Theory in Biosciences (2018) 137:51–65
https://doi.org/10.1007/s12064-018-0259-4

ORIGINAL ARTICLE

Self-complementary circular codes in coding theory

Elena Fimmel[1] · Christian J. Michel[2] · Martin Starman[1] · Lutz Strüngmann[1]

Received: 11 July 2017 / Accepted: 10 February 2018 / Published online: 12 March 2018

Abstract
Self-complementary circular codes are involved in pairing genetic processes. A maximal C^3 self-complementary circular code X of trinucleotides was identified in genes of bacteria, archaea, eukaryotes, plasmids and viruses (Michel in Life 7(20):1–16 2017, J Theor Biol 380:156–177, 2015; Arquès and Michel in J Theor Biol 182:45–58 1996). In this paper, self-complementary circular codes are investigated using the graph theory approach recently formulated in Fimmel et al. (Philos Trans R Soc A 374:20150058, 2016). A directed graph $\mathcal{G}(X)$ associated with any code X mirrors the properties of the code. In the present paper, we demonstrate a necessary condition for the self-complementarity of an arbitrary code X in terms of the graph theory. The same condition has been proven to be sufficient for codes which are circular and of large size $|X| \geq 18$ trinucleotides, in particular for maximal circular codes ($|X| = 20$ trinucleotides). For codes of small-size $|X| \leq 16$ trinucleotides, some very rare counterexamples have been constructed. Furthermore, the length and the structure of the longest paths in the graphs associated with the self-complementary circular codes are investigated. It has been proven that the longest paths in such graphs determine the reading frame for the self-complementary circular codes. By applying this result, the reading frame in any arbitrary sequence of trinucleotides is retrieved after at most 15 nucleotides, i.e., 5 consecutive trinucleotides, from the circular code X identified in genes. Thus, an X motif of a length of at least 15 nucleotides in an arbitrary sequence of trinucleotides (not necessarily all of them belonging to X) uniquely defines the reading (correct) frame, an important criterion for analyzing the X motifs in genes in the future.

Keywords Self-complementary circular codes · Graph properties · Translation process · Reading frame · Genetic code

Introduction

There is a consensus of opinion that the standard genetic code conserves vestiges of earlier, simpler codes, that may have been used to code fewer amino acids than the modern set of 20. Many examples of such ancient genetic codes have been proposed, including the trinucleotide codes *RRY* of size 8 (Crick et al. 1976) and *RNY* of size 16 ($R = \{A, G\}$, $Y = \{C, T\}$, $N = \{A, C, G, T\}$) (Eigen and Schuster 1978; Shepherd 1981), the trinucleotide codes *GNC* of size 4 and *SNS* of size 16 ($S = \{C, G\}$) (Ikehara 2002), *GHN* of size 12 ($H = \{A, C, T\}$) (Trifonov 1987), etc. Among the trinucleotide codes proposed, some of them have the important property to be circular. A circular code X is a set of words such that any motif (sequence) from X, called *X circular code motif* or more simply *X motif*, allows to retrieve, maintain and synchronize the reading (correct, original) frame. All the previously mentioned trinucleotide codes are circular, with the exception of the code *SNS* (as, for example, the periodic trinucleotide $CCC \in SNS$). The codes *RRY*, *RNY*, *GNC* and *GHN* also belong to the more restrictive class of comma-free codes. (The longest path length in their associated graphs is 2, definition given in "Trinucleotide circular codes and their associated graphs" section.) The code *RRY* is in addition strong comma-free. (The longest path length in

✉ Christian J. Michel
c.michel@unistra.fr

Elena Fimmel
e.fimmel@hs-mannheim.de

Martin Starman
m.starman@live.com

Lutz Strüngmann
l.struengmann@hs-mannheim.de

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

[2] Theoretical Bioinformatics, ICube, CNRS, University of Strasbourg, 300 Boulevard Sébastien Brant, 67400 Illkirch, France

its associated graph is 1, definition given in "Trinucleotide circular codes and their associated graphs" section.) A very few trinucleotide circular codes have in addition the property of self-complementarity, i.e., each trinucleotide in the code is complementary to another trinucleotide in the code. The comma-free codes RRY and GHN are not self-complementary (as $\overleftarrow{c(RRY)} = RYY \notin RRY$ and $\overleftarrow{c(GHN)} = NDC \notin GHN$ with $D = \{A, G, T\}$, definition of reversed complemented given in "Self-complementarity as a graph property" section). The comma-free codes RNY and GNC are self-complementary (as $\overleftarrow{c(RNY)} = RNY$ and $\overleftarrow{c(GNC)} = GNC$).

Furthermore, a maximal C^3 self-complementary trinucleotide circular code X was identified in genes of bacteria, archaea, eukaryotes, plasmids and viruses (Michel 2017, 2015; Arquès and Michel 1996). It contains the following 20 trinucleotides

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

In this paper, we study the self-complementary circular codes which are involved in pairing genetic processes. For the first time, all the self-complementary circular codes with words of 3 letters (trinucleotides) on a 4-letter alphabet (genetic alphabet) are identified (see their growth function given in Table 1), i.e., not just the few cases mentioned above, and several new mathematical properties are proven. Thus, genetic properties, in particular amino acid coding, of any self-complementary circular code can be investigated in the future. The paper is structured as follows.

After recalling the definition of a graph associated with a trinucleotide code in "Trinucleotide circular codes and their associated graphs" section and the theorem of acyclic graph of a trinucleotide code which is circular, we demonstrate in "Self-complementarity as a graph property" section that a code X is self-complementary if and only if its graph $\mathcal{G}(X)$ has a self-complementary set of vertices and for any vertex v, the outgoing degree $d^+(v)$ equals the ingoing degree $d^-(\overleftarrow{c(v)})$ of the complementary vertex. It is shown that this statement is true for the self-complementary circular codes of sizes 18 and 20 trinucleotides and for the self-complementary comma-free codes of sizes 14 and 16 trinucleotides. (There are no self-complementary comma-free codes of sizes 18 and 20 trinucleotides.)

In "Longest paths in the graphs associated with self-complementary circular codes" section, we investigate the length of longest paths in the graphs $\mathcal{G}(X)$ associated with self-complementary circular codes X. The longest path lengths (maximal arrow-length of paths) belong to the set $\{1, 2, 3, 4, 6, 8\}$ for the self-complementary circular codes and to the set $\{4, 6, 8\}$ for the class of 528 maximal (of size 20 trinucleotides) self-complementary circular codes. The growth function of all self-complementary circular codes of cardinality $n = 2, 4, \ldots, 20$ as a function of the longest path length $l = 1, 2, \ldots, 8$ is given. We also determine the structure of the longest paths for the self-complementary circular codes.

In "The reading frame of circular codes" section, we prove that the longest paths in such graphs $\mathcal{G}(X)$ determine the reading frame for the self-complementary circular codes X.

By applying this result in "Application: Reading frame of the maximal C^3 self-complementary circular code X identified in genes" section, the reading frame in any arbitrary sequence of trinucleotides is retrieved after at most 15 nucleotides, i.e., 5 consecutive trinucleotides, from the circular code X identified in genes. Thus, any X motif of length at least 5 trinucleotides located anywhere in a gene made of a series of any trinucleotide from the 64 possible ones (i.e., not necessarily all of them belonging to X) defines uniquely the reading (correct) frame. In this line of direction, very recent results have shown an enrichment of X motifs in the genes of the yeast *Saccharomyces cerevisiae* (Michel et al. 2017).

Trinucleotide circular codes and their associated graphs

In this section, we recall some notations and results from Fimmel et al. (2016). Let $\mathcal{B} = \{A, C, G, T\}$ be the set of nucleotides, where A stands for *adenine*, C stands for *cytosine*, G stands for *guanine*, and T stands for *thymine*. A *trinucleotide code* is a subset $X \subseteq \mathcal{B}^3$. The following definition relates a directed graph to any trinucleotide code. Recall that a *graph* $\mathcal{G}$ consists of a finite set of *vertices (nodes)* V and a finite set of *edges* E, where an edge is a set $\{v, w\}$ of vertices from V. The graph is called *oriented or directed* if the edges have an orientation, i.e., the edges are considered to be ordered pairs $[v, w]$ (for more details see, for example, Clark and Holton 1991). We now recall the graph theory approach from Fimmel et al. (2016).

Definition 2.1 (Definition 2.1, Fimmel et al. 2016). Let $X \subseteq \mathcal{B}^3$ be a trinucleotide code. We associate a directed graph $\mathcal{G}(X) = (V(X), E(X))$ with X, with set of vertices $V(X)$ and set of edges $E(X)$ as follows

- $V(X) = \{N_1, N_3, N_1N_2, N_2N_3 : N_1N_2N_3 \in X\}$,
- $E(X) = \{[N_1, N_2N_3], [N_1N_2, N_3] : N_1N_2N_3 \in X\}$.

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

The graph $\mathcal{G}(X)$ associated with the code X was used in Fimmel et al. (2016) in order to characterize the *circular codes* among the trinucleotide codes. Recall that a trinucleotide code $X \subseteq \mathcal{B}^3$ is said to be a *circular code* if for any concatenation $x_1 \cdots x_m$ of trinucleotides from X there is

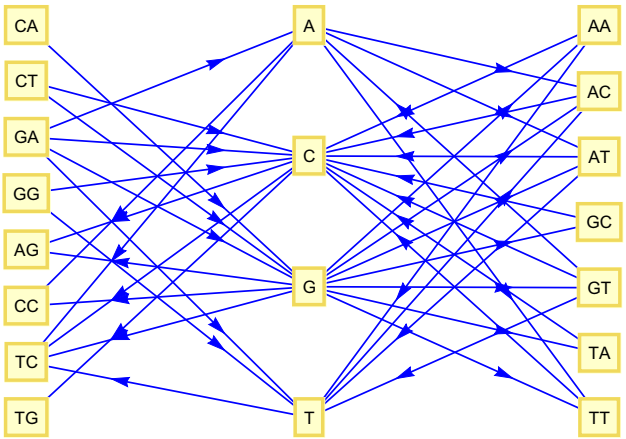

Fig. 1 Graph $\mathcal{G}(X)$ of the maximal C^3 self-complementary trinucleotide circular code $X = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$ of size 20 in genes of bacteria, archaea, eukaryotes, plasmids and viruses (Michel 2017, 2015; Arquès and Michel 1996). The four nucleotides $\{A, C, G, T\}$ have ingoing and outgoing edges. The four dinucleotides $\{AG, CC, TC, TG\}$ of X have no outgoing edge, the four dinucleotides $\{CA, CT, GA, GG\}$ of X have no ingoing edge, and the seven remaining dinucleotides $\{AA, AC, AT, GC, GT, TA, TT\}$ have ingoing and outgoing edges

only one partition into trinucleotides from X when read on a circle. Moreover, the code is called *comma-free* if, given any two trinucleotides $x_1, x_2 \in X$, any subtrinucleotide of the concatenation x_1x_2, except x_1, x_2 themselves, does not belong to X (Crick et al. 1957; Golomb et al. 1958a, b). Roughly speaking, the reading frame is retrieved after the reading of one trinucleotide in a sequence of trinucleotides from a comma-free code, while for circular codes, it is retrieved after the reading of at most four trinucleotides.

