amino acids. However, due to these very specific combinatorial properties, it seems very unlikely that such codes would have been a step in the evolution process of the genetic code. We also note that the maximum number of amino acids coded by the $(2, 2, 2)$-circular codes is 15, compared to 10 for the $(3, 3, 3)$-circular codes, an additional argument that the 2-circular codes would have appeared after the 3-circular codes.

8.1. Amino acids coded by the trinucleotide k-circular codes

The growth function of the trinucleotide k-circular codes is given in Table 1 of the companion article (Michel et al., 2022), allowing the readers to retrieve the corresponding numbers.

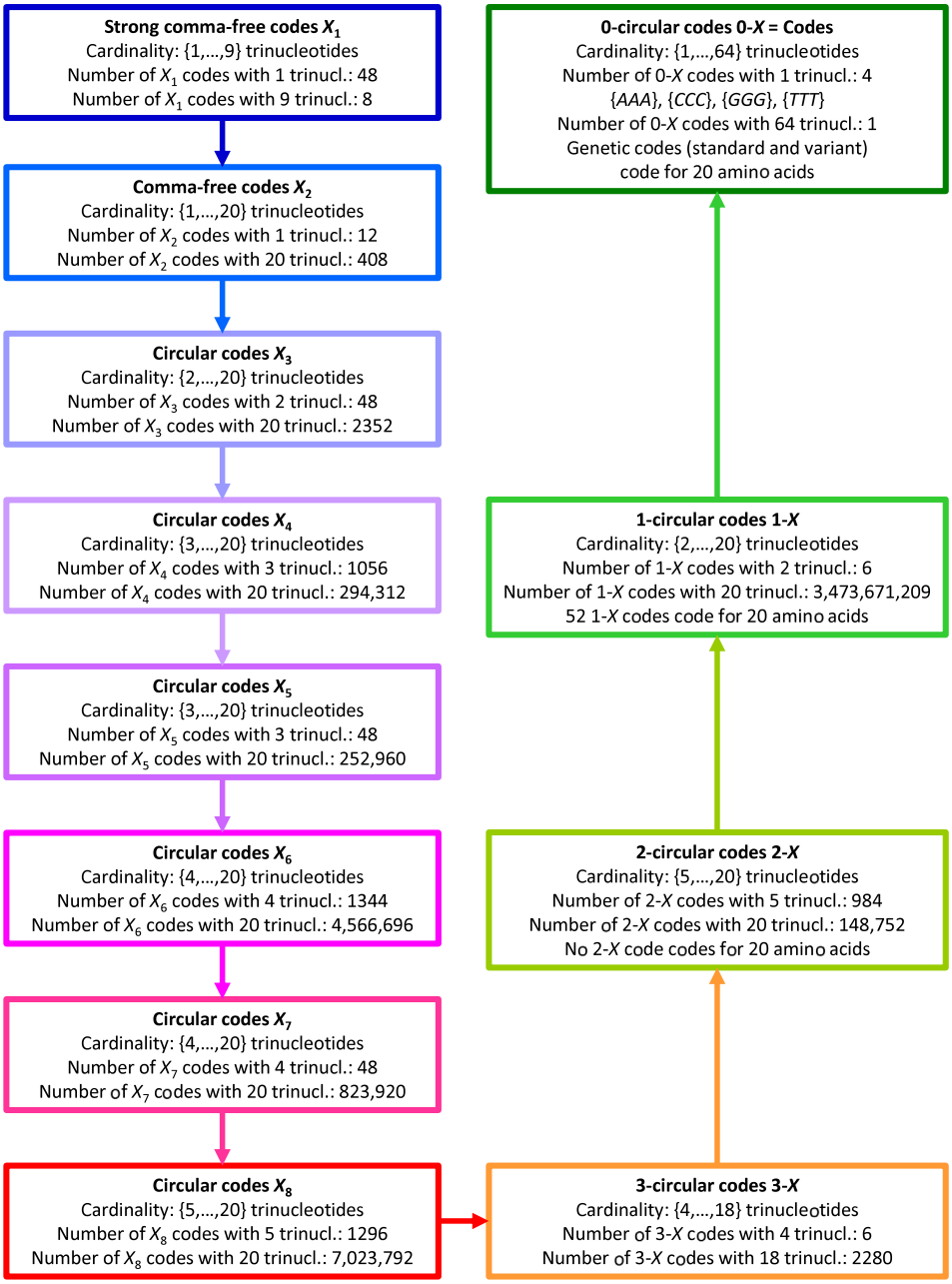


Fig. 8. A combinatorial hierarchy of the trinucleotide k-circular codes, where $k \in \{0, 1, 2, 3, 4\}$. The hierarchy of the trinucleotide circular ($(\geqslant 4)$-circular) codes in X_p is given as a function of the maximal path length p in the associated graph.

8.1.1. Trinucleotide 3-circular codes

Among the 6 minimum trinucleotide 3-circular codes of length 4 (see List 5.8 in Michel et al., 2022), 3 code 3 amino acids and 3 code 4 amino acids.

Observation 8.1. *The maximum number of amino acids coded by a trinucleotide 3-circular code is* $M_3 = 16$.

The number M_3 is already obtained with a code of length 16. Items (1)–(3) and Lists 8.2, 8.3 and 8.4 complete these observations.

(1) 6 trinucleotide 3-circular codes of length 16 (among 788820) code the maximum number 16 of amino acids (see List 8.2).
(2) 2 trinucleotide 3-circular codes of length 17 (among 83520) code the maximum number 16 of amino acids (see List 8.3).
(3) 7 maximum trinucleotide 3-circular codes of length 18 (among 2280) code the largest number 15 of amino acids (see List 8.4).

We point out that no maximum trinucleotide 3-circular code of length 18 codes M_3 amino acids. Indeed, not all trinucleotide 3-circular codes of length 16 or 17 are contained in a trinucleotide 3-circular code of length 18.

