Bull Math Biol (2017) 79:1796–1819
DOI 10.1007/s11538-017-0307-0

Society for Mathematical Biology

ORIGINAL ARTICLE

Strong Comma-Free Codes in Genetic Information

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

Received: 8 January 2017 / Accepted: 2 June 2017 / Published online: 22 June 2017

Abstract Comma-free codes constitute a class of circular codes, which has been widely studied, in particular by Golomb et al. (Biologiske Meddelelser, Kongelige Danske Videnskabernes Selskab 23:1–34, 1958a, Can J Math 10:202–209, 1958b), Michel et al. (Comput Math Appl 55:989–996, 2008a, Theor Comput Sci 401:17–26, 2008b, Inf Comput 212:55–63, 2012), Michel and Pirillo (Int J Comb 2011:659567, 2011), and Fimmel and Strüngmann (J Theor Biol 389:206–213, 2016). Based on a recent approach using graph theory to study circular codes Fimmel et al. (Philos Trans R Soc 374:20150058, 2016), a new class of circular codes, called strong comma-free codes, is identified. These codes detect a frameshift during the translation process immediately after a reading window of at most two nucleotides. We describe several combinatorial properties of strong comma-free codes: enumeration, maximality, self-complementarity and CF^3-property (comma-free property in all the three possible frames). These combinatorial results also highlight some new properties of the genetic code and its evolution. Each amino acid in the standard genetic code is coded by at least one strong comma-free code of size 1. There are 9 amino acids $S = \{Asn, Asp, Gln, Gly, Lys, Met, Phe, Pro, Trp\}$ among 20 such that for each amino acid from S, its synonymous trinucleotide set (excluding the necessary

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

Elena Fimmel
e.fimmel@hs-mannheim.de

Lutz Strüngmann
l.struengmann@hs-mannheim.de

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

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

periodic trinucleotides $\{AAA, CCC, GGG, TTT\}$) is a strong comma-free code. The primeval comma-free RNY code of Eigen and Schuster (Naturwissenschaften 65:341–369, 1978) is a self-complementary CF^3-code of size 16. Furthermore, it is the union of two strong comma-free codes of size 8 which are complementary to each other.

Keywords Strong comma-free codes · Enumeration · Maximality · Self-complementarity · CF^3-property · Genetic code

1 Introduction

The main class of trinucleotide codes, which is involved in the genetic code, are the circular codes and their proper subsets. About 60 years ago, before the discovery of the genetic code, proper subsets of circular codes, called comma-free codes, were proposed by Crick et al. (1957) for explaining how the reading of a sequence of trinucleotides could code the 20 amino acids. In particular, how the correct reading frame can be retrieved and maintained. The four nucleotides $\{A, C, G, T\}$ as well as the 16 dinucleotides {*AA,…,TT*} are simple codes, which are not appropriate for coding the 20 amino acids. However, trinucleotides induce a redundancy in their coding. Thus, Crick et al. (1957) conjectured that only 20 trinucleotides among the 64 possible trinucleotides {*AAA,. . .,TTT*} code the 20 amino acids. Such a bijective code implies that the coding trinucleotides are found only in one frame—the comma-freeness property. The determination of a set of 20 trinucleotides forming a comma-free code has several necessary conditions:

(i) A periodic trinucleotide from the set {*AAA, CCC, GGG, TTT*} must be excluded from such a code. Indeed, the concatenation of *AAA* with itself, for instance, does not allow the (original) reading frame to be retrieved as there are three possible decompositions: *. . .AAA, AAA, AAA. . .* (original frame), *. . .A, AAA, AAA, AA. . .* and *. . .AA, AAA, AAA, A. . .*, the commas showing the adopted decomposition.

(ii) Two non-periodic permuted trinucleotides, i.e., two trinucleotides related by a circular permutation, e.g., *ACG* and *CGA*, must also be excluded from such a code. Indeed, the concatenation of *ACG* with itself, for instance, does not allow the reading frame to be retrieved as there are two possible decompositions: *. . .ACG, ACG, ACG. . .* (original frame) and *. . .A, CGA, CGA, CG. . .*

Therefore, by excluding the four periodic trinucleotides and by gathering the 60 remaining trinucleotides in 20 classes of three trinucleotides such that, in each class, the three trinucleotides are deduced from each other by a circular permutation, e.g., *ACG*, *CGA* and *GAC*, we see that a comma-free code can contain only one trinucleotide from each class and thus has at most 20 trinucleotides. This trinucleotide number is identical to the amino acid number, thus leading to a code assigning one trinucleotide per amino acid without ambiguity. A few combinatorial results on trinucleotide comma-free codes were obtained by Golomb et al. (1958a, b). However, no trinucleotide comma-free code was identified in genes statistically. Furthermore, at the beginning of the 1960s, the discovery that the trinucleotide *TTT* (Nirenberg and Matthaei 1961), an excluded trinucleotide in a comma-free code, codes phenylalanine,

led to the abandonment of the concept of comma-freeness. For several biological reasons, in particular the interaction between mRNA and tRNA, this concept was again taken up later with two trinucleotide comma-free codes: *RRY* (Crick et al. 1976) and *RNY*={*RRY, RYY*} (Eigen and Schuster 1978; Shepherd 1981) with $R = \{A, G\}$, $Y = \{C, T\}$ and $N = \{A, C, G, T\}$.

The circular code theory initiated in 1996 proposes that genes are based on a circular code of 20 trinucleotides for retrieving, maintaining and synchronizing the reading frame as well as for coding amino acids (Michel 2012). It relies on two main results: (i) the identification of a maximal (20 words) C^3 self-complementary trinucleotide circular code X in genes of bacteria (15, 735, 053 genes, 5, 222, 267, 667 trinucleotides), archaea (282, 802 genes, 81, 460, 549 trinucleotides), eukaryotes (4, 356, 391 genes, 2, 406, 844, 838 trinucleotides), plasmids (575, 760 genes, 159, 169, 387 trinucleotides) and viruses (299, 401 genes, 66, 677, 580 trinucleotides) (Michel 2017, 2015; Arquès and Michel 1996) and (ii) the finding of X circular code motifs in tRNAs and rRNAs, in particular in the ribosome decoding center (Michel 2012; El Soufi and Michel 2014, 2015), and in the genomes of eukaryotes (El Soufi and Michel 2016). The universally conserved nucleotides A1492 and A1493 and the conserved nucleotide G530 in the ribosome decoding center are included in X circular code motifs.

The circular code X contains the 20 following trinucleotides

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

and codes the 12 amino acids

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

A trinucleotide circular code has the fundamental property to always retrieve the reading frame in any position of any sequence generated with the circular code. In particular, initiation and stop trinucleotides as well as any frame signals are not necessary to define the reading frame. Indeed, a window of a few nucleotides, whose nucleotide length depends on the class of circular codes, positioned anywhere in a sequence generated with the circular code always retrieves the reading frame (see, e.g., the example given in Fimmel et al. 2016). At most 13 consecutive nucleotides in any sequence generated by the circular code X are enough to always retrieve the reading frame. In a trinucleotide comma-free code, this window is equal to three nucleotides.

The combinatorial properties of comma-free codes and circular codes are important to understand some properties of the genetic code and its encoded amino acids as well as its evolution. Based on a recent approach using graph theory to study circular codes (Fimmel et al. 2016), a new class of circular codes, called strong comma-free codes, is identified. The class of strong comma-free codes is a proper subclass of the class of comma-free codes. The advantage of strong comma-free codes is that two consecutive nucleotides suffice for retrieving the correct reading frame in any sequence generated by the code.

We first describe several combinatorial properties of strong comma-free codes. Then, the combinatorial results obtained highlight some new properties of the genetic code and its evolution.