Here, we recall the main result from Fimmel et al. (2016) on the graph theoretic characterization of circular codes. Recall that a *cycle* in a graph $\mathcal{G}$ is an oriented closed path and that a graph is *acyclic* if it has no cycles (Clark and Holton 1991).

Theorem 2.2 (Theorem 2.6, Fimmel et al. 2016). *Given a trinucleotide code $X \subseteq \mathcal{B}^3$, the following statements are equivalent*

(1) X is circular;
(2) $\mathcal{G}(X)$ is acyclic.

Figure 1 displays the graph $\mathcal{G}(X)$ of the maximal C^3 self-complementary trinucleotide circular code X (1.1) identified in genes.

Recall that a code is called *self-complementary* if for each trinucleotide from X also the *complementary trinucleotide* is in X. (A codon in X implies that its anticodon is also in X.) Moreover, a code has the *C^3-property* if besides X also the two shifted codes $\alpha_1(X)$ and $\alpha_2(X)$ are circular, i.e., X is also circular in frames 1 and 2 (for more details see, for example, Fimmel et al. 2016). Clearly, a circular code can contain at most 20 trinucleotides, so a *maximal* circular code has a size exactly equal to 20.

Finally, we recall the main results from Fimmel et al. (2016, 2017) that characterize the comma-free codes and the strong comma-free codes by the longest paths in their associated graphs.

Theorem 2.3 (Theorem 2.11, Fimmel et al. 2016). *Given a trinucleotide code $X \subseteq \mathcal{B}^3$, the following statements are equivalent*

(1) X is comma-free;
(2) The longest path in $\mathcal{G}(X)$ is of length at most 2.

Theorem 2.4 (Definition 2.7, Fimmel et al. 2017). *Given a trinucleotide code $X \subseteq \mathcal{B}^3$, the following statements are equivalent*

(1) X is strong comma-free;
(2) The longest path in $\mathcal{G}(X)$ is of length at most 1.

In the next section, we will investigate for the first time the important biological property of self-complementarity as a graph property. We will also extend the above Theorem 2.3 by relating the reading frame of self-complementary circular codes to the longest paths in their associated graphs. Note that by Theorem 2.2, any graph associated with a circular code has a bound on the lengths of paths since the graph is finite.

Self-complementarity as a graph property

As we have seen in the above Theorems 2.2 and 2.3 (Theorems 2.6 and 2.11 in Fimmel et al. 2016), graph theory provides a handsome criterion for testing circularity or comma-freeness of codes. In this section, we will show that also the very important biological property of self-complementarity can be deduced from graphs associated with codes.

We first describe self-complementarity of some codes X. At first, we investigate the *reversing (mirroring) transformation* which inverts the order of bases in any trinucleotide, i.e., for $x = N_1N_2N_3 \in \mathcal{B}^3$ we have $\overleftarrow{x} = N_3N_2N_1 \in \mathcal{B}^3$. If X is any trinucleotide code then $\overleftarrow{X} = \{\overleftarrow{x} : x \in X\}$ is the *reversed code* of X. Similarly, the *complementing map* $c : \{A, C, G, T\} \to \{A, C, G, T\}$ that exchanges A and T as well as C and G induces the *complemented code* $c(X) = \{c(x) : x \in X\}$ where $c(N_1N_2N_3) = c(N_1)c(N_2)c(N_3)$ for any trinucleotide $x \in \mathcal{B}^3$. Note that for a trinucleotide

(also called codon) $x = N_1N_2N_3$, the complementary trinucleotide (also called anticodon) of x is exactly $\overleftarrow{c(x)}$ and that a trinucleotide code X is called *self-complementary* if $X = \overleftarrow{c(X)}$. The definitions of complementary and reversed complemented are identical.

The following proposition shows that the graph associated with a self-complementary trinucleotide code satisfies necessary conditions on its set of vertices. Recall that for a vertex $v \in \mathcal{G}$ in some directed graph $\mathcal{G}$, the *ingoing degree* $d^-(v)$ of v is the number of (directed) edges from $\mathcal{G}$ that end in v, while the *outgoing degree* $d^+(v)$ of v is the number of (directed) edges from $\mathcal{G}$ that start from v (Clark and Holton 1991).

Proposition 3.1 *Let $X \subseteq \mathcal{B}^3$ be a self-complementary trinucleotide code and $\mathcal{G}(X) = (V(X), E(X))$ the graph associated with X. Then*

(1) $V(X) = \overleftarrow{c(V(X))}$, i.e., for each nucleotide $v \in V(X)$ its complementary nucleotide $c(v) \in V(X)$ and for each dinucleotide $v \in V(X)$ its complementary dinucleotide $\overleftarrow{c(v)} \in V(X)$;
(2) $d^+(v) = d^-(\overleftarrow{c(v)})$ for any vertex $v \in V(X)$.

Proof Claim (1): Let $N_1N_2N_3 \in X$. Since X is self-complementary we have $c(N_3)c(N_2)c(N_1) \in X$. Thus, by definition of $\mathcal{G}(X)$, $N_1, N_3, c(N_1), c(N_3) \in V(X)$ and $N_1N_2, N_2N_3, c(N_3)c(N_2), c(N_2)c(N_1) \in V(X)$, and hence Claim (1) holds.

Claim (2): $[N_1, N_2N_3], [N_1N_2, N_3] \in E(X)$ is equivalent to $[c(N_3)c(N_2), c(N_1)], [c(N_3), c(N_2)c(N_1)] \in E(X)$. □

It is now tempting to conjecture that the statement in Proposition 3.1 is also sufficient for a code which is circular. But this is not the case—unless for a circular code of size at least 18 as we will see in the next theorem.

Theorem 3.2 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide circular code of size at least* 18. *Then X is self-complementary if and only if*

(1) $\mid X \mid$ is even, i.e., $\mid X \mid = 18$ or $\mid X \mid = 20$ (and hence maximal);
(2) $V(X) = \overleftarrow{c(V(X))}$;
(3) $d^+(v) = d^-(\overleftarrow{c(v)})$ for any vertex $v \in V(X)$.

Proof One direction follows immediately from Proposition 3.1. Note that a self-complementary code has to be of even size since no trinucleotide equals its complementary (reversed complemented) trinucleotide. The opposite direction is proved by computer calculations for all the 12,964,440 maximal circular codes and all the 1,012,099,740 circular codes of size 18. There are 528 maximal self-complementary circular codes among the 12,964,440 maximal circular codes, and 4032 self-complementary circular codes of size 18 among the 1,012,099,740 circular codes of size 18. □

It is a very surprising fact that the above equivalence in Theorem 3.2 only holds for circular codes of sizes 18 or 20. We will show next that one can neither avoid the assumption on circularity nor the assumption on the size of the codes in Theorem 3.2. We start with a constructive process that yields in the end codes satisfying the two conditions (2) and (3) of Theorem 3.2.

Construction method 3.3 Start with a trinucleotide $N_1N_2N_3$ and then choose a next trinucleotide that starts with the complementary of the dinucleotide N_2N_3 but does not end with the complementary of N_1. Continue this process until you get a long sequence of trinucleotides. The code constructed this way will satisfy the two conditions (2) and (3) of Theorem 3.2, but it is not self-complementary.

We give a basic example constructed by Method 3.3.

Example 3.4 The code $X = \{CAC, GAG, CTG, GTC\}$ is not self-complementary since, for example, it does not contain the complementary trinucleotide GTG of CAC, but it is easy to see that its corresponding graph satisfies the two conditions (2) and (3) from Theorem 3.2. The code X is even comma-free and has been constructed using the above construction Method 3.3: $CAC \rightsquigarrow GTC \rightsquigarrow GAG \rightsquigarrow CTG$.

However, Method 3.3 does not yield non-self-complementary codes of size larger than 8 such that the associated graphs satisfy the two conditions (2) and (3) of Theorem 3.2.

Construction method 3.5 In order to construct non-self-complementary codes larger than the size 8 and satisfying the two conditions (2), (3) of Theorem 3.2, a way is to combine codes constructed by Method 3.3.

Using Method 3.5, Example 3.7 below shows that there are even codes of size 20 such that their associated graphs satisfy the two conditions (2) and (3) of Theorem 3.2, but are not self-complementary, and even strongly not self-complementary, and not circular.

Definition 3.6 A code Y is *strongly not self-complementary* if for any trinucleotide $y \in Y$, the complementary trinucleotide $\overleftarrow{c(y)} \notin Y$.