List 8.2 (*The* 6 *Trinucleotide* 3*-circular Codes of Length* 16 *(Among* 788820*) Coding the Maximum Number* $M_3 = 16$ *of Amino Acids*)**.**

$\{AAC, AAG, ATA, ATG, CAC, CAG, CCT, CGG, GAC, GAG, GGT, GTA, TCG, TGC, TTA, TTC\}$,
$\{AAC, AAG, ATA, ATG, CAC, CAG, CCT, CGT, GAC, GAG, GCT, GGC, GTA, TCT, TGG, TTA\}$,
$\{AAC, AAG, ATA, ATG, CAC, CAG, CCT, CGT, GAC, GAG, GCT, GGC, GTA, TGG, TTA, TTC\}$,
$\{AAC, AAG, ATA, ATG, CAC, CAG, CGT, GAC, GAG, GCT,$

$GGC, GTA, TCC, TGG, TTA, TTC\}$,
$\{AAC, AAG, ATC, ATG, CAC, CAG, CCG, CGT, CTC, GAC,$
$GCT, GGA, GTT, TAC, TGG, TTC\}$,
$\{AAC, AAG, ATC, ATG, CAC, CAG, CCT, CGT, GAC, GAG,$
$GCT, GGC, GTA, TGG, TTA, TTC\}$.

List 8.3 (*The* 2 *Trinucleotide* 3*-circular Codes of Length* 17 *(Among* 83520*) Coding the Maximum Number* $M_3 = 16$ *of Amino Acids*)**.**

$\{AAC, AAG, ATA, ATC, ATG, CAC, CAG, CCT, CGT, GAC,$
$GAG, GCT, GGC, GTA, TGG, TTA, TTC\}$,
$\{AAC, AAG, ATC, ATG, CAC, CAG, CCG, CGT, CTC, GAC,$
$GCT, GGA, GTT, TAC, TAG, TGG, TTC\}$.

List 8.4 (*The* 7 *Maximum Trinucleotide* 3*-circular Codes of Length* 18 *(Among* 2280*) Coding the Largest Number* $M_3 - 1 = 15$ *of Amino Acids*)**.**

$\{AAC, AAG, AAT, ACG, CAG, CAT, CCT, CGG, CTA, GAG,$
$GAT, GCC, GTA, GTC, TGC, TGG, TTA, TTC\}$,
$\{AAC, AAG, AAT, ACG, CAG, CAT, CCT, CGG, CTA, GAG,$
$GAT, GCC, GTC, TGC, TGG, TGT, TTA, TTC\}$,
$\{AAC, AAG, ACT, ATA, CAG, CCA, CCT, CGC, GAG, GAT,$
$GGC, GTA, GTT, TCG, TCT, TGC, TGG, TTA\}$,
$\{AAC, AAG, ATA, ATC, ATG, ATT, CAG, CCA, CGC, CTT,$
$GAC, GAG, GGC, GGT, GTT, TAC, TCC, TGC\}$,
$\{AAC, AAG, ATC, ATG, ATT, CAC, CAG, CCG, CCT, CTT,$
$GAC, GCG, GGA, GTA, GTT, TAC, TCG, TGG\}$,
$\{AAC, AAG, ATC, ATG, ATT, CAC, CAG, CCG, CCT, CTT,$
$GAC, GCG, GGA, GTT, TAC, TAG, TCG, TGG\}$,
$\{AAC, AAG, ATC, ATG, ATT, CAG, CCA, CGC, CTT, GAC,$
$GAG, GGC, GGT, GTT, TAC, TAG, TCC, TGC\}$.

8.1.2. Trinucleotide 2-circular codes

Among the 984 minimum trinucleotide 2-circular codes of length 5, there are 50 coding 3 amino acids, 381 coding 4 amino acids and 553 coding 5 amino acids.

Observation 8.5. *The maximum number of amino acids coded by a trinucleotide 2-circular code is* $M_2 = 17$.

The number M_2 is already obtained with a code of length 17. Items (1)–(4) and List 8.6 complete these observations.

(1) 183 trinucleotide 2-circular codes of length 17 (among 142169112) code the maximum number 17 of amino acids.
(2) 183 trinucleotide 2-circular codes of length 18 (among 27843072) code the maximum number 17 of amino acids.
(3) 58 trinucleotide 2-circular codes of length 19 (among 3104832) code the maximum number 17 of amino acids.
(4) 5 maximum trinucleotide 2-circular codes of length 20 (among 148752) code the maximum number 17 of amino acids (see List 8.6).

We point out that the 183 trinucleotide codes in Item (1) are all contained in the 183 trinucleotide codes in Item (2).

List 8.6 (*The* 5 *Maximum Trinucleotide* 2*-circular Codes of Length* 20 *(Among* 148752*) Coding the Maximum Number* $M_2 = 17$ *of Amino Acids*)**.**

$\{AAC, AAG, AAT, ACC, CAG, CAT, CCT, CGC, GAC, GAG,$
$GAT, GGC, GTA, GTT, TAC, TCG, TGC, TGG, TTA, TTC\}$,
$\{AAC, AAG, AAT, ACC, CAG, CAT, CCT, CGC, GAC, GAG,$
$GAT, GGC, GTT, TAC, TAG, TCG, TGC, TGG, TTA, TTC\}$,
$\{AAC, AAG, AAT, ACC, CAG, CAT, CCT, CGC, GAC, GAG,$
$GGC, GTA, GTT, TAC, TCG, TGA, TGC, TGG, TTA, TTC\}$,
$\{AAC, AAG, AAT, ACC, CAG, CAT, CCT, CGC, GAC, GAG,$
$GGC, GTT, TAC, TAG, TCG, TGA, TGC, TGG, TTA, TTC\}$,
$\{AAG, AGC, AGG, AGT, ATA, ATC, ATG, CAA, CAC, CCG,$
$CCT, GAC, GCT, GGC, GTC, TAC, TGG, TGT, TTA, TTC\}$.

Finally, it was (mathematically) established earlier (Fimmel et al., 2020) that there are exactly 52 maximum trinucleotide 1-circular codes of length 20 (among 3473671209) coding for 20 amino acids (see the list in Appendix II in Fimmel et al., 2020). We verified that all these 52 trinucleotide codes actually are $(1, 1, 1)$-circular.

8.2. Amino acids coded by the self-complementary trinucleotide k-circular codes

The growth function of the self-complementary trinucleotide k-circular codes is given in Table 4 of the companion article (Michel et al., 2022), allowing the readers to retrieve the corresponding numbers.