2 Notations and Definitions on Circular and Comma-Free Codes

We start by recalling notations and definitions that will be needed in the sequel. Let us denote the nucleotide 4-letter alphabet by $\mathcal{B} := \{A, C, G, T\}$ where A stands for adenine, C stands for cytosine, G stands for guanine and T stands for thymine. Thus, $\mathcal{B}^3$ is the set of the 64 *trinucleotides (codons or triletters)*, and the genetic code table can be seen as an assignment between $\mathcal{B}^3$ and the set of the 20 amino acids plus the stop signal. It had turned out in Fimmel et al. (2014) that the *symmetric group* on the set $\mathcal{B}$ plays an important role when describing error-detecting and error-correcting subcodes of $\mathcal{B}^3$. Recall that the symmetric group is defined as

$$S_{\mathcal{B}} = \{\pi : \mathcal{B} \to \mathcal{B} \mid \pi \text{ is bijective}\}$$

with the usual group operation given by composition of functions. The group $S_{\mathcal{B}}$ has 24 elements, and any bijective mapping $\pi : \mathcal{B} \to \mathcal{B}$ can be applied componentwise to $x \in \mathcal{B}^3$ and thus induces a bijective map $\mathcal{B}^3 \to \mathcal{B}^3$ which we will also denote by π. Regarding the complementary structure of the DNA double helix, the so-called *Strong/Weak (SW) or complementary* (c) transformation from $S_{\mathcal{B}}$

$$\text{SW (or } c) : (A, T, C, G) \to (T, A, G, C)$$

which exchanges A and T as well as C and G, plays an important biological role. A second symmetric group used in the following sections is the group $(S_3, \circ)$, the symmetric group on the three numbers 1, 2 and 3, where

$$S_3 := \{\alpha : \{1, 2, 3\} \to \{1, 2, 3\} \mid \alpha \text{ is bijective}\}$$

and $\circ$ denotes the composition of mappings. For instance, $(132) \in S_3$ is the permutation such that $1 \mapsto 3, 2 \mapsto 1, 3 \mapsto 2$. Clearly, any such α induces a mapping on the set of trinucleotides $\mathcal{B}^3$ by permuting the order of the bases in the trinucleotides, e.g., (132) transforms a trinucleotide (b_1, b_2, b_3) to the trinucleotide (b_3, b_1, b_2). Hence, given a subset $X \subseteq \mathcal{B}^3$ and a transformation $\pi : \mathcal{B}^3 \to \mathcal{B}^3$, $\pi(X)$ is the set of trinucleotides that are base transformations of triletters from X, whereas $\alpha(X)$ with $\alpha \in S_3$ contains the triletters that are obtained by permutations of the positions of bases in trinucleotides from X. As we will show below, this difference plays a crucial role from the biological point of view, e.g., by transformations from S_3 one can obtain the different frames of a code. Let us focus on the subgroup of *cyclical permutations* of $(S_3, \circ)$ denoted by

$$\mathcal{A}_3 := \{\alpha_0 = (1)(2)(3), \alpha_1 = (231), \alpha_2 = (312)\} \subset S_3.$$

$(\mathcal{A}_3, \circ)$ is known as the *alternating subgroup* of $(S_3, \circ)$ and contains the two shift operations α_1 and α_2. However, it does not contain the second important biologi-

cal transformation related to the antiparallel structure of the DNA double helix: the reversing permutation of the indices (31)(2). We will indicate this permutation by $\overleftarrow{}$ so that a given trinucleotide $x = (b_1, b_2, b_3) \in \mathcal{B}^3$ leads to $\overleftarrow{x} := (b_3, b_2, b_1)$. Obviously, for a trinucleotide (codon) $x \in \mathcal{B}^3$ the reversed complementary trinucleotide (anticodon) can now be described as $\overleftarrow{c(x)}$. We are now in the position to recall the definitions of self-complementarity, circularity (C^3-property), and comma-freeness of codes $X \subseteq \mathcal{B}^3$.

Definition 2.1 A trinucleotide code $X \subseteq \mathcal{B}^3$ is *self-complementary* if for each trinucleotide $x \in X$ its reversed complementary trinucleotide $\overleftarrow{c(x)}$ also belongs to X: $x \in X \Leftrightarrow \overleftarrow{c(x)} \in X$. We will also use the notation $X = \overleftarrow{c(X)}$.

As described in the introduction, there are two classes of codes that allow to detect frameshifts in the reading process of the ribosome and that have been studied recently.

Definition 2.2 A trinucleotide code $X \subseteq \mathcal{B}^3$ is *comma-free* if given any two trinucleotides $x_1, x_2 \in X$, any trinucleotide from the concatenation x_1x_2, except x_1, x_2 themselves, does not belong to X.

Recall that comma-free codes allow to retrieve the correct frame with at most three nucleotides.

Definition 2.3 A trinucleotide code $X \subseteq \mathcal{B}^3$ is *circular* if any word over the alphabet $\mathcal{B}$ written on a circle has at most one decomposition into words from X, i.e., after the last letter the word starts again from its first letter. A trinucleotide circular code X is called *maximal* if it contains 20 trinucleotides (i.e., $|X| = 20$).

Circular codes have weaker error-detecting properties than comma-free codes since they need at most 13 nucleotides to retrieve the correct reading frame. The maximal circular code found by Arquès and Michel (1996) even allows to retrieve the correct frame in the two shifted frames.

Definition 2.4 A trinucleotide code $X \subseteq \mathcal{B}^3$ is a C^3-*code* if X as well as X_1 and X_2 are circular, where $X_1 := \alpha_1(X)$ and $X_2 := \alpha_2(X)$. Similarly, a trinucleotide code $X \subseteq \mathcal{B}^3$ is CF^3-*code* if X, as well as X_1 and X_2 are comma-free, where $X_1 := \alpha_1(X)$ and $X_2 := \alpha_2(X)$.

In Fimmel et al. (2016), the three authors had introduced a new graph theoretical approach which relates a directed graph to any trinucleotide code. Recall from graph theory (Clark and Holton 1991) that a *graph* $\mathcal{G}$ consists of a finite set of *vertices (nodes)* V and a finite set of *edges* E. Here, an edge is a set $\{v, w\}$ of vertices from V. The graph is called *oriented* if the edges have an orientation, i.e., edges are considered to be ordered pairs $[v, w]$ in this case.

Definition 2.5 Let $X \subseteq \mathcal{B}^3$ be a trinucleotide code. We define a directed graph $\mathcal{G}(X) = (V(X), E(X))$ with set of vertices $V(X)$ and set of edges $E(X)$ as follows:

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

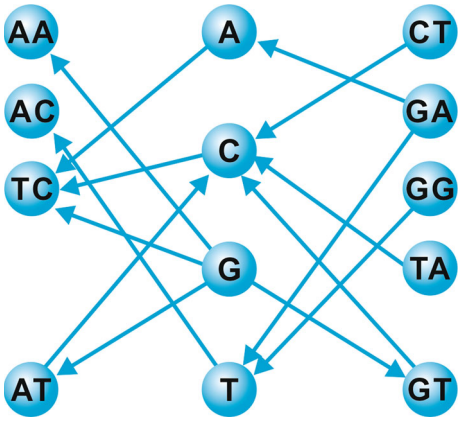

Fig. 1 The associated graph $\mathcal{G}(X)$ of the trinucleotide circular code $X = \{ATC, CTC, GAA, GAT, GGT, GTC, TAC\}$ (Color figure online)

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

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

The main results from Fimmel et al. (2016) stated that (i) a trinucleotide code X is circular if and only if its associated graph $\mathcal{G}(X)$ is *acyclic*, i.e., does not contain any cycles, and (ii) a trinucleotide code X is comma-free if and only if its graph $\mathcal{G}(X)$ is acyclic and the maximal length of a path in $\mathcal{G}(X)$ is at most 2.

Example 2.6 Figure 1 gives a trinucleotide circular code X and its associated graph $\mathcal{G}(X)$.

Motivated by the graph approach, we define a new class of codes.

Definition 2.7 A trinucleotide code $X \subseteq \mathcal{B}^3$ is called *strong comma-free* code if its associated graph $\mathcal{G}(X)$ has only oriented paths of length 1. A trinucleotide strong comma-free code X is *of maximal size* if there is *no* trinucleotide strong comma-free code X' with $|X'| > |X|$.