Example 3.7 The code Y of size 20

$$Y = \{AAT, ACA, AGT, ATC, CAC, CCG, CGA, CTG, GAA, GAG, GCA, GGC, GTC, GTT, TAC, TCC, TCT, TGA, TGG, TTA\}$$

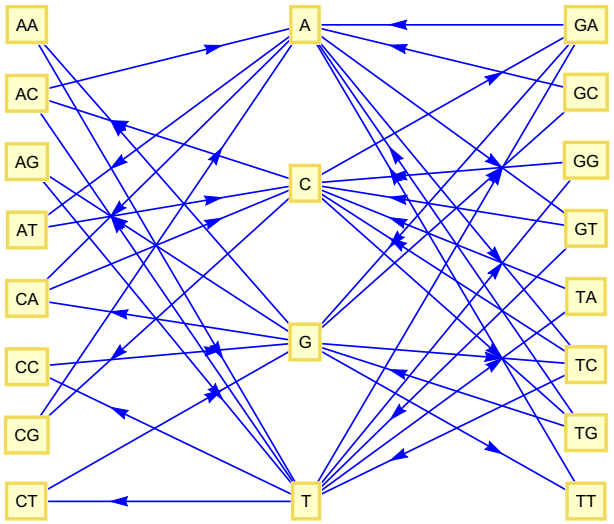

Fig. 2 Graph $\mathcal{G}(Y)$ of the strongly not self-complementary and not circular code $Y = \{AAT, ACA, AGT, ATC, CAC, CCG, CGA, CTG, GAA, GAG, GCA, GGC, GTC, GTT, TAC, TCC, TCT, TGA, TGG, TTA\}$ of size 20 satisfying the two conditions (2) and (3) of Theorem 3.2

is strongly not self-complementary and not circular, but its graph $\mathcal{G}(Y)$ satisfies the two conditions (2) and (3) of Theorem 3.2. Figure 2 displays the graph $\mathcal{G}(Y)$ associated with Y.

Strongly not self-complementary circular codes satisfying the two conditions (2) and (3) of Theorem 3.2 exist. Example 3.8 shows a strongly not self-complementary circular code of size 10.

Example 3.8 The code X_1 of size 10

$$X_1 = \{AAT, ATC, CAC, CTG, GAA, GAG, GTC, GTT, TAC, TTA\}$$

is a strongly not self-complementary circular code with its graph $\mathcal{G}(X_1)$ satisfying the two conditions (2) and (3) of Theorem 3.2. Figure 3 displays the graph $\mathcal{G}(X_1)$ associated with X_1.

Codes of large sizes that are not circular, not self-complementary and strongly not self-complementarity satisfying the two conditions (2) and (3) of Theorem 3.2 can easily be constructed, as shown in Example 3.9.

Example 3.9 The addition to the code $X = \{CAC, GAG, CTG, GTC\}$ from Example 3.4 of pairs of trinucleotide–complementary trinucleotide which are not contained in X can build not self-complementary codes of every even size between 4 and 60 such that their associated graphs have the two conditions (2) and (3) of Theorem 3.2. Note that adding such trinucleotide pairs does not violate the

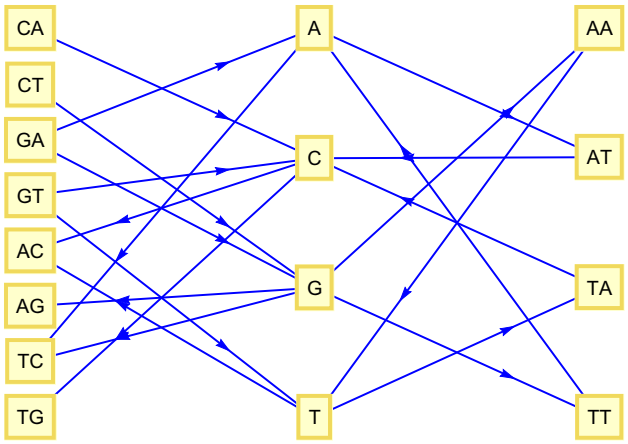

Fig. 3 Graph $\mathcal{G}(X_1)$ of the strongly not self-complementary circular code $X_1 = \{AAT, ATC, CAC, CTG, GAA, GAG, GTC, GTT, TAC, TTA\}$ of size 10 satisfying the two conditions (2) and (3) of Theorem 3.2

conditions (2) and (3) of Theorem 3.2 by the next Lemma 3.10.

We continue with a few closure properties of the class of graphs that satisfy the two conditions (2) and (3) of Theorem 3.2.

Lemma 3.10 *Let $X_1, X_2 \subseteq \mathcal{B}^3$ with $X_1 \cap X_2 = \emptyset$ be trinucleotide codes such that their associated graphs $\mathcal{G}(X_1)$ and $\mathcal{G}(X_2)$ satisfy the two conditions* (2) *and* (3) *of Theorem 3.2. Then the following statements hold*

(1) The graph $\mathcal{G}(X_1^c)$ where $X_1^c := \mathcal{B}^3 \setminus X_1$ satisfies both conditions (2) *and* (3) *as well;*
(2) The graph $\mathcal{G}(Z)$ where $Z := X_1 \cup X_2$ satisfies both conditions (2) *and* (3) *as well.*

Proof Let $X_1, X_2 \subseteq \mathcal{B}^3$ with $X_1 \cap X_2 = \emptyset$ be codes such that their associated graphs $\mathcal{G}(X_1)$ and $\mathcal{G}(X_2)$ satisfy the two conditions (2) and (3) of Theorem 3.2.

Claim (1): It follows from the fact that the graph $\mathcal{G}(\mathcal{B}^3)$ satisfies the two conditions (2) and (3) of Theorem 3.2 and[1] $\mathcal{G}(\mathcal{B}^3) = \mathcal{G}(X_1) \cup \mathcal{G}(X_1^c)$ and $E(X_1) \cap E(X_1^c) = \emptyset$.

Claim (2): Condition (2) of Theorem 3.2 is obviously true since $V(Z) = V(X_1) \cup V(X_2)$. Let us show that Condition (3) of Theorem 3.2 also holds. Since $X_1 \cap X_2 = \emptyset$, it follows that $E(X_1) \cap E(X_2) = \emptyset$. Two cases are considered: (i) If $v \notin V(X_1) \cap V(X_2)$ then also $\overleftarrow{c(v)} \notin V(X_1) \cap V(X_2)$ and Condition (3) of Theorem 3.2 is satisfied since it

[1] Recall that the union $G_1 \cup G_2$ of two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is defined as $G = (V_1 \cup V_2, E_1 \cup E_2)$ (Clark and Holton 1991).

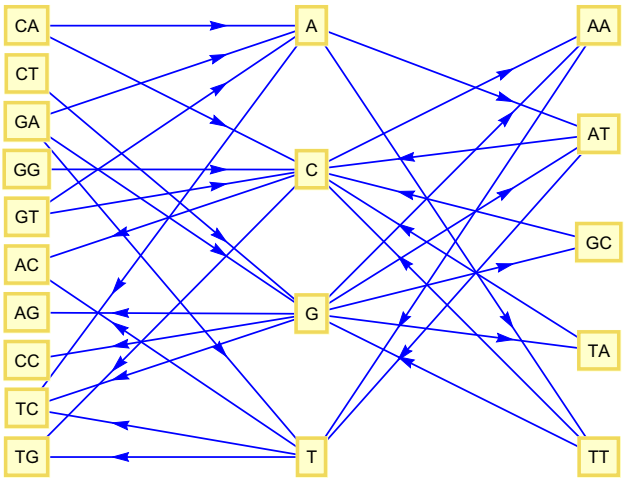

Fig. 4 Graph $\mathcal{G}(X_2)$ of the not self-complementary, circular code $X_2 = \{AAT, ATC, ATT, CAA, CAC, CTG, GAA, GAG, GAT, GCC, GGC, GTA, GTC, TAC, TTC, TTG\}$ of size 16 satisfying the two conditions (2) and (3) of Theorem 3.2

holds in $\mathcal{G}(X_1)$ or $\mathcal{G}(X_2)$; (ii) if $v \in V(X_1) \cap V(X_2)$ then also $\overleftarrow{c(v)} \in V(X_1) \cap V(X_2)$ and Condition (3) of Theorem 3.2 is also satisfied since in- and out-degrees which are equal in $\mathcal{G}(X_1)$ and $\mathcal{G}(X_2)$, respectively, are added. □

In general, the graphs $\mathcal{G}(X_1 \cap X_2)$ and $\mathcal{G}(X_1 \cup X_2)$ do not satisfy the two conditions (2) and (3) of Theorem 3.2 even though both graphs $\mathcal{G}(X_1)$ and $\mathcal{G}(X_2)$ do, as shown in Example 3.11.

Example 3.11 The two codes

$$X_1 = \{CAC, GAG, CTG, GTC\}, \quad X_2 = \{CAC, GTG\}$$

lead to

$$X_1 \cap X_2 = \{CAC\}, \quad X_1 \cup X_2 = \{CAC, GAG, CTG, GTC, GTG\}.$$

Then $\mathcal{G}(X_1 \cap X_2)$ does not even satisfy Condition (2) of Theorem 3.2 and $\mathcal{G}(X_1 \cup X_2)$ does not satisfy Condition (3) of Theorem 3.2, while both graphs $\mathcal{G}(X_1)$ and $\mathcal{G}(X_2)$ do.

Neither the assumption on the size of the code nor its circularity can be omitted in Theorem 3.2, as shown in Example 3.12.

Example 3.12 The code X_2 of size 16

$$X_2 = \{AAT, ATC, ATT, CAA, CAC, CTG, GAA, GAG, GAT, GCC, GGC, GTA, GTC, TAC, TTC, TTG\}$$

is circular (even C^3), but not self-complementary even if its graph $\mathcal{G}(X_2)$ satisfies the two conditions (2) and (3) of Theorem 3.2. Figure 4 displays the graph $\mathcal{G}(X_2)$ associated with X_2.

After having stated a theorem for the self-complementarity of circular codes of large sizes, we aim a similar one for comma-free codes.

Theorem 3.13 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide comma-free code of size at least* 14. *Then X is self-complementary if and only if*

(1) $| X | = 14$ *or* $| X | = 16$;[2]
(2) $V(X) = \overleftarrow{c(V(X))}$;
(3) $d^+(v) = d^-(\overleftarrow{c(v)})$ *for any vertice* $v \in V(X)$.

Proof As in the proof of Theorem 3.2, one direction follows immediately from Proposition 3.1. The opposite direction is proved by means of computer calculations for all the 25,473,732 comma-free codes of size 14 and all the 2,743,080 comma-free codes of size 16. The fact that there are no self-complementary comma-free codes of size 18 or 20 (Michel et al. 2008) completes the proof. □

In the next section, we provide a characterization of the longest paths in the graphs associated with self-complementary circular codes.