8.2.1. Self-complementary trinucleotide 3-circular codes

Among the 4 minimum self-complementary trinucleotide 3-circular codes of length 4 (see List 5.21 in Michel et al., 2022), 1 codes 3 amino acids and 3 code 4 amino acids.

Observation 8.7. *The maximum number of amino acids coded by a self-complementary trinucleotide 3-circular code is* $M_3^{sc} = 14$.

The number M_3^{sc} is already obtained with a code of length 14. Items (1)–(2) and Lists 8.8 and 8.9 complete these observations.

(1) 1 self-complementary trinucleotide 3-circular code of length 14 (among 464) codes the maximum number 14 of amino acids (see List 8.8).
(2) 2 self-complementary trinucleotide 3-circular codes of length 16 (among 80) code the maximum number 14 of amino acids (see List 8.9).

List 8.8 (*The Unique Self-complementary Trinucleotide* 3*-circular Code of Length* 14 *(Among* 464*) Coding the Maximum Number* $M_3^{sc} = 14$ *of Amino Acids*)**.**

$\{ACG, CGT, AGC, GCT, ATC, GAT, CAA, TTG, CCA, TGG,$
$GAA, TTC, GTA, TAC\}$.

List 8.9 (*The* 2 *Maximum Self-complementary Trinucleotide* 3*-circular Codes of Length* 16 *(Among* 80*) Coding the Maximum Number* $M_3^{sc} = 14$ *of Amino Acids*)**.**

$\{ACG, CGT, AGC, GCT, AGG, CCT, ATC, GAT, CAA, TTG,$
$CCA, TGG, GAA, TTC, GTA, TAC\}$,
$\{ACG, CGT, AGC, GCT, ATC, GAT, CAA, TTG, CCA, TGG,$
$CTC, GAG, GAA, TTC, GTA, TAC\}$.

8.2.2. Self-complementary trinucleotide 2-circular codes

Among the 8 minimum self-complementary trinucleotide 2-circular codes of length 6 (see List 5.24 in Michel et al., 2022), 1 codes 4 amino acids, 4 code 5 amino acids and 3 code 6 amino acids.

Observation 8.10. *The maximum number of amino acids coded by a self-complementary trinucleotide 2-circular code is* $M_2^{sc} = 14$.

The number M_2^{sc} is already obtained with a code of length 14. Items (1)–(4) and Lists 8.11, 8.12 and 8.13 complete these observations.

(1) 4 self-complementary 2-circular codes of length 14 (among 1704) code the maximum number 14 of amino acids (see List 8.11).
(2) 17 self-complementary trinucleotide 2-circular codes of length 16 (among 780) code the maximum number 14 of amino acids.
(3) 5 self-complementary trinucleotide 2-circular code of length 18 (among 176) code the maximum number 14 of amino acids (see List 8.12).
(4) 1 maximum self-complementary trinucleotide 2-circular code of length 20 (among 16) codes the largest number 13 of amino acids (see List 8.13).

We point out that no maximum self-complementary trinucleotide 2-circular code of length 20 codes $M_2^{sc} = 14$ amino acids.

List 8.11 (*The* 4 *Self-complementary Trinucleotide* 2*-circular Codes of Length* 14 *(Among* 1704*) Coding the Maximum Number* $M_2^{sc} = 14$ *of Amino Acids*)**.**

$\{ACA, TGT, ATC, GAT, CAG, CTG, CGA, TCG, GAA, TTC,$
$GCC, GGC, GTA, TAC\},$
$\{ACG, CGT, ATC, GAT, CAA, TTG, CCA, TGG, GAA, TTC,$
$GCC, GGC, GTA, TAC\},$
$\{ACT, AGT, AGG, CCT, ATG, CAT, CAA, TTG, GAA, TTC,$
$GAC, GTC, GCA, TGC\},$
$\{ACT, AGT, ATG, CAT, CAA, TTG, CCG, CGG, GAA, TTC,$
$GAC, GTC, GCA, TGC\}.$

List 8.12 (*The* 5 *Self-complementary Trinucleotide* 2*-circular Codes of Length* 18 *(Among* 176*) Coding the Maximum Number* $M_2^{sc} = 14$ *of Amino Acids*)**.**

$\{ACT, AGT, AGG, CCT, ATC, GAT, CAA, TTG, CAC, GTG,$
$GAA, TTC, GAC, GTC, GCC, GGC, TAA, TTA\},$
$\{ACT, AGT, AGG, CCT, ATG, CAT, CAA, TTG, CCG, CGG,$
$GAA, TTC, GAC, GTC, GCA, TGC, TAA, TTA\},$
$\{ACT, AGT, AGG, CCT, ATG, CAT, CCA, TGG, CCG, CGG,$
$GAA, TTC, GAC, GTC, GCA, TGC, TAA, TTA\},$
$\{ACT, AGT, ATG, CAT, CAA, TTG, CCG, CGG, CTC, GAG,$
$GAA, TTC, GAC, GTC, GCA, TGC, TAA, TTA\},$
$\{ACT, AGT, ATG, CAT, CCA, TGG, CCG, CGG, CTC, GAG,$
$GAA, TTC, GAC, GTC, GCA, TGC, TAA, TTA\}.$

List 8.13 (*The Unique Maximum Self-complementary Trinucleotide* 2*-circular Code of Length* 20 *(Among* 16*) Coding the Largest Number* $M_2^{sc} - 1 = 13$ *of Amino Acids*)**.**

$\{AAC, GTT, AAG, CTT, AAT, ATT, CAC, GTG, CAG, CTG,$
$CTC, GAG, GAC, GTC, GCC, GGC, GTA, TAC, TCA, TGA\}.$

8.3. Amino acids coded by the trinucleotide (k, k, k)-codes

The growth function of the trinucleotide (k, k, k)-circular codes is given in Table 6 of the companion article (Michel et al., 2022), allowing the readers to retrieve the corresponding numbers.

8.3.1. Trinucleotide $(3, 3, 3)$-circular codes

All the trinucleotide $(3, 3, 3)$-circular codes have length 10.

Observation 8.14. *The maximum number of amino acids coded by a trinucleotide* $(3, 3, 3)$*-circular code is* $M_{(3,3,3)} = 10$.