It is easily seen that strong comma-free codes have a stronger error-detecting property than comma-free codes in the sense that they can retrieve the correct reading frame within at most 2 nucleotides while comma-free codes need 3 nucleotides in general.

Example 2.8 Figure 2 gives a strong comma-free code X and its associated graph $\mathcal{G}(X)$.

We would like to remark that even a single trinucleotide (code of size 1) may not be strong comma-free. This is obviously true for trinucleotides of the form NNN but also for those of the form NMN since they yield a path $NM \rightarrow N \rightarrow MN$, i.e., a path of length 2.

Example 2.9 Figure 3 displays two graphs $\mathcal{G}(X)$ associated with two trinucleotide codes X of size 1, one code X which is not strong comma-free and one code X which is strong comma-free.

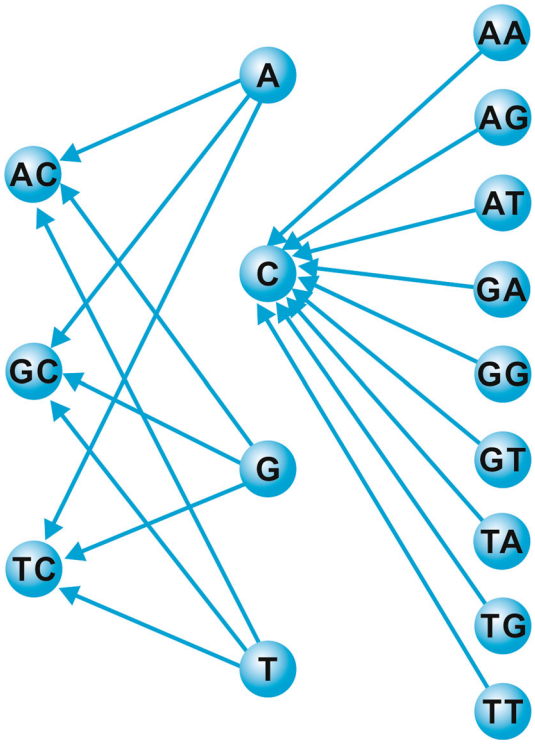

Fig. 2 The associated graph $\mathcal{G}(X)$ of the strong comma-free code $X = \{AAC, AGC, ATC, GAC, GGC, GTC, TAC, TGC, TTC\}$ (Color figure online)

(a) **(b)**

Fig. 3 The associated graphs $\mathcal{G}(X)$ of **a** the non-strong comma-free code $\{ACA\}$ and **b** the strong comma-free code $\{GCT\}$ (Color figure online)

3 Properties and Classification of Strong Comma-Free Codes

In this section, we completely classify the strong comma-free codes of maximal size as well as those that are self-complementary. Moreover, we give the growth function of strong comma-free codes and some handy criteria to test strong comma-freeness.

3.1 Combinatorial Properties of Strong Comma-Free Codes

We first aim for some criteria for testing strong comma-freeness. Let us start with a technical definition that will also be helpful to determine the structure of certain strong comma-free codes in the next sections.

Definition 3.1 We will denote by π_i ($i = 1, 2, 3$) the *projection* $\pi_i : \mathcal{B}^3 \to \mathcal{B}$ which assigns to each trinucleotide $x = N_1N_2N_3 \in \mathcal{B}^3$ its ith coordinate $N_i \in \mathcal{B}$. Moreover, for a subset $X \subseteq \mathcal{B}^3$, the image $\pi_i(X) \subseteq \mathcal{B}$ of X under π_i denotes the set of all ith coordinates of the elements of X.

Note that for any subset $X \subseteq \mathcal{B}^3$, the sets $\pi_1(X)$, $\pi_2(X)$ and $\pi_3(X)$ can contain at most 4 elements. For instance, $\pi_1(CCG) = C$, $\pi_2(TGA) = G$, $\pi_3(TAG) = G$ and $\pi_1(\{CCG, TGA, GAA, GTC, TCT\}) = \{C, G, T\}$. While in general, the sets $\pi_i(X)$ may not be disjoint for a code X, a strong comma-free code forces disjointness of the first and last coordinates.

Lemma 3.2 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide strong comma-free code. Then $\pi_1(X) \cap \pi_3(X) = \emptyset$.*

Proof Let us assume that $U = \pi_1(X) \cap \pi_3(X) \neq \emptyset$ and assume that $N_1 \in U$. Then $N_1X_1Y_1, X_2Y_2N_1 \in X$ for some $X_1, X_2, Y_1, Y_2 \in \mathcal{B}$. Hence, there is an oriented path of length at least 2 in the associated graph $\mathcal{G}(X)$, namely $X_2Y_2 \to N_1 \to X_1Y_1$, and thus, X is not strongly comma-free—a contradiction. □

The converse in the above Lemma 3.2 is false as the following example shows.

Example 3.3 Let $X = \{AAC, ACC\}$. Then $\pi_1(X) \cap \pi_3(X) = \emptyset$ but X is not a strong comma-free code since its associated graph $\mathcal{G}(X)$ has an oriented path of length 2: $A \to AC \to C$.

Lemma 3.2 shows that strong comma-freeness is a stronger property than comma-freeness since the condition $\pi_1(X) \cap \pi_3(X) = \emptyset$ is necessary for strong comma-free codes but already sufficient for being comma-free. Precisely, the following statement is true (compare also Lemma 4.4 in Fimmel and Strüngmann 2015).

Lemma 3.4 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide code. If $\pi_1(X) \cap \pi_3(X) = \emptyset$ or $\pi_1(X) \cap \pi_2(X) = \emptyset$ or $\pi_2(X) \cap \pi_3(X) = \emptyset$, then X is comma-free.*

Proof Without loss of generality, we assume that $X \subseteq \mathcal{B}^3$ is a trinucleotide code such that $\pi_1(X) \cap \pi_3(X) = \emptyset$ since the two other cases are symmetric. Let $t_1 = X_1X_2X_3 \in X$ and $t_2 = Y_1Y_2Y_3 \in X$ be two trinucleotides of X and consider the concatenation $X_1X_2X_3Y_1Y_2Y_3$. Obviously, $X_2X_3Y_1 \notin X$ since $Y_1 \notin \pi_3(X)$ and $X_3Y_1Y_2 \notin X$ since $X_3 \notin \pi_1(X)$. Thus, X is comma-free. □

However, the condition $\pi_1(X) \cap \pi_3(X) = \emptyset$ (as well as the two other conditions) is not necessary for comma-freeness as we demonstrate next. Nevertheless, note that for strong comma-free codes we do not need to have $\pi_1(X) \cap \pi_2(X) = \emptyset$ or $\pi_2(X) \cap \pi_3(X) = \emptyset$. Therefore, we now focus on the condition $\pi_1(X) \cap \pi_3(X) = \emptyset$. However, we will come back to codes satisfying one of the disjointness conditions from Lemma 3.4 in Sect. 3.4.

Example 3.5 The code $X = \{CAG, GTC\}$ is obviously comma-free but $\pi_1(X) \cap \pi_3(X) \neq \emptyset$.

Putting Lemmas 3.2 and 3.4 together, we see that the class of strong comma-free codes is strictly contained in the class of all codes satisfying $\pi_1(X) \cap \pi_3(X) = \emptyset$, which is again strictly contained in the class of comma-free codes:

$$X \text{ strong comma-free} \Rightarrow \pi_1(X) \cap \pi_3(X) = \emptyset \Rightarrow X \text{ comma-free}.$$

We aim now for a sufficient condition for strong comma-freeness.