Longest paths in the graphs associated with self-complementary circular codes

In this section, we study the structure of the longest paths in graphs associated with self-complementary circular codes. Here, the length of a path may have different meanings, namely either the number of edges in the path or the length of the word obtained by concatenating its labels (vertices). In the sequel of this section, we will only look at the so-called *arrow-length*.

Definition 4.1 Let $X \subseteq \mathcal{B}^3$ be a trinucleotide circular code and $\mathcal{G}(X)$ its associated graph. Let $p : t_1 \to \cdots \to t_n$ be a path in $\mathcal{G}(X)$ where $t_i \in \mathcal{B} \cup \mathcal{B}^2$ for $i = 1, \ldots, n$. Then the *arrow-length* $l_a(p)$ is defined as $n - 1$. Moreover, by $l_{max}(X)$ we denote the maximal arrow-length of a path, i.e., the length of a longest path, in the associated graph $\mathcal{G}(X)$.

We would like to remark that the assumption on the circularity of the code in Definition 4.1 has only be made

[2] Due to self-complementarity of X, $| X |$ must be even, but in opposite to circular codes, there are no self-complementary comma-free codes of sizes 18 or 20.

Table 1 Growth function of self-complementary circular codes X of even cardinality $n = 2, 4, \ldots, 20$ as a function of the longest path length $l_{max}(X) = 1, \ldots, 8$ in their associated graph $\mathcal{G}(X)$

l_{max}	$n=2$	$n=4$	$n=6$	$n=8$	$n=10$	$n=12$	$n=14$	$n=16$	$n=18$	$n=20$
1	12	8	0	0	0	0	0	0	0	0
2	16	202	556	642	396	152	36	4	0	0
3	0	16	152	336	280	80	0	0	0	0
4	0	108	1344	5808	12,048	14,032	9800	4116	964	96
5	0	0	0	0	0	0	0	0	0	0
6	0	0	68	684	2352	3896	3568	1872	532	64
7	0	0	0	0	0	0	0	0	0	0
8	0	0	56	824	4024	9104	10,920	7248	2536	368
Total	28	334	2176	8294	19,100	27,264	24,324	13,240	4032	528

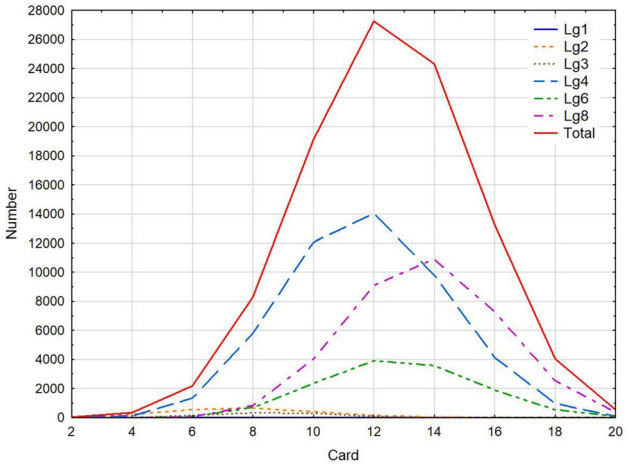

Fig. 5 Growth function of self-complementary circular codes X of even cardinality $n = 2, 4, \ldots, 20$ as a function of the longest path length $l_{max}(X) = 1, \ldots, 8$ in their associated graph $\mathcal{G}(X)$

in order to ensure that there is no cycle in the graph; otherwise, the longest path has an infinite length. Recall that the longest path for the comma-free codes has length $l_{max}(X) = 2$ (Theorem 2.3) and that the longest path for the strong comma-free codes has length $l_{max}(X) = 1$ (Theorem 2.4).

Table 1 gives the number of self-complementary circular codes X of different cardinality n (number of trinucleotides in the code) depending on the length of a longest path $l_{max}(X)$. Figure 5 is a graphical representation of Table 1.

Surprisingly, according to the computational results in Table 1, $l_{max}(X)$ is always bounded by 8. Theorem 4.2 below explains this issue and characterizes completely the possible values of $l_{max}(X)$ for non-maximal and maximal self-complementary circular codes.

Theorem 4.2 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide circular code. The following statements about the maximal arrow-length $l_{max}(X)$ of a path are true*

(1) $1 \leq l_{max}(X) \leq 8$;
(2) If X is self-complementary, then $l_{max}(X) \in \{1, 2, 3, 4, 6, 8\}$, i.e., $l_{max}(X) = 5, 7$ are excluded;
(3) If X is maximal and self-complementary, then $l_{max}(X) \in \{4, 6, 8\}$, i.e., in addition to (2), $l_{max}(X) = 1, 2, 3$ are impossible.

Proof Claim (1): It is immediate since for $l_{max}(X) \geq 9$ in a graph $\mathcal{G}(X)$ associated with a circular code, there is a path containing at least 5 vertices labeled by nucleotides. Note that by construction of $\mathcal{G}(X)$, the labels of the vertices alternate between nucleotides and dinucleotides. However, there are only 4 different bases in the alphabet $\mathcal{B}$; hence, 2 of the vertices must have the same label which yields a cycle in $\mathcal{G}(X)$, in contradiction to circularity. Thus, $1 \leq l_{max}(X) \leq 8$.

Claim (2): Let X be a self-complementary circular code.

(i) Assume that $l_{max}(X) \geq 4$ is odd. By construction of $\mathcal{G}(X)$, any path in $\mathcal{G}(X)$ starts with either a nucleotide or a dinucleotide. Moreover, the vertices of the path alternate between nucleotides and dinucleotides. Thus, if $l_{max}(X)$ is odd, then a longest path in $\mathcal{G}(X)$ must either be of the form

$$(I) \quad l_1 \to d_1 \to l_2 \to d_2 \to \cdots \to d_{n-1} \to l_n \to d_n$$

starting with a nucleotide l_1 and ending with a dinucleotide d_n or

$$(II) \quad d_1 \to l_1 \to d_2 \to l_2 \to \cdots \to l_{n-1} \to d_n \to l_n$$

starting with a dinucleotide d_1 and ending with a nucleotide l_n. In fact, the following argument shows that actually both cases hold. Assume without loss of generality that a longest path is of the first form (I). By self-complementarity, we then obtain a *complementary and reversed* path

$$(III) \quad \overleftarrow{c(d_n)} \to c(l_n) \to \overleftarrow{c(d_{n-1})} \to \cdots \to \overleftarrow{c(d_2)} \to c(l_2) \to \overleftarrow{c(d_1)} \to c(l_1).$$

Since we have assumed that $l_{max}(X) \geq 4$ in (I), there are at least 3 nucleotides $l_1, l_2, l_3, \ldots$ appearing. By circularity of the code X, all these nucleotides have to be different since otherwise the path would contain a cycle. Similarly, the path (III) has also at least 3 different nucleotides $c(l_n), c(l_{n-1}), c(l_{n-2}), \ldots$. However, there are only 4 nucleotides in the alphabet $\mathcal{B}$, so there must be $i, j \leq n$ such that $l_i = c(l_j)$. Since the path (I) starts with a nucleotide and the path (III) starts with a dinucleotide, the two paths

$$(I') \quad l_1 \to d_1 \to l_2 \to d_2 \to \cdots \to d_{i-1} \to l_i$$
$$(III') \quad \overleftarrow{c(d_n)} \to c(l_n) \to \overleftarrow{c(d_{n-1})} \to \cdots \to \overleftarrow{c(d_j)} \to c(l_j)$$

must have different lengths. Without loss of generality, assume that (III') is the longer path, but then the path

$$\overleftarrow{c(d_n)} \to c(l_n) \to \overleftarrow{c(d_{n-1})} \to \cdots \to \overleftarrow{c(d_j)} \to c(l_j) = l_i \to d_i \to \cdots \to d_{n-1} \to l_n \to d_n$$

has length greater than $l_{max}(X)$—a contradiction.

(ii) The following Examples 4.3, 4.4 and 4.5 show that $l_{max}(X) = 1, 2, 3$ exist for self-complementary circular codes that are not maximal.

Claim (3): Let X be a maximal self-complementary circular code.

(i) If $l_{max}(X) \leq 2$ is true then X is comma-free. However, there are no maximal self-complementary comma-free codes (Table 7 in Michel et al. 2008). Thus, according to the claim (2), $l_{max}(X) \in \{3, 4, 6, 8\}$.

(ii) Assume now that $l_{max}(X) = 3$. By maximality and circularity, X must contain exactly one element in each equivalence class $\{N_1N_2N_3, N_2N_3N_1, N_3N_1N_2\}$ for every trinucleotide $N_1N_2N_3$. Thus, X must contain one trinucleotide from $\{AAT, ATA, TAA\}$ and one complementary trinucleotide from $\{ATT, TTA, TAT\}$. It is easy to see that each combination yields a path of the form $A \to d_1 \to T$ or $T \to d_1 \to A$ for some dinucleotide d_1. Similarly, we get a path of the form $C \to d_2 \to G$ or $G \to d_2 \to C$ for some dinucleotide d_2. Without loss of generality, assume that $A \to d_1 \to T$ and $C \to d_2 \to G$ are paths in $\mathcal{G}(X)$. Clearly, the four trinucleotides Ad_1, d_1T, Cd_2, d_2G are all different; hence, $X' = X \backslash \{Ad_1, d_1T, Cd_2, d_2G\}$ has 16 elements. Assume that there is a trinucleotide $dC \in X'$, d being a dinucleotide, and then also $Gc(d) \in X'$ by self-complementarity. So we get a path $d \to C \to d_2 \to G \to c(d)$ of length 4—a contradiction.