Among the 96 trinucleotide $(3, 3, 3)$-circular codes, there are 4 coding 6 amino acids, 14 coding 7 amino acids, 45 coding 8 amino acids, 25 coding 9 amino acids, and 8 coding $M_{(3,3,3)} = 10$ amino acids (see List 8.15).

List 8.15 (*The* 8 *Trinucleotide* $(3, 3, 3)$*-circular Codes of Length* 10 *(Among* 96*) Coding the Maximum Number* $M_{(3,3,3)} = 10$ *of Amino Acids*)**.**

$\{AAC, ACG, ATA, CAT, CCT, GAG, GCC, GTA, TGC, TGG\},$
$\{AAG, ACC, ATA, CAT, CCT, GAC, GCG, GTA, TGC, TGG\},$
$\{ACG, AGA, ATT, CAA, CAT, GCG, GGT, GTA, TGC, TTC\},$
$\{ACG, AGT, ATA, CAA, CAT, GCG, GGA, GTT, TGC, TTC\},$
$\{ATA, ATG, CAA, CCG, CGT, GAG, GCA, GGT, TAC, TCC\},$
$\{ATA, ATG, CAG, CCA, CGT, GAA, GCG, GGT, TAC, TCC\},$
$\{ATG, ATT, CAC, CCG, CGT, GCA, GGA, TAC, TCT, TGG\},$
$\{ATG, ATT, CAC, CCT, CGG, GCA, GGA, TAC, TCG, TGT\}.$

8.3.2. Trinucleotide $(2, 2, 2)$-circular codes

Among the 72 minimum trinucleotide $(2, 2, 2)$-circular codes of length 6, there are 6 coding 4 amino acids, 33 coding 5 amino acids and 33 coding 6 amino acids.

Observation 8.16. *The maximum number of amino acids coded by a trinucleotide* $(2, 2, 2)$*-circular code is* $M_{(2,2,2)} = 15$.

The number $M_{(2,2,2)}$ is already obtained with a code of length 15. Items (1)–(5) and Lists 8.17 and 8.18 complete these observations.

(1) 4 trinucleotide $(2, 2, 2)$-circular codes of length 15 (among 224832) code the maximum number 15 of amino acids (see List 8.17).
(2) 74 trinucleotide $(2, 2, 2)$-circular codes of length 16 (among 55620) code the largest number 14 of amino acids.
(3) 59 trinucleotide $(2, 2, 2)$-circular codes of length 17 (among 12312) code the largest number 14 of amino acids.
(4) 34 trinucleotide $(2, 2, 2)$-circular codes of length 18 (among 1944) code the largest number 14 of amino acids.
(5) 8 trinucleotide $(2, 2, 2)$-circular codes of length 19 (among 144) code the largest number 14 of amino acids (see List 8.18).

We point out that no trinucleotide $(2, 2, 2)$-circular code of length greater than 15 codes the maximum number $M_{(2,2,2)} = 15$ of amino acids.

List 8.17 (*The* 4 *Trinucleotide* $(2, 2, 2)$*-circular Codes of Length* 15 *(Among* 224832*) Coding the Maximum Number* $M_{(2,2,2)} = 15$ *of Amino Acids*)**.**

$\{AAG, AAT, ACT, AGG, ATC, CAG, CCA, GAC, GGC, GTA,$
$TAT, TCC, TGG, TGT, TTC\},$
$\{AAG, AAT, ACT, AGG, CAA, CCA, CTG, GAC, GCG, GTA,$
$TAT, TCA, TGG, TGT, TTC\},$
$\{AAG, AAT, ACT, ATC, CAG, CCA, CGT, GAC, GAG, GCC,$
$GTG, TAT, TCC, TTC, TTG\},$
$\{AAT, AGC, ATG, ATT, CAA, CAC, CCT, CGT, CTT, GAC,$
$GAG, GCG, GTA, TAC, TGG\}.$

List 8.18 (*The* 8 *Maximum Trinucleotide* $(2, 2, 2)$*-circular Codes of Length* 19 *(Among* 144*) Coding the Largest Number* $M_{(2,2,2)} - 1 = 14$ *of Amino Acids*)**.**

$\{AAC, AAG, AAT, ACC, ACT, AGT, ATT, CAT, CCT, CGA,$
$CGT, CTT, GAG, GAT, GCC, GCG, GCT, GGT, GTT\},$
$\{AAC, AAG, AAT, ACC, AGC, AGG, ATC, GAC, GCC, GGC,$

$GTA, GTC, TAC, TAT, TCC, TGC, TGG, TGT, TTC\}$,
$\{AAC, AAG, AAT, AGC, AGG, ATC, CAC, CGC, CTC, GAC,$
$GGC, GTA, GTC, TAC, TAT, TGC, TGG, TGT, TTC\}$,
$\{AAC, AAG, AAT, AGC, ATC, ATT, CAC, CGC, CTC, GAC,$
$GAG, GGC, GTC, GTG, GTT, TAC, TGA, TGC, TTC\}$,
$\{AAC, ACC, AGA, AGC, AGG, ATA, ATC, GAC, GAT, GCC,$
$GGC, GTC, TAC, TCC, TGC, TGG, TTA, TTC, TTG\}$,
$\{AAC, AGC, ATC, ATG, CAC, CGC, CTC, GAA, GAC, GGA,$
$GGC, GGT, GTC, TAA, TAC, TAT, TGC, TGT, TTC\}$,
$\{AAT, ACG, ACT, AGT, ATT, CAA, CAC, CAT, CCT, CGC,$
$CGT, CTT, GAA, GAT, GCT, GGA, GGC, GGT, GTT\}$,
$\{AAT, ACT, AGC, AGT, ATT, CAA, CAT, CCA, CCG, CCT,$
$CGT, CTT, GAA, GAG, GAT, GCG, GCT, GGT, GTT\}$.

8.4. Amino acids coded by the self-complementary trinucleotide (k, k, k)-codes

The growth function of the self-complementary trinucleotide (k, k, k)-circular codes is given in Table 7 of the companion article (Michel et al., 2022), allowing the readers to retrieve the corresponding numbers.

There is no self-complementary trinucleotide $(3, 3, 3)$-circular code.