Lemma 3.6 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide code such that $\pi_1(X) \cap \pi_3(X) = \emptyset$. If in addition one of the two conditions $\pi_1(X) \cap \pi_2(X) = \emptyset$ or $\pi_2(X) \cap \pi_3(X) = \emptyset$ is satisfied, then X is strong comma-free.*

Proof Let $X \subseteq \mathcal{B}^3$ be a trinucleotide code. There are two possibilities to get an oriented path of length at least 2 in the associated graph $\mathcal{G}(X) : N_1N_2 \to N_3 \to N_2'N_3'$ or $N_1 \to N_2N_3 \to N_3'$ where $N_i, N_i' \in \mathcal{B}$. The first possibility is excluded by the condition $\pi_1(X) \cap \pi_3(X) = \emptyset$ and the second one by each of the additional conditions $\pi_1(X) \cap \pi_2(X) = \emptyset$ or $\pi_2(X) \cap \pi_3(X) = \emptyset$. □

Again, the converse direction in Lemma 3.6 does not hold:

Example 3.7 Let $X = \{AAC, ATG, AGC\}$. X is a strong comma-free code but $\pi_1(X) \cap \pi_2(X) \neq \emptyset$ and $\pi_2(X) \cap \pi_3(X) \neq \emptyset$.

Now we focus on the question how symmetrical transformations of a given code affect its strong comma-freeness. As for circular or comma-free codes (Fimmel et al. 2014), cyclical permutations of a trinucleotide strong comma-free code X do not preserve strong comma-freeness in general. In fact, the four permutations of the letters $\alpha \in S_3 \setminus \{id, \leftarrow\}$ (except the identical and reversing permutations) do not guarantee comma-freeness of $\alpha(X)$. Let us denote the four permutations by $\alpha_1 = (231)$, $\alpha_2 = (312)$, $p_1 = (132)$ and $p_2 = (213)$, and consider the following code $X = \{ACC, AAG\}$ with $\alpha_1(X) = \{CCA, AGA\}$, $\alpha_2(X) = \{CAC, GAA\}$, $p_1(X) = \{ACC, AGA\}$ and $p_2(X) = \{CAC, AAG\}$. According to Lemma 3.2, the code X is strong comma-free, but none of the codes $\alpha_1(X)$, $\alpha_2(X)$, $p_1(X)$, $p_2(X)$ are strong comma-free. However, by reversing each trinucleotide of a strong comma-free code, the reversed code is also strong comma-free code (obvious check).

Moreover, any transformation $\pi \in S_{\mathcal{B}}$ turns a graph $\mathcal{G}(X)$ associated with a trinucleotide code X naturally into an isomorphic graph $\mathcal{G}(\pi(X))$ of the image code $\pi(X)$. Thus, if $\mathcal{G}(X)$ has no path longer than one edge, i.e., X is strong comma-free, then this also holds for $\mathcal{G}(\pi(X))$, and hence, $\pi(X)$ is strong comma-free. The theorem below collects these facts.

Theorem 3.8 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide strong comma-free code. Then, the following statements are true.*

(1) *The identical and reversing permutations are the sole permutations of the positions of trinucleotide letters which preserve strong comma-freeness of a code.*
(2) *For every $\pi \in S_{\mathcal{B}}$, $\pi(X)$ is a also a trinucleotide strong comma-free code.*

Proof Clear. □

The above Theorem 3.8 gives an easy way to construct new strong comma-free codes once a strong comma-free code X is given, just by applying the 24 bijective transformations π from $S_{\mathcal{B}}$. However, some of the new codes $\pi(X)$ may coincide. Also, an important property of complementary codes related to the DNA double helix follows as an immediate consequence of Theorem 3.8.

Corollary 3.9 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide strong comma-free code. Then the complementary code $\overleftarrow{c(X)}$ of X is also strong comma-free.*

Proof Follows from Theorem 3.8. □

In the next subsections, we will investigate strong comma-free codes of maximal size and their structure, self-complementary strong comma-free codes and so-called CF^3-codes.

3.2 Strong Comma-Free Codes of Maximal Size

A computer calculation determined the numbers of strong comma-free codes of cardinalities 1–9. This growth function is given in Table 1. We will prove in Theorem 3.11 that 9 is the maximal size of a strong comma-free code.

We first explain the first row of Table 1. Recall from the first section that the trinucleotides of the form NMN cannot belong to a strong comma-free code.

Lemma 3.10 *The number of trinucleotide strong comma-free codes of size* 1 *is* 48.

Proof According to Lemma 3.2, the necessary condition for the codes with a single trinucleotide is at the same time the sufficient one. Thus, we have four possibilities to choose the first nucleotide and three possibilities for the third one, and the second nucleotide can be chosen arbitrarily. Consequently, there are $4 \times 3 \times 4 = 48$ trinucleotide strong comma-free codes of size 1. □

Now we describe all the trinucleotide strong comma-free codes of maximal cardinality explaining the last row of Table 1.

Table 1 Growth function (numbers) of trinucleotide strong comma-free codes of cardinalities 1–9

Cardinality	Number
1	48
2	564
3	2432
4	4968
5	5424
6	3288
7	1080
8	168
9	8

Theorem 3.11 *All trinucleotide strong comma-free codes of maximal size have the following form*

$$X = \{N_1YZ \mid Z, Y \in \mathcal{B} \setminus \{N_1\}\} \text{ or } X = \{YZN_1 \mid Z, Y \in \mathcal{B} \setminus \{N_1\}\}$$

for some $N_1 \in \mathcal{B}$. *In particular, the maximal size of a strong comma-free code is* 9, *and there are exactly* 8 *trinucleotide strong comma-free codes of this maximal cardinality.*

Proof Clearly, codes of the stated form $X = \{N_1YZ \mid Z, Y \in \mathcal{B} \setminus \{N_1\}\}$ or $X = \{YZN_1 \mid Z, Y \in \mathcal{B} \setminus \{N_1\}\}$, $N_1 \in \mathcal{B}$, are strongly comma-free by Lemma 3.6 since $\pi_1(X) \cap \pi_3(X) = \emptyset$ and $\pi_1(X) \cap \pi_2(X) = \emptyset$.

Now, let X be a strong comma-free code of maximal size and assume that $|X| > 9$. By Lemma 3.2, we must have $\pi_1(X) \cap \pi_3(X) = \emptyset$. Hence, either $|\pi_1(X)| = |\pi_3(X)| = 2$ or $\{|\pi_1(X)|, |\pi_3(X)|\} = \{1, 3\}$.

Case 1 $\pi_1(X) = \{N_1\}$ ($\pi_3(X) = \{N_1\}$, respectively) and thus $|X| \leq 12$. Since $|X| > 9$, there must be a trinucleotide $N_1N_1Y \in X$ ($YN_1N_1 \in X$, respectively) for some $Y \in \mathcal{B}$. However, then none of the trinucleotides N_1YZ (ZYN_1, respectively) for $Z \in \mathcal{B}\setminus\{N_1\}$ can be in X since otherwise we obtain an oriented path of length 2 in $\mathcal{G}(X)$, namely $N_1 \to N_1Y \to Z$ ($Z \to YN_1 \to N_1$, respectively), which contradicts the strong comma-freeness of X. Hence, $|X| \leq 12 - 3 = 9$—a contradiction.

Case 2 $\pi_1(X) = \{N_1, N_2\}$ and $\pi_3(X) = \{N_3, N_4\}$. Then, $|X| \leq 4 \cdot |\pi_2(X)|$ and $|X| > 9$ implies that $|\pi_2(X)| \geq 3$. Assume that $N \in \pi_2(X)$, then there is a trinucleotide $N_1NY \in X$ or $N_2NY \in X$ for some $Y \in \mathcal{B}$. As in Case 1, this excludes two trinucleotides from X, namely NYN_3 and NYN_4 if $N = N_1$ or $N = N_2$, N_1N_1N and N_2N_1N or N_1N_2N and N_2N_2N if $N = N_3$ or $N = N_4$. All these trinucleotides are different, and hence, six trinucleotides are excluded if $|\pi_2(X)| = 3$ and eight trinucleotides if $|\pi_2(X)| = 4$. Thus, $|X| \leq 4 \cdot |\pi_2(X)| - 2 \cdot |\pi_2(X)| \leq 8$—a contradiction. Therefore, case 2 cannot exist for strong comma-free codes of maximal size. □

As a consequence of Theorem 3.11, we can now explicitly list all strong comma-free codes of maximal size.