Similarly, we cannot have trinucleotides of the form $dA, Td, Gd \in X'$. So no trinucleotide in X' starts with T or G and no trinucleotide ends with C or A. Hence, $X' \subseteq S = \{N_1N_2N_3 \mid N_2 \in \mathcal{B}, N_1 \in \{A, C\}, N_3 \in \{G, T\}\}$. Clearly, $\mid S \mid = 16$. However, the 4 trinucleotides Ad_1, d_1T, Cd_2, d_2G are also in S, but excluded from X', so $\mid X' \mid \leq 12$ —a contradiction. □

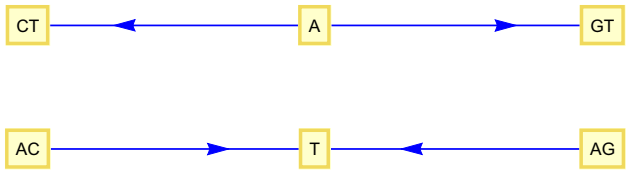

Fig. 6 Graph $\mathcal{G}(X_3)$ of the self-complementary circular code $X_3 = \{ACT, AGT\}$ of size 2 with longest path length $l_{max}(X_3) = 1$

Example 4.3 The code $X_3 = \{ACT, AGT\}$ of size 2 is a self-complementary circular code with longest path length $l_{max}(X_3) = 1$, e.g., $A \to CT$, $AG \to T$, etc. (Figure 6).

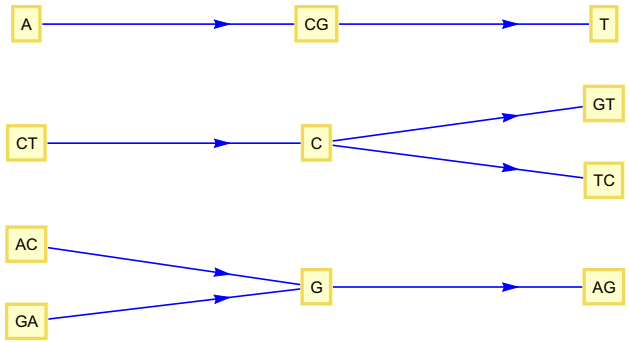

Fig. 7 Graph $\mathcal{G}(X_4)$ of the self-complementary circular code $X_4 = \{ACG, CGT, CTC, GAG\}$ of size 4 with longest path length $l_{max}(X_4) = 2$

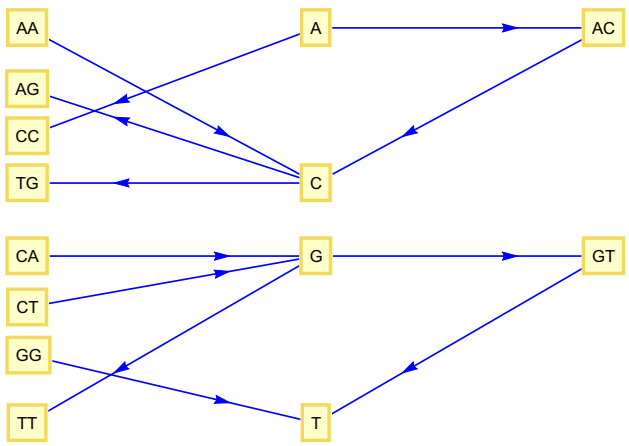

Fig. 8 Graph $\mathcal{G}(X_5)$ of the self-complementary circular code $X_5 = \{AAC, ACC, CAG, CTG, GGT, GTT\}$ of size 6 with longest path length $l_{max}(X_5) = 3$

Example 4.4 The code $X_4 = \{ACG, CGT, CTC, GAG\}$ of size 4 is a self-complementary circular (even comma-free) code with longest path length $l_{max}(X_4) = 2$, e.g., $A \to CG \to T$, $CT \to C \to GT$, $GA \to G \to AG$, etc. (Figure 7).

Example 4.5 The code $X_5 = \{AAC, ACC, CAG, CTG, GGT, GTT\}$ of size 6 is a self-complementary circular code with longest path length $l_{max}(X_5) = 3$, e.g., $A \to AC \to C \to AG, CA \to C \to GT \to T$, etc. (Fig. 8).

In summary, Theorem 4.2 proves that the longest paths for the maximal self-complementary circular codes are always symmetric (nucleotide–nucleotide $l_1 \to \cdots \to l_n$ or dinucleotide–dinucleotide $d_1 \to \cdots \to d_n$), while the longest paths for the non-maximal self-complementary circular codes can in addition be asymmetric (nucleotide–dinucleotide $l_1 \to \cdots \to d_n$ or dinucleotide–nucleotide $d_1 \to \cdots \to l_n$).

The structure of longest paths in Examples 4.3, 4.4 and 4.5 is not unique. Indeed, the longest paths can start with either a nucleotide or a dinucleotide. However, for the maximal comma-free codes, the longest path of length 2 needs to start with a dinucleotide (see Theorem 4.8 below). For the convenience of the reader, Example 4.6 gives maximal self-complementary circular codes that have longest path lengths equal to 4, 6 and 8 (without displaying their associated graphs).

Example 4.6 The maximal self-complementary circular codes

(1) $X_6 = \{AAC, AAT, ACC, ACT, AGA, AGC, AGG, AGT, ATC, ATT, CCT, GAC, GAT, GCC, GCT, GGC, GGT, GTC, GTT, TCT\}$ with $l_{max}(X_6) = 4$ has a longest path

$$AG \to A \to AC \to C \to CT;$$

(2) $X_7 = \{AAG, AGG, CAA, CAG, CCA, CCG, CCT, CGA, CGG, CTA, CTG, CTT, TAA, TAG, TCA, TCG, TGA, TGG, TTA, TTG\}$ with $l_{max}(X_7) = 6$ has a longest path

$$C \to CT \to T \to CA \to A \to AG \to G;$$

(3) $X_8 = \{AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGT, ATC, ATT, CGT, CTT, GAT, GCC, GCT, GGA, GGC, GGT, GTT, TCC\}$ with $l_{max}(X_8) = 8$ has a longest path

$$GG \to A \to AC \to G \to AT \to C \to GT \to T \to CC.$$

The structure of the longest paths in Example 4.6 is not arbitrary. Indeed, the longest paths in the associated graphs of maximal self-complementary circular codes have a unique structure, as shown in Theorem 4.7.

Theorem 4.7 *Let $X \subseteq \mathcal{B}^3$ be a maximal self-complementary trinucleotide circular code. Then the following statements hold*

(1) If $l_{max}(X) = 4$, then the longest paths are of the form $d_1 \to l_1 \to d_2 \to l_2 \to d_3$;
(2) If $l_{max}(X) = 6$, then the longest paths are of the form $l_1 \to d_1 \to l_2 \to d_2 \to l_3 \to d_3 \to l_4$;
(3) If $l_{max}(X) = 8$, then the longest paths are of the form $d_1 \to l_1 \to d_2 \to \cdots \to d_4 \to l_4 \to d_5$

where the nucleotide $l_i \in \mathcal{B}$ and the dinucleotide $d_i \in \mathcal{B}^2$ for any i.

Proof See "Appendix." □

We now turn to the comma-free codes. By Theorem 2.3, any comma-free code satisfies $l_{max}(X) = 2$. Theorem 4.8 below will show that the longest paths always have to start and end with dinucleotides if the comma-free code is maximal.

Theorem 4.8 *Let $X \subseteq \mathcal{B}^3$ be a maximal trinucleotide comma-free code. Then $l_{max}(X) = 2$ and the longest paths are of the form $d_1 \to l_1 \to d_2$ where the nucleotide $l_1 \in \mathcal{B}$ and the dinucleotides $d_1, d_2 \in \mathcal{B}^2$.*

Proof Let $l_{max}(X) = 2$ and assume that $l_1 \to d_1 \to l_2$ is the maximal path in $\mathcal{G}(X)$. Clearly, $l_1 \neq l_2$ and there is no trinucleotide in X starting with l_2 or ending with l_1 since otherwise the path could be extended. Let b_1, b_2 be the remaining 2 nucleotides. By maximality, X must contain one trinucleotide of the class $\{l_1l_1b, bl_1l_1, l_1bl_1\}$ for each nucleotide $b \neq l_1$. Since trinucleotides ending in l_1 are forbidden, we conclude that $l_1l_1l_2, l_1l_1b_1$ and $l_1l_1b_2$ are in X. Similarly, it follows that $l_1l_2l_2, b_1l_2l_2$ and $b_2l_2l_2$ are in X. So $l_1l_1l_2$ and $l_1l_2l_2$ are in X and yield the path $l_1 \to l_1l_2 \to l_2$. Now consider the class $\{l_1l_2b_1, b_1l_1l_2, l_2b_1l_1\}$. Again by maximality, one of the trinucleotides of this class must be in X. However, $l_2b_1l_1$ is excluded since it starts with l_2 and ends with l_1. If $l_1l_2b_1 \in X$ then we get the path $l_1 \to l_1l_2 \to b_1$. Hence, no trinucleotide in X is allowed to start with b_1 but $b_1l_2l_2 \in X$ —a contradiction. Similarly, $b_1l_2l_2 \in X$ yields a contradiction. □

In the next section, we will show that the length of the *reading frame* of a circular code can be deduced from the longest path in the associated graph.

The reading frame of circular codes

Circular codes do not always recognize a frameshift immediately, but may be only after reading a few trinucleotides. From the formal definition, in a sequence consisting entirely of trinucleotides from a given circular code X, there are arbitrarily long subsequences that could be read in two ways, namely in the reading (correct) frame and also in the frames 1 or 2, respectively. We will show that the reading in two frames is impossible and prove that the length of such subsequences is bounded. We will prove that this bound, called the *reading frame number*, is determined by the longest path in the graph $\mathcal{G}(X)$ associated with X. Moreover, most importantly, we will show that the reading frame number of a circular code even allows to retrieve the reading frame in arbitrary sequences of trinucleotides whenever a subsequence of at least five consecutive trinucleotides from X (called an X motif) is read. This supports strongly the idea that the ribosome may retrieve the reading frame using X motifs from the circular code X identified in genes (detailed in "Application: Reading frame of the maximal C^3 self-complementary circular code X identified in genes" section).