Among the 96 minimum self-complementary trinucleotide $(2, 2, 2)$-circular codes of length 10, there are 1 coding 5 amino acids, 6 coding 6 amino acids, 17 coding 7 amino acids, 28 coding 8 amino acids, 29 coding 9 amino acids and 15 coding 10 amino acids.

Observation 8.19. *The maximum number of amino acids coded by a self-complementary trinucleotide $(2, 2, 2)$-circular code is $M^{sc}_{(2,2,2)} = 12$.*

The number $M^{sc}_{(2,2,2)}$ is already obtained with a code of length 12. Items (1)–(3) and Lists 8.20, 8.21 and 8.22 complete these observations.

(1) 1 self-complementary trinucleotide $(2, 2, 2)$-circular code of length 12 (among 184) codes the maximum number 12 of amino acids (see List 8.20).
(2) 4 self-complementary trinucleotide $(2, 2, 2)$-circular codes of length 14 (among 56) code the maximum number 12 of amino acids (see List 8.21).
(3) 1 self-complementary trinucleotide $(2, 2, 2)$-circular code of length 16 (among 4) codes the largest number 10 of amino acids (see List 8.22).

We point out that no maximum self-complementary trinucleotide $(2, 2, 2)$-circular code of length 16 codes the maximum number $M^{sc}_{(2,2,2)} = 12$ of amino acids.

List 8.20 (*The Unique Self-complementary Trinucleotide $(2, 2, 2)$-circular Code of Length 12 (Among 184) Coding the Maximum Number $M^{sc}_{(2,2,2)} = 12$ of Amino Acids*).

$\{ACG, CGT, ATC, GAT, CAA, TTG, CCA, TGG, GAA, TTC,$
$GCC, GGC\}$.

List 8.21 (*The 4 Self-complementary Trinucleotide $(2, 2, 2)$-circular Codes of Length 14 (Among 56) Coding the Maximum Number $M^{sc}_{(2,2,2)} = 12$ of Amino Acids*).

$\{AAG, CTT, AAT, ATT, ACA, TGT, CTC, GAG, GAC, GTC,$
$GCC, GGC, TCA, TGA\}$,
$\{AAG, CTT, AAT, ATT, CCA, TGG, CTC, GAG, GAC, GTC,$
$GCC, GGC, TCA, TGA\}$,
$\{AAT, ATT, ACA, TGT, ACT, AGT, CCA, TGG, CCG, CGG,$
$CTC, GAG, GAC, GTC\}$,
$\{AAT, ATT, ACA, TGT, ACT, AGT, CCA, TGG, CCG, CGG,$
$GAA, TTC, GAC, GTC\}$.

List 8.22 (*The Unique Maximum Self-complementary Trinucleotide $(2, 2, 2)$-circular Code of Length 16 (Among 4) Coding the Largest Number $M^{sc}_{(2,2,2)} - 2 = 10$ of Amino Acids*).

$\{AAC, GTT, AAG, CTT, AAT, ATT, CAC, GTG, CAG, CTG,$
$CTC, GAG, GAC, GTC, TCA, TGA\}$.

9. Conclusion

The theory of trinucleotide k-circular codes developed in the companion article (Michel et al., 2022), has open several new biological fields studied in this work.

A method was proposed to determine the ambiguous sequences from a trinucleotide k-circular code. It also hinted at classifying the genetic sequences into three classes: (i) sequences with reading frame retrieval; (ii) sequences with ambiguous frame; and (iii) sequences without frame (frameless). Furthermore, this approach applied to the different classes of trinucleotide k-circular codes led to new properties for determining the reading frame of a genetic sequence as a function of its trinucleotide length.

In contrast to the classical view in the circular code theory, we showed that the circularity property, i.e. the property of reading frame retrieval, is an ordinary property in genes as almost all the $2^{64} \approx 10^{19}$ trinucleotide codes have a partial circularity (except the empty set and the 15 trinucleotide codes identified in Proposition 4.2). In particular "random" trinucleotides codes have a partial circularity. The complete circularity is achieved with the $115,606,988,558 \approx 10^{11}$ trinucleotide circular codes. For coding the 20 amino acids, life could have constructed an alphabet of 20 (different) nucleotides in bijection with the 20 amino acids, avoiding thus the problem of reading frame retrieval. Due to chemical reasons, this mathematical structure was not selected. Thus, a reduced alphabet of only 4 nucleotides has required codes with words of length greater than 1, e.g. trinucleotide codes, that automatically has led to a process of reading frame retrieval.

A new formula is derived to measure the reading frame loss in the trinucleotide k-circular codes. It ranges from 0 with a circular code to 301056 with the genetic code. Furthermore, it allowed, for the first time, to develop a model of evolution from a trinucleotide code to the genetic code, i.e. an evolution of trinucleotide codes of cardinality greater than 20.

Three properties are identified in the evolution of primitive codes to the genetic code: the two classical properties of circularity and self-complementarity, and the new property of trinucleotide code balance. A method based on linear algebra is proposed to compute the balanced trinucleotide codes, in the general case and in the self-complementarity case. The definition of a probability ratio based on the numbers of trinucleotides codes that are balanced or not, showed that the self-complementarity of the trinucleotide codes decreases the balanceness loss occurring when their cardinalities increase during evolution.

The hierarchy of the trinucleotide k-circular codes is updated according to the growth functions obtained.

Finally, the numbers of amino acids coded by the different classes of trinucleotide k-circular codes are determined. All results converge to the evolutionary hypothesis that the 2-circular codes would have appeared after the 3-circular codes.

We point out that the concept of trinucleotide k-circular codes could be interesting to construct genes such that their translation is efficient, i.e. simultaneously without reading frame error and fast, for example by replacing motifs that lose the reading frame with motifs that retrieve it. It could be applied to bioengineering research as recently mentioned (see Section 2.5 of a recent work by Štambuk, Konjevoda and Pavan (Štambuk et al., 2021)).

Cardinality 22: $\{AAA,TTT\}:\{\{2,2\}\}$, $\{AAG,CTT\}:\{\{2,2\}\}$, $\{AGC,GCT\}:\{\{2,2\}\}$, $\{CAC,GTG\}:\{\{2,2\}\}$, $\{CCC,GGG\}:\{\{2,2\}\}$, $\{GGA,TCC\}:\{\{2,2\}\}$.