List of strong comma-free codes of maximal size 3.12 *The 8 strong comma-free codes of maximal size 9 are:*

$$\begin{aligned}
X_1 &= \{AAC, AGC, ATC, GAC, GGC, GTC, TAC, TGC, TTC\},\\
X_2 &= \{AAG, ACG, ATG, CAG, CCG, CTG, TAG, TCG, TTG\},\\
X_3 &= \{AAT, ACT, AGT, CAT, CCT, CGT, GAT, GCT, GGT\},\\
X_4 &= \{ACC, ACG, ACT, AGC, AGG, AGT, ATC, ATG, ATT\},\\
X_5 &= \{CAA, CAG, CAT, CGA, CGG, CGT, CTA, CTG, CTT\},\\
X_6 &= \{CCA, CGA, CTA, GCA, GGA, GTA, TCA, TGA, TTA\},\\
X_7 &= \{GAA, GAC, GAT, GCA, GCC, GCT, GTA, GTC, GTT\},\\
X_8 &= \{TAA, TAC, TAG, TCA, TCC, TCG, TGA, TGC, TGG\}.
\end{aligned}$$

Remark 3.13 The graph given in Fig. 2 is associated with the strong comma-free code X_1 from List 3.12.

We now turn to strong comma-free codes of size 8. We will see in particular that not every such codes can be embedded into a strong comma-free code of maximal cardinality.

Lemma 3.14 *The number of trinucleotide strong comma-free codes of size* 8 *is* 168.

Proof Clearly, removing one trinucleotide from a strong comma-free code yields again a strong comma-free code. Thus, we obtain $72 = 8 \times 9$ strong comma-free codes of size 8 by simply removing an arbitrary trinucleotide from one of the 8 strong comma-free codes $X_1, \ldots, X_8$ of maximal size. It is easy to see that all these codes are different by the structure of the strong comma-free codes of maximal size.

The remaining $96 = 2^4 \times 6$ strong comma-free codes of size 8 are generated by the codes of the form

$$X = \{N_1 Y N_3, N_1 Y N_4, N_2 Y N_3, N_2 Y N_4 | Y \in B\}$$

where $\mathcal{B} = \{N_1, N_2, N_3, N_4\}$. From each of these 6 codes, $16 = 2^4$ strong comma-free codes of size 8 can be generated by removing the two forbidden trinucleotides whenever $Y \in \pi_2(X)$—see Case 2 in Theorem 3.11. Thus, there are $72 + 96 = 168$ strong comma-free codes of size 8, and obviously, not all of them can be embedded into a strong comma-free code of maximal size. □

We finish this section by giving an example of a strong comma-free code that cannot be embedded into a larger strong comma-free code.

Example 3.15 Figure 4 shows the associated graph $\mathcal{G}(X)$ of a strong comma-free code X of size 8 which is not included in any of the strong comma-free codes of maximal size 9 from List 3.12.

3.3 Self-Complementary Strong Comma-Free Codes

We now study strong comma-free codes which are self-complementary (see Definition 2.1). A computer calculation shows that such codes must be of size either 2 or 4. Table 2 gives the numbers of self-complementary strong comma-free codes of cardinalities 2 and 4. In particular, none of the strong comma-free codes of size 6 and 8 can be self-complementary. Our main results in this subsection give a constructive proof of the numbers in Table 2.

The following theorem gives a theoretical basis for Table 2.

Theorem 3.16 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide code.*

(1) *If X is a self-complementary strong comma-free code with $|X| = 2$, then*

$$X = \{N_1 N_2 N_3, c(N_3)c(N_2)c(N_1)\}$$

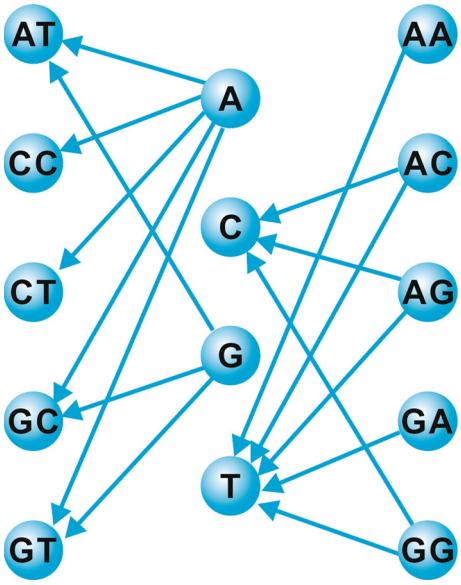

Fig. 4 The associated graph $\mathcal{G}(X)$ of the strong comma-free code $X = \{AAT, ACC, ACT, AGC, AGT, GAT, GGC, GGT\}$ of size 8 which is not included in any of the strong comma-free codes of size 9 from List 3.12 (Color figure online)

Table 2 Growth function (numbers) of trinucleotide self-complementary strong comma-free codes of cardinalities 2 and 4

Cardinality	Number
2	12
4	8

for some $N_1, N_2, N_3 \in \mathcal{B}$ *with* $N_1 \neq N_3, c(N_1) \neq N_2 \neq c(N_3)$. *In total, there are 12 self-complementary strong comma-free codes of size* 2.

(2) *If* X *is a self-complementary strong comma-free code with* $|X| = 4$, *then*

$$X = \{N_1N_2c(N_1), N_1c(N_2)c(N_1), N_2N_2c(N_1), N_1c(N_2)c(N_2)\}$$

for some $N_1, N_2, N_3 \in \mathcal{B}$ *with* $N_1 \neq N_2 \neq c(N_1)$. *In total, there are* 8 *self-complementary strong comma-free codes of size* 4.

(3) *There is no trinucleotide self-complementary strong comma-free code* $X \subseteq \mathcal{B}^3$ *with* $|X| > 4$.

Proof Let X be a trinucleotide self-complementary strong comma-free code. Certainly, the size of X must be even by self-complementarity. We distinguish cases. □

Claim 1 If X is of size 2, then $X = \{N_1N_2N_3, c(N_3)c(N_2)c(N_1)\}$ for some $N_1, N_2, N_3 \in \mathcal{B}$. Since X is strong comma-free, Lemma 3.2 implies that $N_1 \neq N_3$ which is equivalent to $c(N_1) \neq c(N_3)$. Therefore, we do not have any directed paths of the form $M_1M_2 \to M_3 \to M_4M_5$ in $\mathcal{G}(X)$. If there is an oriented path of the form $M_1 \to M_1M_2 \to M_3$ in $\mathcal{G}(X)$, then this implies that $N_2N_3 = c(N_3)c(N_2)$ or

$N_1N_2 = c(N_2)c(N_1)$. Hence, strong comma-freeness implies that $N_2 \neq c(N_3)$ and $N_2 \neq c(N_1)$. Clearly, if a code is of the stated form, then it is strong comma-free.

Thus, there are $(4 \cdot 2 \cdot 3) : 2 = 12$ self-complementary strong comma-free codes of size 2. Indeed, we choose the first position of a 'first' trinucleotide arbitrary, for the third position there are three possibilities, for the second position only two possibilities left and the second trinucleotide in the code is automatically deduced. The same code can be obtained if we begin with the 'second' trinucleotide; hence, we have to divide by 2 the number obtained.