Definition 5.1 Let $X \subseteq \mathcal{B}^3$ be a trinucleotide code and $b_1 \cdots b_n$ a sequence of nucleotides where $b_i \in B$. A *possible X-frame* for the sequence $b_1 \cdots b_n$ is a division

$$b_1 \cdots b_n = t_b c_1 \cdots c_l t_e$$

where $l \geq 0$, $c_i \in X$ for $i = 1, \ldots, l$ and $t_b \in \{\epsilon, b_1, b_1 b_2\}$ and $t_e \in \{\epsilon, b_{n-1} b_n, b_n\}$, ϵ being the empty word.

Note that for $l = 0$ in the above Definition 5.1, we include the case $b_1 b_2 b_3 b_4$ for which $t_b = b_1 b_2$ and $t_e = b_3 b_4$ is a possible X-frame.

Example 5.2 Let $X = \{ACG, CGT, TAT, ATG, GAC\}$. Then the sequence $ACGTATGAC$ has 2 possible X-frames, namely

$$ACG \quad TAT \quad GAC \quad \text{with } t_b = \epsilon = t_e$$

and

$$A \quad CGT \quad ATG \quad AC \text{ with } t_b = A, t_e = AC.$$

Remark 5.3 From the above Definition 5.1, we require that the middle part in a possible X-frame consists of trinucleotides from the code X (may be empty), but we do not make any hypothesis on the beginning and the end of the sequence, i.e., we do not require that the beginning of the sequence is a suffix of a trinucleotide of X and also that the end of the sequence does not need to be a prefix of a trinucleotide of X. This approach contrasts the notion of unambiguous words defined in Michel (2012) and makes the notion of X-frame and later on reading frame applicable to arbitrary sequences, i.e., not entirely consisting of trinucleotides from X.

We have a first observation.

Observation 5.4 Let $X \subseteq \mathcal{B}^3$ be a trinucleotide circular code and $b_1 \cdots b_k$ a sequence of nucleotides where $b_i \in B$. If we assume that $b_1 \cdots b_k$ has 2 different possible X-frames

$$t_b u_1 \cdots u_l t_e \quad \text{and} \quad t'_b u'_1 \cdots u'_m t'_e$$

with $u_i, u'_i \in X$ and $t_b, t_e, t'_b, t'_e \in \left(\{\epsilon\} \cup \mathcal{B} \cup \mathcal{B}^2\right)$, then there exists a path in $\mathcal{G}(X)$ associated with the overlapping sequences $u_1 \cdots u_l$ and $u'_1 \cdots u'_m$. The word associated with this path (see Definition 5.6 below) covers exactly the smallest subsequence of $b_1 \cdots b_k$ that contains both $u_1 \cdots u_l$ and $u'_1 \cdots u'_m$.

Example 5.5 In the above Example 5.2, we obtain the path $A \to CG \to T \to AT \to G \to AC$.

Before we proceed, we need a few more definitions. Recall from Definition 4.1 that the *arrow-length* $l_a(p)$ of a path p is the number of edges in this path. For the sake of completeness, we include this definition again in the next definition.

Definition 5.6 Let $X \subseteq \mathcal{B}^3$ be a trinucleotide circular code and $\mathcal{G}(X)$ its associated graph. Let $p : t_1 \to \cdots \to t_n$ be a path in $\mathcal{G}(X)$ where $t_i \in \mathcal{B} \cup \mathcal{B}^2$ for $i = 1, \ldots, n$. Then

- the *word associated with* p is defined as $w(p) = t_1 \cdots t_n$, the concatenation of the labels of p;
- the *arrow-length* $l_a(p)$ is defined as $n - 1$ (see Definition 4.1);
- the *word-length* $l_w(p)$ is defined as $| w(p) |$, the length of the word associated with p.

We would like to remark that in general two paths of the same arrow-length can have different word-lengths, as shown in Example 5.7.

Example 5.7 $A \to CG \to T$ with $l_a(p) = 2$ and $l_w(p) = 4$, and $AT \to G \to TT$ with $l_a(p) = 2$ but $l_w(p) = 5$.

However, this case only happens if the two paths start with different labels, i.e., one is a nucleotide and the other one is a dinucleotide. Since we know that the different paths of maximal arrow-length in a maximal self-complementary circular code always have the same structure (see Theorem 4.7 above), we deduce that a path of maximal arrow-length in a maximal self-complementary circular code is also a path of maximal word-length and vice versa.

Fig. 9 A model of reading frame retrieval in genes using the X circular code motifs, i.e., motifs from the circular code X (1.1)

Definition 5.8 Let $X \subseteq \mathcal{B}^3$ be a trinucleotide circular code. We define the *reading frame number* n_X of X as

$$n_X := min\{n \in \mathbb{N} \mid \textit{for all sequences of nucleotides } b_1 \cdots b_n \textit{ there is at most one possible X-frame for } b_1 \cdots b_n\}.$$

For any sequence $b_1 \cdots b_n$ as in the above Definition 5.8, there is not always a possible X-frame, but the main point is that if there is one, e.g., in a subsequence of a sequence of trinucleotides of a circular code X, we want it to be unique. We start with a general upper bound for the reading frame number of any circular code.

Theorem 5.9 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide circular code and $\mathcal{G}(X)$ its associated graph. Then the reading frame number n_X satisfies $n_X \leq 2 \cdot l_{max}(X) + 3$.*

Proof Assume that there is a sequence $b_1 \cdots b_n$ of nucleotides that has 2 different possible X-frames: $t_b u_1 \cdots u_l t_e$ and $t'_b u'_1 \cdots u'_m t'_e$ with $u_i, u'_i \in X$ and $t_b, t_e, t'_b, t'_e \in \left(\{\epsilon\} \cup \mathcal{B} \cup \mathcal{B}^2\right)$. Then the 2 overlapping sequences $u_1 \cdots u_l$ and $u'_1 \cdots u'_m$ cover at least $n-2$ nucleotides of the sequence $b_1 \cdots b_n$. Each pair of overlapping trinucleotides from the 2 sequences $u_1 \cdots u_l$ and $u'_1 \cdots u'_m$ yield a path of length 2 in $\mathcal{G}(X)$. Putting these paths together we cannot exceed the arrow-length $l_a(p)$ of a maximal path in $\mathcal{G}(X)$ (which exists by Theorem 4.2 since X is circular). Thus in total, we obtain $n \leq 2 \cdot l_{max}(X) + 3$. □

The above proof can easily be generalized to circular codes over any arbitrary finite alphabet and any arbitrary word-length.

Theorem 5.10 *Let $X \subseteq \mathcal{A}^l$ be a circular code where $\mathcal{A}$ is a finite alphabet, l is a positive integer, and $\mathcal{G}(X)$ its associated graph. Then the reading frame number n_X satisfies $n_X \leq int(\frac{l}{2}) \cdot l_{max}(X) + l$ where $int(\frac{n}{2})$ denotes the smallest natural number greater than or equal to $\frac{n}{2}$.*

Proof Clear. □

We now determine explicitly the reading frame numbers for maximal self-complementary circular codes.

Theorem 5.11 *Let $X \subseteq \mathcal{B}^3$ be a maximal self-complementary trinucleotide circular code and $\mathcal{G}(X)$ its associated graph. Let $p = p_{max}(X)$ be a path of maximal arrow-length (and hence word-length) in $\mathcal{G}(X)$, and let $l_w(p)$ be its word-length. Then the following statements about the reading frame number n_X are true*

(1) $n_X = l_w(p) + 2$, *if* $p = d_1 \to b_1 \to \cdots \to b_k$ *or* $p = b_1 \to d_1 \to \cdots \to d_k$;
(2) $n_X = l_w(p) + 1$, *if* $p = d_1 \to b_1 \to \cdots \to d_k$;
(3) $n_X = l_w(p) + 3$, *if* $p = b_1 \to d_1 \to \cdots \to b_k$,

where the nucleotide $b_i \in \mathcal{B}$ and the dinucleotide $d_i \in \mathcal{B}^2$ for any i.

Proof See "Appendix." □

In Michel (2012), a slightly different definition of the reading frame number n_X was used where the words t_b and t_e in a possible X-frame have to be suffix and prefix, respectively, of some trinucleotides of X. This definition is a stronger requirement and thus yields smaller reading frame numbers n_X. For instance, $n_X = 13$ nucleotides for the maximal C^3 self-complementary code X from (1.1), $n_X = 3$ nucleotides for the comma-free codes and $n_X = 2$ nucleotides for the strong comma-free codes.

Application: Reading frame of the maximal C^3 self-complementary circular code X identified in genes

The longest paths in $\mathcal{G}(X)$ (Fig. 1) of the maximal C^3 self-complementary circular code X (1.1) identified in genes are:

$[CA, G, GT, A, AT, T, AC, C, AG]$,
$[CA, G, GT, A, AT, T, AC, C, TC]$,
$[CA, G, GT, A, AT, T, AC, C, TG]$,
$[CT, G, GT, A, AT, T, AC, C, AG]$,
$[CT, G, GT, A, AT, T, AC, C, TC]$,
$[CT, G, GT, A, AT, T, AC, C, TG]$,
$[GA, G, GT, A, AT, T, AC, C, AG]$,
$[GA, G, GT, A, AT, T, AC, C, TC]$,
$[GA, G, GT, A, AT, T, AC, C, TG]$.

These nine longest paths have the form $p = d_1 \to b_1 \to \cdots \to d_k$ with $l_w(p) = 14$ nucleotides. Thus, by application of Claim (2) of Theorem 5.11, the reading frame number n_X of X (1.1) is equal to 15 nucleotides (5 trinucleotides).

A model of frame retrieval was proposed in Fimmel et al. (2017) where the ribosome pairs with the X motifs located at different positions in the genes (Fig. 9).

The X motifs from the circular code X (1.1) occur preferentially in genes compared to genomes (noncoding regions of eukaryotes) with a factor of about 8 (Tables 4 and 5, Figures 7 and 8 in Soufi and Michel (2016)). Furthermore, very recent results have shown an enrichment of X motifs in the genes of the yeast *Saccharomyces cerevisiae* (Michel et al. 2017). Precisely, several basic statistical analyses comparing X motifs and R motifs (random motifs from random codes) demonstrated that:

(i) No significant difference is observed between the frequencies of X and R motifs in the noncoding regions of *S. cerevisiae*.