Cardinality 24: $\{AAA,TTT,AGC,GCT\}:\{\{2,4\}\}$, $\{AAA,TTT,CAC,GTG\}:\{\{2,4\}\}$, $\{AAA,TTT,CCC,GGG\}:\{\{2,4\}\}$, $\{AAA,TTT,GGA,TCC\}:\{\{2,4\}\}$,
$\{AAG,CTT,AGC,GCT\}:\{\{2,4\}\}$, $\{AAG,CTT,CAC,GTG\}:\{\{2,4\}\}$, $\{AAG,CTT,CCC,GGG\}:\{\{2,4\}\}$, $\{AAG,CTT,GGA,TCC\}:\{\{2,4\}\}$,
$\{AGC,GCT,CAC,GTG\}:\{\{2,4\}\}$, $\{AGC,GCT,CCC,GGG\}:\{\{2,4\}\}$, $\{AGC,GCT,GGA,TCC\}:\{\{2,4\}\}$, $\{CAC,GTG,CCC,GGG\}:\{\{2,4\}\}$,
$\{CAC,GTG,GGA,TCC\}:\{\{2,4\}\}$, $\{CCC,GGG,GGA,TCC\}:\{\{2,4\}\}$.

Cardinality 26: $\{AAA,TTT,AGC,GCT,CAC,GTG\}:\{\{2,6\}\}$, $\{AAA,TTT,AGC,GCT,CCC,GGG\}:\{\{2,6\}\}$, $\{AAA,TTT,AGC,GCT,GGA,TCC\}:\{\{2,6\}\}$,
$\{AAA,TTT,CAC,GTG,CCC,GGG\}:\{\{2,6\}\}$, $\{AAA,TTT,CAC,GTG,GGA,TCC\}:\{\{2,6\}\}$, $\{AAA,TTT,CCC,GGG,GGA,TCC\}:\{\{2,6\}\}$,
$\{AAG,CTT,AGC,GCT,CAC,GTG\}:\{\{2,6\}\}$, $\{AAG,CTT,AGC,GCT,CCC,GGG\}:\{\{2,6\}\}$, $\{AAG,CTT,AGC,GCT,GGA,TCC\}:\{\{2,6\}\}$,
$\{AAG,CTT,CAC,GTG,CCC,GGG\}:\{\{2,6\}\}$, $\{AAG,CTT,CAC,GTG,GGA,TCC\}:\{\{2,6\}\}$, $\{AAG,CTT,CCC,GGG,GGA,TCC\}:\{\{2,6\}\}$,
$\{AGC,GCT,CAC,GTG,CCC,GGG\}:\{\{2,6\}\}$, $\{AGC,GCT,CAC,GTG,GGA,TCC\}:\{\{2,6\}\}$, $\{AGC,GCT,CCC,GGG,GGA,TCC\}:\{\{2,6\}\}$,
$\{CAC,GTG,CCC,GGG,GGA,TCC\}:\{\{2,6\}\}$.

Cardinality 28: $\{AAA,TTT,AGC,GCT,CAC,GTG,CCC,GGG\}:\{\{2,8\}\}$, $\{AAA,TTT,AGC,GCT,CAC,GTG,GGA,TCC\}:\{\{2,8\}\}$, $\{AAA,TTT,AGC,GCT,CCC,GGG,GGA,TCC\}:\{\{2,8\}\}$,
$\{AAA,TTT,CAC,GTG,CCC,GGG,GGA,TCC\}:\{\{2,8\}\}$, $\{AAG,CTT,AGC,GCT,CAC,GTG,CCC,GGG\}:\{\{2,8\}\}$, $\{AAG,CTT,AGC,GCT,CAC,GTG,GGA,TCC\}:\{\{2,8\}\}$,
$\{AAG,CTT,AGC,GCT,CCC,GGG,GGA,TCC\}:\{\{2,8\}\}$, $\{AAG,CTT,CAC,GTG,CCC,GGG,GGA,TCC\}:\{\{2,8\}\}$, $\{AGC,GCT,CAC,GTG,CCC,GGG,GGA,TCC\}:\{\{2,8\}\}$.

Cardinality 30: $\{AAA,TTT,AGC,GCT,CAC,GTG,CCC,GGG,GGA,TCC\}:\{\{2,10\}\}$, $\{AAG,CTT,AGC,GCT,CAC,GTG,CCC,GGG,GGA,TCC\}:\{\{2,10\}\}$.

Cardinality 32: $\{AAG,CTT,AGC,GCT,ATA,TAT,CAC,GTG,CCC,GGG,GGA,TCC\}:\{\{2,12\},\{4,2\}\}$.

Cardinality 34: $\{AAG,CTT,AGC,GCT,ATA,TAT,CAC,GTG,CCC,GGG,GCA,TGC,GGA,TCC\}:\{\{2,16\},\{4,4\}\}$.

Cardinality 36: $\{AAA,TTT,AGC,GCT,ATA,TAT,CAC,GTG,CCC,GGG,GCA,TGC,GGA,TCC,TAA,TTA\}:\{\{2,20\},\{4,20\}\}$.

Cardinality 38: $\{AAG,CTT,ACT,AGT,ATG,CAT,CAA,TTG,CAC,GTG,CCC,GGG,CCG,CGG,CGC,GCG,CTA,TAG\}:\{\{2,22\},\{4,80\},\{6,74\},\{8,23\}\}$.

Cardinality 40: $\{AAG,CTT,ACT,AGT,AGC,GCT,AGG,CCT,ATG,CAT,CAA,TTG,CAC,GTG,CCC,GGG,CCG,CGG,CGC,GCG\}:\{\{2,22\},\{4,113\},\{6,130\},\{8,48\}\}$.

Cardinality 42: $\{AAG,CTT,ACG,CGT,ACT,AGT,AGC,GCT,AGG,CCT,ATG,CAT,CAA,TTG,CAC,GTG,CCC,GGG,CCG,CGG,CGC,GCG\}:\{\{2,24\},\{4,157\},\{6,356\},\{8,233\}\}$.