Claim 2 Let X be of size 4. We have $\pi_1(X) = c(\pi_3(X))$ due to self-complementarity. Assume that $|\pi_1(X)| = |c(\pi_3(X))| = 1$, e.g., $\pi_1(X) = \{N_1\}$, then $X = \{N_1N_2c(N_1)|N_2 \in \mathcal{B}\}$. Thus, there is an oriented path of length 2, namely $N_1 \rightarrow N_1c(N_1) \rightarrow c(N_1)$ in $\mathcal{G}(X)$—contradicting strong comma-freeness. We conclude that $|\pi_1(X)| = |c(\pi_3(X))| = 2$ since $\pi_1(X) \cap \pi_3(X) = \emptyset$. w.l.o.g., we can assume that $\pi_1(X) = \{N_1, N_2\}$ and $\pi_3(X) = \{c(N_1), c(N_2)\}$ for $N_1, N_2 \in \mathcal{B}$. Moreover, since a subcode of a strong comma-free code is also strong comma-free, the conditions from (1) stay true also in the case of a code with $|X| = 4$.

Next we observe that X cannot have the two forms

$$X = \{N_1N_2c(N_1), N_1c(N_2)c(N_1), N_2N_1c(N_2), N_2c(N_1)c(N_2)|N_1, N_2 \in \mathcal{B},\ N_1 \neq N_2 \neq c(N_1)\} \quad (*)$$

since there is an oriented path of length 2, namely $N_1 \rightarrow N_2c(N_1) \rightarrow c(N_2)$ in $\mathcal{G}(X)$, and

$$X = \{N_1N_2c(N_2), N_2c(N_2)c(N_1), N_2N_2c(N_1), N_1c(N_2)c(N_2)|N_1, N_2 \in \mathcal{B},\ N_1 \neq N_2 \neq c(N_1)\} \quad (**)$$

since there is an oriented path of length 2, namely $N_1 \rightarrow N_2c(N_2) \rightarrow c(N_1)$ in $\mathcal{G}(X)$. The only possibility left is that X has the form

$$X = \{N_1N_2c(N_1), N_1c(N_2)c(N_1), N_2N_2c(N_1), N_1c(N_2)c(N_2)|N_1, N_2 \in \mathcal{B},\ N_1 \neq N_2 \neq c(N_1)\}$$

as stated above. X is self-complementary and strong comma-free simultaneously. To construct such a code, we have four possibilities to choose N_1 and then two possibilities to choose N_2. In total, there are $4 \cdot 2 = 8$ self-complementary strong comma-free codes of size 4.

Claim 3 Assume that X has size ≥ 6. As in (2), we conclude that $|\pi_1(X)| = |c(\pi_3(X))| = 2$ and $\pi_1(X) = c(\pi_3(X))$. Assume w.l.o.g. that $\pi_1(X) = \{N_1, N_2\}$ and $\pi_3(X) = \{c(N_1), c(N_2)\}$ for $N_1, N_2 \in \mathcal{B}$. Then, there exists a self-complementary strong comma-free subcode $Y \subseteq X$ with $|Y| = 4$. Y must have the form described in (2). According to (1), only two pairs $N_2N_1c(N_2)$ and $N_2c(N_1)c(N_2)$ or $N_1N_2c(N_2)$ and $N_2c(N_2)c(N_1)$ can be considered for the choice of the last pair trinucleotide—complementary trinucleotide. However, due to arguments (*) and (**) above, in both cases there are oriented paths of length 2 in $\mathcal{G}(X)$. This leads us to a contradiction. □

We now give a complete list of all self-complementary strong comma-free codes.

List of self-complementary comma-free codes 3.17 *The 12 self-complementary strong comma-free codes of size 2 are:*

$$\{AAC, GTT\},\ \{AAG, CTT\},\ \{ACC, GGT\},\ \{ACT, AGT\},$$
$$\{AGG, CCT\},\ \{CAA, TTG\},\ \{CAG, CTG\},\ \{CCA, TGG\},$$
$$\{GAA, TTC\},\ \{GAC, GTC\},\ \{GGA, TCC\},\ \{TCA, TGA\}.$$

The 8 self-complementary strong comma-free codes of size 4 are:

$$\{AAC, GAC, GTC, GTT\},\ \{AAG, CAG, CTG, CTT\},$$
$$\{ACC, ACT, AGT, GGT\},\ \{ACT, AGG, AGT, CCT\},$$
$$\{CAA, CAG, CTG, TTG\},\ \{CCA, TCA, TGA, TGG\},$$
$$\{GAA, GAC, GTC, TTC\},\ \{GGA, TCA, TCC, TGA\}.$$

We conclude this subsection with some examples of graphs associated with the self-complementary strong comma-free codes.

Example 3.18 Figures 5 and 6 give the associated graphs of some of the self-complementary strong comma-free codes from List 3.17.

3.4 CF^3-Property of Strong Comma-Free Codes

As discussed in Introduction, the maximal self-complementary circular code X identified in genes (Arquès and Michel 1996) has the additional property that the cyclically permuted codes $\alpha_1(X)$ and $\alpha_2(X)$ are circular as well. This property is called the C^3-property. Similarly, one can ask whether the strong comma-free codes admit a

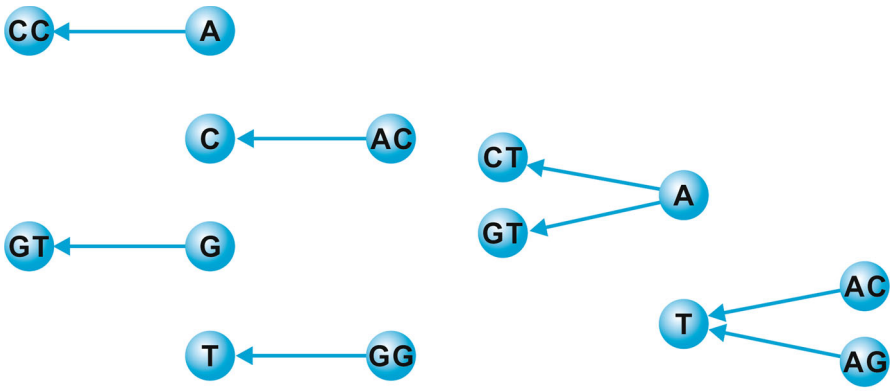

Fig. 5 The associated graphs $\mathcal{G}(X)$ of the self-complementary strong comma-free codes **a** $X = \{ACC, GGT\}$ of size 2 and **b** $X = \{ACT, AGT\}$ of size 2 (Color figure online)

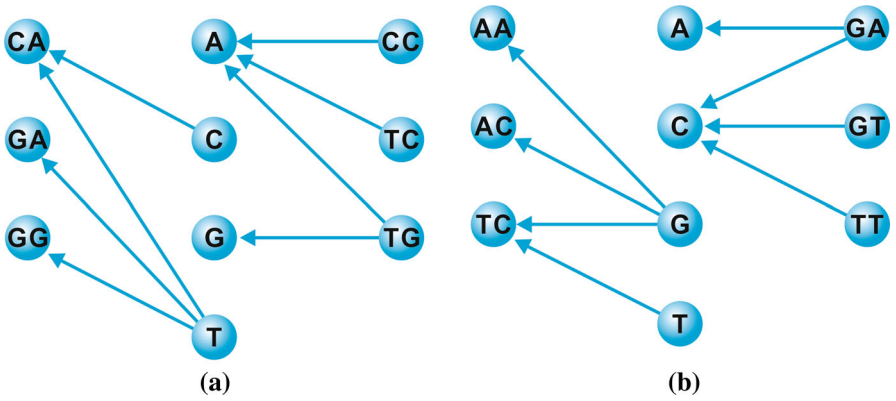

Fig. 6 The associated graphs $\mathcal{G}(X)$ of the self-complementary strong comma-free codes **a** $\{CCA, TCA, TGA, TGG\}$ of size 4 and **b** $\{GAA, GAC, GTC, TTC\}$ of size 4 (Color figure online)

corresponding property. The main result will show that indeed a strong comma-free code is a CF^3-code, i.e., its cyclically permuted codes are even comma-free—but not strong comma-free.

We start with an example showing that in general a cyclically permuted code of a strong comma-free code is not necessarily strong comma-free again.

Example 3.19 Consider $X = \{ACG, AAG\}$ with $\alpha_1(X) = \{CGA, AGA\}$. According to Lemma 3.2, X is strong comma-free while $\alpha_1(X)$ is not.