(ii) The frequency of X motifs is significantly greater than that of R motifs in the genes (protein-coding regions) of *S. cerevisiae*. This property is true for all cardinalities of X motifs (from 4 to 20 trinucleotides) and for all 16 chromosomes of *S. cerevisiae*.

(iii) The X motifs in the three frames of *S. cerevisiae* genes occur more frequently in the reading frame, regardless of their cardinality or their length.

(iv) The ratio of X genes, i.e., genes with at least one X motif, to non-X genes in the set of verified genes is significantly different to that observed in the set of putative or dubious genes with no experimental evidence.

The ribosome contains the circular code information for pairing with the X motifs in genes. Indeed, the X motifs are also identified in tRNAs of prokaryotes and eukaryotes (Michel 2012, 2013) and in rRNAs of prokaryotes (16S) and eukaryotes (18S), in particular in the ribosome decoding center where the universally conserved nucleotides G530, A1492 and A1493 are included in the X motifs (Michel 2012; Soufi and Michel 2014, 2015). Pairing of X motifs between mRNAs–rRNAs, mRNAs–tRNAs and rRNAs–tRNAs, shown with a 3D visualization of the ribosome (Michel 2012, 2013; Soufi and Michel 2014, 2015), may be involved in maintaining and synchronizing the reading frame during the translation process. However, the experimental biological mechanism by which the ribosome uses the X motifs for maintaining and synchronizing the reading frame during genome decoding and protein synthesis is still unknown.

Conclusion

Self-complementary circular codes are investigated here with the graph theory approach recently formulated in Fimmel et al. (2016). Self-complementary circular codes are involved in several pairing genetic processes, mainly DNAs–DNAs, DNAs–mRNAs, mRNAs–rRNAs, mRNAs–tRNAs and rRNAs–tRNAs. For the first time, all the self-complementary trinucleotide circular codes (words of 3 letters on a 4-letter alphabet) are identified here and several new mathematical properties are proven.

A code X is self-complementary if and only if its graph $\mathcal{G}(X)$ has a self-complementary set of vertices and for any vertex v, the outgoing degree $d^+(v)$ equals the ingoing degree $d^-(\overleftarrow{c(v)})$ of the complementary vertex. This statement is true for the self-complementary circular codes of sizes 18 and 20 trinucleotides and for the self-complementary comma-free codes of sizes 14 and 16 trinucleotides. (There are no self-complementary comma-free codes of sizes 18 and 20 trinucleotides.) For the self-complementary circular codes of sizes strictly less than 18 trinucleotides and for the self-complementary comma-free codes of sizes strictly less than 14 trinucleotides, this statement is not true. Despite a deep investigation from the authors, no explanation has been found for this interesting graph combinatorial problem which therefore remains open.

The lengths of the longest paths belong to the set $\{1, 2, 3, 4, 6, 8\}$ for the self-complementary circular codes and to the set $\{4, 6, 8\}$ for the 528 maximal (of size 20) self-complementary circular codes. The growth function of all self-complementary circular codes is also given. The structure of the longest paths is also determined for the maximal self-complementary circular codes.

The longest paths in the graphs $\mathcal{G}(X)$ determine the reading frame of self-complementary circular codes X. By applying this new theorem, the reading frame of the circular code X (1.1) identified in genes is retrieved after 15 nucleotides, i.e., 5 trinucleotides. The importance of this result lies in the fact that the reading frame number of a circular code even allows to retrieve the reading frame in arbitrary sequences of trinucleotides whenever a subsequence of at least 5 consecutive trinucleotides from X (called an X motif) is read. This theoretical result again suggests that the ribosome may retrieve the reading (correct) frame (circularity property of X) by using the X motifs from the circular code X in genes (Michel et al. 2017 and Fig. 9) which can pair (self-complementary property of X) with the X motifs found in tRNAs and rRNAs, in particular in the ribosome decoding center (Michel 2012, 2013; Soufi and Michel 2014, 2015). However, the experimental biological mechanism by which the ribosome involves the X motifs during genome decoding and protein synthesis is still unknown.

Appendix

Proof of Theorem 4.7 Claim (1): Let $l_{max}(X) = 4$ and assume that $l_1 \rightarrow d_1 \rightarrow l_2 \rightarrow d_2 \rightarrow l_3$ is a longest path in $\mathcal{G}(X)$. Since the path is maximal, there is no trinucleotide of the form

dl_1 and no trinucleotide of the form l_3d in X. It follows that $c(l_3) = l_1$ and $d_1, d_2 \in \{l_2, c(l_2)\}^2$. Note that all the nucleotides l_1, l_2, l_3 must be different by circularity. Thus, we have 4 possibilities for d_1, d_2, namely l_2l_2, $l_2c(l_2)$, $c(l_2)l_2$ and $c(l_2)c(l_2)$. As $l_2l_2l_2 \notin X$ by circularity, we have the following options for the 2 trinucleotides $d_1l_2 \in X$ and $l_2d_2 \in X$

$$\begin{array}{lllll} d_1l_2: & & l_2c(l_2)l_2 & c(l_2)l_2l_2 & c(l_2)c(l_2)l_2; \\ l_2d_2: & l_2l_2c(l_2) & l_2c(l_2)l_2 & l_2c(l_2)c(l_2). & \end{array}$$

If d_1l_2 or l_2d_2 is equal to $l_2c(l_2)l_2$ then self-complementarity yields $c(l_2)l_2c(l_2) \in X$ and the word $c(l_2)l_2c(l_2)l_2c(l_2)l_2$ contradicts circularity. Excluding the combinations $c(l_2)l_2l_2$, $l_2l_2c(l_2)$ and $c(l_2)c(l_2)l_2$, $l_2c(l_2)c(l_2)$ since the trinucleotides are obviously circular permutations of each other, only 2 combinations remain: $c(l_2)l_2l_2$, $l_2c(l_2)c(l_2)$ and $c(l_2)c(l_2)l_2$, $l_2l_2c(l_2)$. But also here, self-complementarity yields a contradiction to circularity since, for example, the complementary trinucleotide of $c(l_2)c(l_2)l_2$ is in the same equivalence class as $l_2l_2c(l_2)$.

Claim (2): Let $l_{max}(X) = 6$ and assume that $d_1 \to l_1 \to d_2 \to l_2 \to d_3 \to l_3 \to d_4$ is a longest path in $\mathcal{G}(X)$. By self-complementarity, there is the reversed complemented path

$$\overleftarrow{c(d_4)} \to c(l_3) \to \overleftarrow{c(d_3)} \to c(l_2) \to \overleftarrow{c(d_2)} \to c(l_1) \to \overleftarrow{c(d_1)}.$$

Now, the middle nucleotides l_2 and $c(l_2)$ of the 2 paths are either the pair A and T, or C and G. Therefore, it suffices to show that there are paths $A \to d \to T$ or $T \to d \to A$ and $C \to d \to G$ or $G \to d \to C$ in $\mathcal{G}(X)$; since then, we will obtain a path of length 8 combining the 2 paths, e.g.,

$$d_1 \to l_1 \to d_2 \to l_2 \to d \to c(l_2) \to \overleftarrow{c(d_2)} \to c(l_1) \to \overleftarrow{c(d_1)}$$

contradicting $l_{max}(X) = 6$. However, by maximality, the code X must contain exactly one trinucleotide of the class $\{ATT, TTA, TAT\}$ and its complementary trinucleotide as well as exactly one trinucleotide from the class $\{GCC, CCG, CGC\}$ and its complementary trinucleotide. It is easy to verify that in each case we obtain either a path of the form $A \to d \to T$ or $T \to d \to A$ and $C \to d \to G$ or $G \to d \to C$, e.g., if $ATT \in X$ then also $AAT \in X$ and we get the path $A \to AT \to T$ in $\mathcal{G}(X)$.

Claim (3): Let $l_{max}(X) = 8$ and assume that $l_1 \to d_1 \to l_2 \to d_2 \to l_3 \to d_3 \to l_4 \to d_4 \to l_5$ is the longest path in $\mathcal{G}(X)$. Then obviously, 2 out of the 5 nucleotides l_1, l_2, l_3, l_4, l_5 must be equal, which yields a cycle in $\mathcal{G}(X)$ contradicting the circularity of X. □

Proof of Theorem 5.11 Let $X \subseteq \mathcal{B}^3$ be a maximal self-complementary circular code and $\mathcal{G}(X)$ its associated graph. Since X is circular then $\mathcal{G}(X)$ is acyclic, so it has a path $p = p_{max}(X)$ of maximal length $l(p)$.

Claim (1): Assume that $p = d_1 \to b_1 \to \cdots \to b_k$, then any concatenation $d_ib_i \in X$. Choose any trinucleotide $c = s_1s_2s_3 \in X$. Then $(d_1b_1)\cdots(d_kb_k)(s_1s_2s_3) \in X^{k+1}$ and hence $(d_1b_1)\cdots(d_kb_k)s_1$ is a possible X-frame (for itself) with $t_b = \epsilon$ and $t_e = s_1$. Moreover, each concatenation b_id_{i+1} is also a trinucleotide in X, so $d_1(b_1d_2)\cdots(b_{k-1}d_k)b_ks_1$ is a second possible X-frame with $t_b = d_1$ and $t_e = b_ks_1$. Thus, $n_X \geq l_w(p) + 2$ since the sequence $d_1b_1 \cdots d_kb_ks_1$ has length $l_w(p) + 1$.