Cardinality 44: $\{AAG,CTT,ACG,CGT,ACT,AGT,AGC,GCT,AGG,CCT,ATG,CAT,CAA,TTG,CAC,GTG,CCC,GGG,CCG,CGG,CGC,GCG,CTA,TAG\}:\{\{2,28\},\{4,204\},\{6,776\},\{8,960\}\}$.

Cardinality 46: $\{AAA,TTT,AAG,CTT,ACG,CGT,ACT,AGT,AGC,GCT,AGG,CCT,ATG,CAT,CAA,TTG,CAC,GTG,CCC,GGG,CCG,CGG,CGC,GCG,CTA,TAG\}$:
$\{\{2,30\},\{4,252\},\{6,1362\},\{8,2754\}\}$.

Cardinality 48: $\{AAA,TTT,AAG,CTT,ACG,CGT,ACT,AGT,AGC,GCT,AGG,CCT,ATA,TAT,ATG,CAT,CAA,TTG,CAC,GTG,CCC,GGG,CCG,CGG,CGC,GCG,CTA,TAG\}$:
$\{\{2,32\},\{4,319\},\{6,2280\},\{8,7380\}\}$.

Cardinality 50: $\{AAA,TTT,AAG,CTT,ACG,CGT,ACT,AGT,AGC,GCT,AGG,CCT,ATA,TAT,ATG,CAT,CAA,TTG,CAC,GTG,CCA,TGG,CCC,GGG,CCG,CGG,CGC,GCG,CTA,TAG\}$:
$\{\{2,36\},\{4,421\},\{6,3650\},\{8,14473\}\}$,
$\{AAA,TTT,AAG,CTT,ACG,CGT,ACT,AGT,AGC,GCT,AGG,CCT,ATA,TAT,ATG,CAT,CAA,TTG,CAC,GTG,CCC,GGG,CCG,CGG,CGC,GCG,CTA,TAG,GGA,TCC\}$:
$\{\{2,36\},\{4,421\},\{6,3650\},\{8,14473\}\}$.

Cardinality 52: $\{AAA,TTT,AAG,CTT,ACG,CGT,ACT,AGT,AGC,GCT,AGG,CCT,ATA,TAT,ATG,CAT,CAA,TTG,CAC,GTG,CCA,TGG,CCC,GGG,CCG,CGG,CGA,TCG,CGC,GCG,CTA,TAG\}:\{\{2,40\},\{4,538\},\{6,5528\},\{8,27104\}\}$,
$\{AAA,TTT,AAG,CTT,ACG,CGT,ACT,AGT,AGC,GCT,AGG,CCT,ATA,TAT,ATG,CAT,CAA,TTG,CAC,GTG,CCC,GGG,CCG,CGG,CGC,GCG,CTA,TAG,GCA,TGC,GGA,TCC\}:\{\{2,40\},\{4,538\},\{6,5528\},\{8,27104\}\}$.

Cardinality 54: $\{AAA,TTT,AAG,CTT,ACA,TGT,ACG,CGT,ACT,AGT,AGC,GCT,AGG,CCT,ATA,TAT,ATG,CAT,CAA,TTG,CAC,GTG,CCA,TGG,CCC,GGG,CCG,CGG,CGC,GCG,CTA,TAG,GCA,TGC\}:\{\{2,44\},\{4,650\},\{6,7680\},\{8,45408\}\}$,
$\{AAA,TTT,AAG,CTT,ACG,CGT,ACT,AGT,AGA,TCT,AGC,GCT,AGG,CCT,ATA,TAT,ATG,CAT,CAA,TTG,CAC,GTG,CCC,GGG,CCG,CGG,CGA,TCG,CGC,GCG,CTA,TAG,GGA,TCC\}:\{\{2,44\},\{4,650\},\{6,7680\},\{8,45408\}\}$.

Box III.

Cardinality 56: $\{AAA, TTT, AAG, CTT, ACA, TGT, ACG, CGT, ACT, AGT, AGC, GCT, AGG, CCT, ATA, TAT, ATG, CAT, CAA, TTG, CAC, GTG, CCA, TGG, CCC, GGG, CCG, CGG, CGA, TCG, CGC, GCG, CTA, TAG, GCA, TGC\} : \{\{2, 48\}, \{4, 789\}, \{6, 10458\}, \{8, 70153\}\}$,

$\{AAA, TTT, AAG, CTT, ACG, CGT, ACT, AGT, AGA, TCT, AGC, GCT, AGG, CCT, ATA, TAT, ATG, CAT, CAA, TTG, CAC, GTG, CCC, GGG, CCG, CGG, CGA, TCG, CGC, GCG, CTA, TAG, GCA, TGC, GGA, TCC\} : \{\{2, 48\}, \{4, 789\}, \{6, 10458\}, \{8, 70153\}\}$.

Cardinality 58: $\{AAA, TTT, AAG, CTT, ACA, TGT, ACG, CGT, ACT, AGT, AGA, TCT, AGC, GCT, AGG, CCT, ATA, TAT, ATG, CAT, CAA, TTG, CAC, GTG, CCA, TGG, CCC, GGG, CCG, CGG, CGA, TCG, CGC, GCG, CTA, TAG, GCA, TGC\} : \{\{2, 52\}, \{4, 923\}, \{6, 13580\}, \{8, 103285\}\}$,

$\{AAA, TTT, AAG, CTT, ACA, TGT, ACG, CGT, ACT, AGT, AGA, TCT, AGC, GCT, AGG, CCT, ATA, TAT, ATG, CAT, CAA, TTG, CAC, GTG, CCC, GGG, CCG, CGG, CGA, TCG, CGC, GCG, CTA, TAG, GCA, TGC, GGA, TCC\} : \{\{2, 52\}, \{4, 923\}, \{6, 13580\}, \{8, 103285\}\}$.