The main theorem now shows that the class of codes with disjoint first and third coordinates (first and second, or second and third coordinates, respectively) are in the class of CF^3-codes (see Definition 2.4).

Theorem 3.20 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide code such that $\pi_1(X) \cap \pi_3(X) = \emptyset$ or $\pi_1(X) \cap \pi_2(X) = \emptyset$ or $\pi_2(X) \cap \pi_3(X) = \emptyset$. Then X is a CF^3-code.*

Proof The code X itself is comma-free according to Lemma 3.4. Moreover, if one of the conditions is satisfied for X, then the two other conditions are satisfied for $\alpha_1(X)$ and $\alpha_2(X)$. Thus, $\alpha_1(X)$ and $\alpha_2(X)$ are also comma-free by Lemma 3.4. □

The next example shows that the converse implication in Theorem 3.20 does not hold.

Example 3.21 Let $X = \{AAG, CTA\}$, then X is a CF^3-code but X satisfies the three conditions $\pi_1(X) \cap \pi_2(X) = \pi_2(X) \cap \pi_3(X) = \pi_1(X) \cap \pi_3(X) = \{A\} \neq \emptyset$.

Since any transformation of nucleotides preserves comma-freeness, we immediately obtain

Lemma 3.22 *If X is a CF^3-code then and $\pi \in S_{\mathcal{B}}$ is a transformation, then $\pi(X)$ is also a CF^3-code.*

Proof For all $\alpha \in S_3$ and $\pi \in S_{\mathcal{B}}$, the property $\alpha(\pi(X)) = \pi(\alpha(X))$ is true (see also Fimmel et al. 2014). □

This argument is in the line of a number of results, showing that systematical exchanges of nucleotides in a code X preserve one-to-one its error-detecting properties (compare also Theorem 3.8 and the corresponding results for circular and C^3-codes from Fimmel et al. 2014).

Corollary 3.23 *Let $X \subseteq \mathcal{B}^3$ be a trinucleotide strong comma-free code. Then X is a CF^3-code.*

Proof Follows from Lemma 3.2 and Theorem 3.20. □

The next example shows that there are CF^3-codes where X and its cyclically permuted codes $\alpha_1(X)$ and $\alpha_2(X)$ are not strong comma-free; hence, the class of CF^3-codes is strictly larger than the class of strong comma-free codes.

Example 3.24 Let us consider $X = \{AAG, GCC\}$. Then $X, \alpha_1(X) = \{AGA, CCG\}$ and $\alpha_2(X) = \{GAA, CGC\}$ are comma-free but not strong comma-free (see Lemma 3.2).

However, the class of all CF^3-codes obviously lies in the class of all C^3-codes. As a consequence, we have the following result.

Corollary 3.25 *Any strong comma-free code is C^3.*

Proof The claim obviously follows from Corollary 3.23 since every comma-free code is also circular. □

We have seen that the classes of strong comma-free codes ($SCFC$), comma-free codes (CFC), circular codes (CC), CF^3-codes (CF^3C) and C^3-codes (C^3C) form the following hierarchy:

(i) $SCFC \subset CF^3C \subset CFC \subset CC$.
(ii) $SCFC \subset CF^3C \subset C^3C \subset CC$.

This hierarchy is also displayed in Fig. 7.

3.5 Strong Comma-Free Codes and the Standard Genetic Code

In this section, we investigate which amino acids are coded by strong comma-free codes. Since even codes of size one can be non-strong comma-free, the first result is important.

Proposition 3.26 *Each amino acid in the standard genetic code is coded by at least one strong comma-free code of size 1.*

Proof By excluding the four periodic trinucleotides $\{AAA, CCC, GGG, TTT\}$ and the 12 trinucleotides $\{ACA\}$ coding Thr, $\{AGA\}$ coding Arg, $\{ATA\}$ coding Ile, $\{CAC\}$ coding His, $\{CGC\}$ coding Arg, $\{CTC\}$ coding Leu, $\{GAG\}$ coding Glu, $\{GCG\}$ coding Ala, $\{GTG\}$ coding Val, $\{TAT\}$ coding Tyr, $\{TCT\}$ coding Ser and $\{TGT\}$ coding Cys, we see by inspection of the genetic code table that each of the 20 amino acids is coded by at least one trinucleotide which is a strong comma-free code of size 1. □

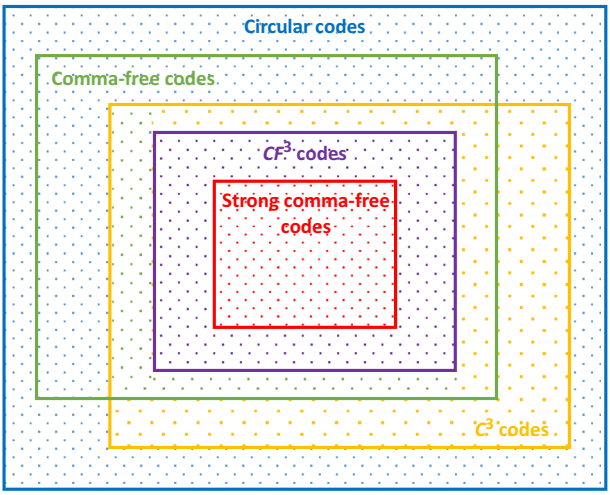

Fig. 7 Hierarchy of the five classes of strong comma-free codes, comma-free codes, CF^3-codes, C^3-codes and circular codes (Color figure online)

We now turn our attention to amino acids which are encoded by sets of trinucleotides that form a strong comma-free code.

Proposition 3.27 *There are 9 amino acids* $S = \{Asn, Asp, Gln, Gly, Lys, Met, Phe, Pro, Trp\}$ *among 20 such that for each amino acid from S, its synonymous trinucleotide set (excluding the necessary periodic trinucleotides* $\{AAA, CCC, GGG, TTT\}$*) is a strong comma-free code.*

Proof $\{ATG\}$ coding Met, $\{TGG\}$ coding Trp, $\{AAC, AAT\}$ coding Asn, $\{GAC, GAT\}$ coding Asp and $\{CAA, CAG\}$ coding Gln are strong comma-free codes. Moreover, the three trinucleotides $\{TAA, TAG, TGA\}$ coding the stop signal also build a strong comma-free code.

If the four periodic trinucleotides are excluded, four additional amino acids are coded by a strong comma-free code: $\{GGA, GGC, GGT\}$ coding Gly, $\{AAG\}$ coding Lys, $\{TTC\}$ coding Phe and $\{CCA, CCG, CCT\}$ coding Pro. □

Table 3 shows that the 8 trinucleotide strong comma-free codes of maximal size (called $MSCFC$ and listed in List 3.12) allow to code at least 4 amino acids (in case of X_5, X_7 and X_8) and at most 9 amino acids (in case of X_1 and X_3).

We finally investigate the primeval RNY code (Eigen and Schuster 1978), which is supposed to be one possible predecessor of the current standard genetic code. Our results show that the RNY code had very strong error-detecting properties and might have evolved from strong comma-free codes.

Proposition 3.28 *The primeval RNY code*

$$\{AAC, AAT, ACC, ACT, AGC, AGT, ATC, ATT,$$
$$GAC, GAT, GCC, GCT, GGC, GGT, GTC, GTT\}$$

is a self-complementary CF^3-code of size 16. Moreover, it is the union of two strong comma-free codes of size 8 which are complementary to each other.