Now assume that $b_1 \cdots b_k$ is a sequence of nucleotides and assume that $k \geq l_w(p) + 2$ but $b_1 \cdots b_k$ has 2 different possible X-frames. We have to show a contradiction to conclude that $n_X = l_w(p) + 2$. Assume that $t_bu_1 \cdots u_lt_e$ and $t'_bu'_1 \cdots u'_mt'_e$ with $u_i, u'_i \in X$ and $t_b, t_e, t'_b, t'_e \in \left(\{\epsilon\} \cup \mathcal{B} \cup \mathcal{B}^2\right)$ are the 2 different possible X-frames. Obviously, $|\, t_bt_e \,| \leq 4$. If $|\, t_bt_e \,| = 4$ then by the difference of the 2 possible X-frames, we conclude that at least one of t'_b or t'_e has to have length ≥ 3, a contradiction to the definition of possible X-frame, or $|\, t'_bt'_e \,| \leq 3$. Hence, w.l.o.g. we assume that $|\, t_bt_e \,| \leq 3$. Consequently, $|\, u_1 \cdots u_l \,| \geq k - 3 \geq l_w(p) + 2 - 3 = l_w(p) - 1$ and hence $|\, u_1 \cdots u_l \,| \geq l_w(p) + 1$. We now have to distinguish cases:

(a) If $|\, t_bt_e \,| \leq 1$ then we even get $|\, u_1 \cdots u_l \,| \geq k - 1 \geq l_w(p) + 2 - 1 = l_w(p) + 1$ and hence $|\, u_1 \cdots u_l \,| \geq l_w(p)$. Thus, the path associated with the 2 possible X-frames has word-length at least $l_w(p) + 1$, a contradiction to the maximality of $l_w(p)$. In this case, the sequence $u_1 \cdots u_l$ could contain the sequence $u'_1 \cdots u'_m$ as a subsequence.

(b) If $|\, t_bt_e \,| \geq 2$ then the second possible X-frame is at least shifted by one with respect to the first possible X-frame, i.e., it must extend the sequence $u_1 \cdots u_l$ to the left or to the right. In this case, the sequence $u_1 \cdots u_l$ cannot contain the sequence $u'_1 \cdots u'_m$ as a subsequence. The path associated with the 2 possible X-frames has word-length at least $|\, u_1 \cdots u_l \,| + 1 \geq l_w(p) + 1$, again a contradiction to the maximality of $l_w(p)$.

Thus, $n_X = l(p) + 2$.

The case $p = b_1 \to d_1 \to \cdots \to d_k$ is symmetric and can be similarly dealt with.

Claim (2): Assume that $p = d_1 \to b_1 \to \cdots \to d_k$, then any concatenation $d_ib_i \in X$. As in Claim (1), $(d_1b_1)\cdots(d_{k-1}b_{k-1})d_k$ is a possible X-frame (for itself) with $t_b = \epsilon$ and $t_e = d_k$. Moreover, each concatenation b_id_{i+1} is a trinucleotide in X, so $d_1(b_1d_2)\cdots(b_{k-2}d_{k-1})(b_{k-1}d_k)$ is a second possible X-frame with $t_b = d_1$ and $t_e = \epsilon$. Thus, $n_X \geq l_w(p)$ since the sequence $d_1b_1 \cdots d_{k-1}b_{k-1}d_k$ has length $l_w(p)$.

Now assume that $b_1 \cdots b_k$ is a sequence of nucleotides and assume that $k \geq l_w(p) + 1$ but $b_1 \cdots b_k$ has 2 different possible

X-frames: $t_b u_1 \cdots u_l t_e$ and $t'_b u'_1 \cdots u'_m t'_e$ with $u_i, u'_i \in X$ and $t_b, t_e, t'_b, t'_e \in \left(\{\epsilon\} \cup \mathcal{B} \cup \mathcal{B}^2\right)$. As in Claim (1), we assume w.l.o.g. that $| t_b t_e | \leq 3$. We distinguish cases:

(a) If $| t_b t_e | = 0$ then $| u_1 \cdots u_l | \geq l_w(p) + 1$ and $u'_1 \cdots u'_m$ is a subsequence of $u_1 \cdots u_l$. Thus, the path associated with the 2 possible X-frames has word-length $l_w(p) + 1$ with the associated word $u_1 \cdots u_l$, a contradiction to the maximality of $l_w(p)$.

(b) If $| t_b t_e | = 1$ then $| u_1 \cdots u_l | \geq l_w(p)$. If the second possible X-frame is shifted by one with respect to the first one, then the path associated with the 2 possible X-frames has word-length $l_w(p) + 1$, again a contradiction to the maximality of $l_w(p)$. If the second possible X-frame is shifted by two, then the path associated with the 2 possible X-frames has word-length $l_w(p)$. However, in this case, the path starts with a dinucleotide and ends with a nucleotide, a contradiction to the structure of maximal paths which have to start and end with a dinucleotide.

(c) If $| t_b t_e | = 2$ then $| u_1 \cdots u_l | \geq l_w(p) - 1$. Again, we have to distinguish cases:

  (i) $| t_b | = 2$ and $| t_e | = 0$. Then the associated path to the 2 possible X-frames has word-length $l_w(p)$ and starts with a nucleotide but ends with a dinucleotide, a contradiction to the structure of maximal paths, or has word-length $l_w(p) + 1$, a contradiction to the maximality of $l_w(p)$.

  (ii) $| t_b | = 0$ and $| t_e | = 2$, as (i).

  (iii) $| t_b | = 1$ and $| t_e | = 1$. As above, if the second possible X-frame is shifted by one, then the path associated with the 2 possible X-frames has word-length $l_w(p)$ again starting with a nucleotide (u_1) and ending with a dinucleotide, a contradiction to the structure of maximal paths. If the second possible X-frame is shifted by two, then again the path associated with the 2 possible X-frames has word-length $l_w(p)$ starting with a nucleotide (u'_1) and ends with a dinucleotide.

(d) If $| t_b t_e | = 3$ then $| u_1 \cdots u_l | \geq l_w(p) - 2$. We distinguish two symmetric cases:

  (i) $| t_b | = 2$ and $| t_e | = 1$. If the second possible X-frame is shifted by one, then the path associated with the 2 possible X-frames has word-length $l_w(p) + 1$, a contradiction to the maximality of $l_w(p)$, or has word-length $l_w(p)$ but starting with a nucleotide and ending with a dinucleotide, a contradiction to the structure of maximal paths. If the second possible X-frame is shifted by two, then either the path associated with the 2 possible X-frames has word-length $l_w(p) + 1$, a contradiction to the maximality of $l_w(p)$, or has word-length $l_w(p) - 1$ starting with a nucleotide and ending with a nucleotide. But this case cannot exist unless the arrow-length of this path is at least the arrow-length of p, a contradiction to the maximality of p.

  (ii) $| t_b | = 1$ and $| t_e | = 2$, as (i).

Claim (3): Assume that $p = b_1 \to d_1 \to \cdots \to b_k$, then any concatenation $b_i d_i \in X$. Choose any 2 trinucleotides $c = s_1 s_2 s_3, c' = s'_1 s'_2 s'_3 \in X$. Then $(s'_1 s'_2 s'_3)(b_1 d_1) \cdots (d_k b_k)(s_1 s_2 s_3) \in X^{k+2}$ and hence $s'_3(b_1 d_1) \cdots (b_{k-1} d_{k-1}) b_k s_1$ is a possible X-frame (for itself) with $t_b = s'_3$ and $t_e = b_k s_1$. Moreover, each concatenation $d_i b_{i+1}$ is a trinucleotide in X, so $s'_3 b_1 (d_1 b_2) \cdots (d_{k-1} b_k) s_1$ is a second possible X-frame with $t_b = s'_3 b_1$ and $t_e = s_1$. Thus, $n_X \geq l_w(p) + 3$ since the sequence $s'_3 b_1 d_1 \cdots b_{k-1} d_{k-1} b_k s_1$ has length $l_w(p) + 2$.

Now assume that $b_1 \cdots b_k$ is a sequence of nucleotides with $k \geq l_w(p) + 3$ but $b_1 \cdots b_k$ has 2 different possible X-frames: $t_b u_1 \cdots u_l t_e$ and $t'_b u'_1 \cdots u'_m t'_e$ with $u_i, u'_i \in X$ and $t_b, t_e, t'_b, t'_e \in \left(\{\epsilon\} \cup \mathcal{B} \cup \mathcal{B}^2\right)$. As in Claim (1), we conclude that w.l.o.g. $| t_b t_e | \leq 3$ and hence $| u_1 \cdots u_l | \geq k - 3 \geq l_w(p) + 3 - 3 = l_w(p)$. Similar arguments as above show that the path associated with the 2 possible X-frames has word-length greater than $l_w(p)$, in contradiction to the maximality of p and $l_w(p)$. □

References

Arquès DG, Michel CJ (1996) A complementary circular code in the protein coding genes. J Theor Biol 182:45–58

Clark J, Holton DA (1991) A first look at graph theory. World Scientific, New Jersey

Crick FH, Brenner S, Klug A, Pieczenik G (1976) A speculation on the origin of protein synthesis. Orig Life 7:389–397

Crick FH, Griffith JS, Orgel LE (1957) Codes without commas. Proc Natl Acad Sci USA 43:416–421

Eigen M, Schuster P (1978) The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle. Naturwissenschaften 65:341–369

El Soufi K, Michel CJ (2014) Circular code motifs in the ribosome decoding center. Comput Biol Chem 52:9–17

El Soufi K, Michel CJ (2015) Circular code motifs near the ribosome decoding center. Comput Biol Chem 59:158–176

El Soufi K, Michel CJ (2016) Circular code motifs in genomes of eukaryotes. J Theor Biol 408:198–212

Fimmel E, Michel CJ, Strüngmann L (2016) n-Nucleotide circular codes in graph theory. Philos Trans R Soc A 374:20150058

Fimmel E, Michel CJ, Strüngmann L (2017) Strong comma-free codes in genetic information. Bull Math Biol 79:1796–1819

Golomb SW, Delbruck M, Welch LR (1958a) Construction and properties of comma-free codes. Biol Medd K Dan Vidensk Selsk 23:1–34

Golomb SW, Gordon B, Welch LR (1958b) Comma-free codes. Can J Math 10:202–209

Ikehara K (2002) Origins of gene, genetic code, protein and life: comprehensive view of life systems from a GNC-SNS primitive genetic code hypothesis. J Biosci 27:165–186

Michel CJ (2012) Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. Comput Biol Chem 37:24–37

Michel CJ (2013) Circular code motifs in transfer RNAs. Comput Biol Chem 45:17–29

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

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

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

Michel CJ, Pirillo G, Pirillo MA (2008) Varieties of comma free codes. Comput Math Appl 55:989–996

Shepherd JCW (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc Natl Acad Sci USA 78:1596–1600

Trifonov EN (1987) Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. J Mol Biol 194:643–652