Cardinality 60: $\{AAA, TTT, AAG, CTT, ACA, TGT, ACG, CGT, ACT, AGT, AGA, TCT, AGC, GCT, AGG, CCT, ATA, TAT, ATG, CAT, CAA, TTG, CAC, GTG, CCA, TGG, CCC, GGG, CCG, CGG, CGA, TCG, CGC, GCG, CTA, TAG, GCA, TGC, TAA, TTA\} : \{\{2, 56\}, \{4, 1084\}, \{6, 17472\}, \{8, 146640\}\}$,

$\{AAA, TTT, AAG, CTT, ACA, TGT, ACG, CGT, ACT, AGT, AGA, TCT, AGC, GCT, AGG, CCT, ATA, TAT, ATG, CAT, CAA, TTG, CAC, GTG, CCC, GGG, CCG, CGG, CGA, TCG, CGC, GCG, CTA, TAG, GCA, TGC, GGA, TCC, TAA, TTA\} : \{\{2, 56\}, \{4, 1084\}, \{6, 17472\}, \{8, 146640\}\}$,

$\{AAA, TTT, AAG, CTT, ACA, TGT, ACG, CGT, AGA, TCT, AGC, GCT, ATA, TAT, ATG, CAT, CAA, TTG, CAC, GTG, CCA, TGG, CCC, GGG, CCG, CGG, CGA, TCG, CGC, GCG, CTA, TAG, GCA, TGC, GGA, TCC, TAA, TTA, TCA, TGA\} : \{\{2, 56\}, \{4, 1084\}, \{6, 17472\}, \{8, 146640\}\}$.

Cardinality 62: $\{AAA, TTT, AAG, CTT, ACA, TGT, ACG, CGT, ACT, AGT, AGA, TCT, AGC, GCT, AGG, CCT, ATA, TAT, ATG, CAT, CAA, TTG, CAC, GTG, CCA, TGG, CCC, GGG, CCG, CGG, CGA, TCG, CGC, GCG, CTA, TAG, GCA, TGC, GGA, TCC, TAA, TTA\} : \{\{2, 60\}, \{4, 1260\}, \{6, 21952\}, \{8, 199472\}\}$,

$\{AAA, TTT, AAG, CTT, ACA, TGT, ACG, CGT, AGA, TCT, AGC, GCT, AGG, CCT, ATA, TAT, ATG, CAT, CAA, TTG, CAC, GTG, CCA, TGG, CCC, GGG, CCG, CGG, CGA, TCG, CGC, GCG, CTA, TAG, GCA, TGC, GGA, TCC, TAA, TTA, TCA, TGA\} : \{\{2, 60\}, \{4, 1260\}, \{6, 21952\}, \{8, 199472\}\}$.

Cardinality 64: $\{AAA, TTT, AAG, CTT, ACA, TGT, ACG, CGT, ACT, AGT, AGA, TCT, AGC, GCT, AGG, CCT, ATA, TAT, ATG, CAT, CAA, TTG, CAC, GTG, CCA, TGG, CCC, GGG, CCG, CGG, CGA, TCG, CGC, GCG, CTA, TAG, GCA, TGC, GGA, TCC, TAA, TTA, TCA, TGA\} : \{\{2, 64\}, \{4, 1440\}, \{6, 26880\}, \{8, 262080\}\}$.

Box III. (*continued*).

Declaration of competing interest

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

Appendix. List of self-complementary additions to X minimising the reading frame loss

List of pairs of complementary codons added to the trinucleotide circular code X (1.1) up to the genetic code X_g, minimising the reading frame loss function f (Eq. (5.1)). After the trinucleotide 0-circular code, the type $2, 4, 6, 8$ and the number of corresponding directed cycles in the associated graph are given as cardinality 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64 in Box III.

References

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

Dila, G., Michel, C.J., Poch, O., Ripp, R., Thompson, J.D., 2019a. Evolutionary conservation and functional implications of circular code motifs in eukaryotic genomes. Biosystems 175, 57–74.

Dila, G., Ripp, R., Mayer, C., Poch, O., Michel, C.J., Thompson, J.D., 2019b. Circular code motifs in the ribosome: a missing link in the evolution of translation? RNA 25, 1714–1730.

Fimmel, E., Michel, C.J., Pirot, F., Sereni, J.-S., Starman, M., Strüngmann, L., 2020. The relation between k-circularity and circularity of codes. Bull. Math. Biol. 82, 105, 1–34.

Fimmel, E., Michel, C.J., Strüngmann, L., 2016. n-Nucleotide circular codes in graph theory. Phil. Trans. R. Soc. A 374, 20150058, 1–19.

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

Gumbel, E., Wiedemann, P., 2021. Motif lengths of circular codes in coding sequences. J. Theoret. Biol. 523, 110708, 1–9.

Hsiao, C., Mohan, S., Kalahar, B.K., Williams, L.D., 2009. Peeling the onion: Ribosomes are ancient molecular fossils. Mol. Biol. Evol. 26, 2415–2425.

Koonin, E.V., 2017. Frozen accident pushing 50: Stereochemistry, expansion, and chance in the evolution of the genetic code. Life 7, 22.

Michel, C.J., 2008. A 2006 review of circular codes in genes. Comput. Math. Appl. 55, 984–988.

Michel, C.J., 2015. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. J. Theoret. Biol. 380, 156–177.

Michel, C.J., 2017. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, archaea, eukaryotes, plasmids and viruses. Life 7 (2), 1–16.

Michel, C.J., 2020. The maximality of circular codes in genes statistically verified. Biosystems 197, 104201, 1–7.

Michel, C.J., Mouillon, B., Sereni, J.-S., 2022. Trinucleotide k-circular codes I: Theory. Biosystems 104667, http://dx.doi.org/10.1016/j.biosystems.2022.104667.

Michel, C.J., Nguefack Ngoune, V., Poch, O., Ripp, R., Thompson, J.D., 2017. Enrichment of circular code motifs in the genes of the yeast *Saccharomyces cerevisiae*. Life 7 (52), 1–20.

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

Michel, C.J., Pirillo, G., 2013. A permuted set of a trinucleotide circular code coding the 20 amino acids in variant nuclear codes. J. Theoret. Biol. 319, 116–121.

Petrov, A.S., Gulen, B., Norris, A.M., Kovacs, N.A., Bernier, C.R., Lanier, K.A., Fox, G.E., Harvey, S.C., Wartell, R.M., Hud, N.V., Williams, L.D., 2015. History of the ribosome and the origin of translation. Proc. Natl. Acad. Sci. USA 112, 15396–15401.

Štambuk, N., Konjevoda, P., Pavan, J., 2021. Antisense peptide technology for diagnostic tests and bioengineering research. Int. J. Mol. Sci. 22 (17), 9106.