Proof The self-complementary comma-free RNY code (see also Michel et al. 2008a; Fimmel and Strüngmann 2015) has 16 trinucleotides with purine $R = \{A, G\}$ as first nucleotide, $N \in \mathcal{B}$ as second nucleotide and pyrimidine $Y = \{C, T\}$ as third nucleotide. Thus, we have $\pi_1(X) \cap \pi_3(X) = \emptyset$. Then, according to Theorem 3.20, the RNY code is also CF^3. It is easy to check that the RNY code can be partitioned into the two complementary subcodes

$$\{AAC, ATT, GGT, AGT, GCT, ATC, GAC, GCC\}$$

and

$$\{AAT, ACC, ACT, AGC, GAT, GGC, GTC, GTT\}$$

which are strong comma-free. □

3.6 A Model of Evolution of Circular Codes

Circular codes allow to retrieve the reading (correct) frame with at most 13 nucleotides, comma-free codes with at most 3 nucleotides and strong comma-free codes with at most 2 nucleotides. The properties of these three classes of circular codes related to the reading frame retrieval, the cardinality range and the numbers of codes allow us to propose a model of evolution of circular codes (Fig. 8).

According to this model, evolution of circular codes is based on an increase in combinatorial flexibility, starting with strong comma-free codes, then comma-free codes up to circular codes which contain the greatest number of codes and, in addition, the longest nucleotide window of reading frame retrieval.

This combinatorial circular code evolution may also be associated with time evolution where strong comma-free codes and comma-free codes are more ancestral than circular codes.

Table 3 Amino acids encoded by the 8 trinucleotide strong comma-free codes of maximal size (called $MSCFC$ and listed in List 3.12)

$MSCFC$	Coded amino acids
X_1	$\{Asn, Asp, Cys, Gly, Ile, Phe, Ser, Tyr, Val\}$
X_2	$\{Gln, Leu, Lys, Met, Pro, Ser, Thr\}$
X_3	$\{Ala, Arg, Asn, Asp, Gly, His, Pro, Ser, Thr\}$
X_4	$\{Arg, Ile, Met, Ser, Thr\}$
X_5	$\{Arg, Gln, His, Leu\}$
X_6	$\{Ala, Arg, Gly, Leu, Pro, Ser, Val\}$
X_7	$\{Ala, Asp, Glu, Val\}$
X_8	$\{Cys, Ser, Trp, Tyr\}$

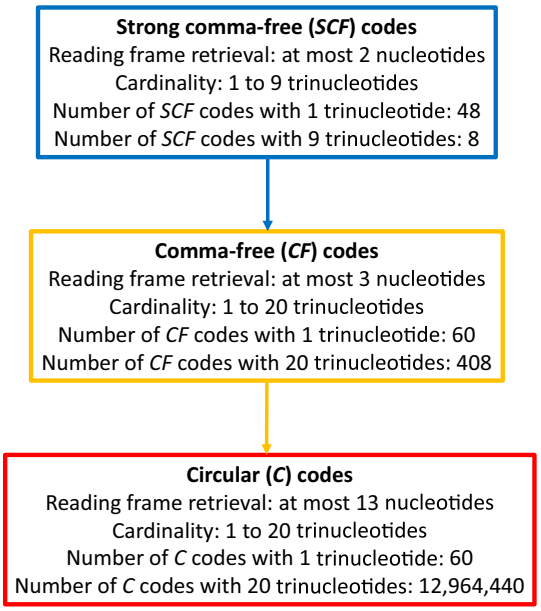

Fig. 8 A model of evolution of circular codes (Color figure online)

Recent results support this hypothesis (El Soufi and Michel 2017). Indeed, pure and long repeated trinucleotides are very common in genomes (noncoding regions) of eukaryotes (e.g., Canapa et al. 2002; Gemayel et al. 2010), e.g., $(AAC)^{2885}$, $(GAA)^{3512}$, $(GTC)^{605}$ and $(TAC)^{576}$ in *S. pennellii*, $(AAT)^{4317}$, $(ATT)^{6425}$ and $(GAT)^{692}$ in *C. sinensis*, $(ACC)^{121}$, $(GCC)^{185}$ and $(GGC)^{70}$ in *O. brachyantha*, $(ATC)^{715}$ in *C. sativa*, $(CAG)^{1185}$, $(CTC)^{355}$, $(CTG)^{241}$ and $(GAG)^{274}$ in *F. albicollis*, $(GAC)^{48}$ in *B. terrestris*, $(GGT)^{210}$ in *H. sapiens*, $(GTA)^{642}$ in *Z. mays*, $(GTT)^{1413}$ and $(TTC)^{1421}$ in *C. arietinum* (from Table 2 in El Soufi and Michel 2017). Note that the longest pure repeated trinucleotide $(ATT)^{6425}$ observed so far has a length of $l = 19275$ nucleotides. These pure and long repeated trinucleotides can be associated with a genetic information of low complexity.

For the pure and long repeated trinucleotides, the statistical analysis developed in El Soufi and Michel (2017) cannot decide between $(N_1N_2N_3)^n$, $(N_2N_3N_1)^n$ and $(N_3N_1N_2)^n$ with $N_1, N_2, N_3 \in \mathcal{B}$. For example, a sequence of the form $N_4N_1N_2N_3...N_1N_2N_3N_1N_2N_4$ with $N_1, N_2, N_3, N_4 \in \mathcal{B}$ and $N_1 \neq N_2 \neq N_3 \neq N_4$ can be assigned to the repeated trinucleotide $(N_1N_2N_3)^n$ or $(N_2N_3N_1)^n$ or $(N_3N_1N_2)^n$. In fact, the purpose of the developed analysis is the identification of pure and long repeated trinucleotides in eukaryotic genomes and their assignment to a class of permuted trinucleotides $[N_1N_2N_3] = \{N_1N_2N_3, N_2N_3N_1, N_3N_1N_2\}$.

From a scientific point of view which has been ignored so far, these pure repeated trinucleotides are in fact X circular code motifs, i.e., motifs from the circular code X identified in genes of minimal cardinality 1. They are generated by: (i) strong comma-free codes (of minimal cardinality 1), precisely the class of trinucleotides $[N_1N_2N_3]$ with $N_1 \neq N_2 \neq N_3$ as their associated graphs $\mathcal{G}(N_1N_2N_3)$ have only oriented paths of length 1, i.e., the eight trinucleotide classes $[ATC]$, $[CAG]$, $[CTG]$, $[GAC]$, $[GAT]$,

$[GTA]$, $[GTC]$ and $[TAC]$; and (ii) comma-free codes (of minimal cardinality 1), precisely the class of trinucleotides $[N_1N_2N_1]$ with $N_1 \neq N_2$ as their associated graphs $\mathcal{G}(N_1N_2N_1)$ have an oriented path of length 2, i.e., the 12 trinucleotide classes $[ACA]$, $[AGA]$, $[ATA]$, $[CAC]$, $[CGC]$, $[CTC]$, $[GAG]$, $[GCG]$, $[GTG]$, $[TAT]$, $[TCT]$ and $[TGT]$. Note that this class contains two strong comma-free codes $N_1N_1N_2$ and $N_2N_1N_1$, which cannot be distinguished statistically from the comma-free code $N_1N_2N_1$.

The pure and long repeated trinucleotides in genomes are very unstable. Mutation with rates up to 100,000 times higher than the genomic average mutation rate increases their evolutionary stability (e.g., Canapa et al. 2002; Gemayel et al. 2010). Mutation increases the cardinality (composition) and decreases the length of the pure and long repeated trinucleotides leading to mixed and short repeated trinucleotides which can be associated with a genetic information of middle complexity. Surprisingly, mixed and short repeated trinucleotides can contain circular code information. For example, X motifs of cardinality 5 are observed in genomes (El Soufi and Michel 2017):

- $CTG, GCC, GTT, GTC, (ACC)^{30}$ of $l = 102$ nucleotides in *H. sapiens*,
- $(GAA)^{11}, (GAC)^3, AAC, (GGT)^2, GAG$ of $l = 54$ nucleotides in *H. sapiens*,
- $(GGT)^7, GAC, AAT, GAT, (GAA)^2$ of $l = 36$ nucleotides in *H. sapiens*,
- $(GGC)^{16}, GTA, GCC, GTA, GAG, GGT, GAG$ of $l = 66$ nucleotides in *H. sapiens*,
- $ACC, GCC, (GTT)^9, ATT, (GTT)^2, ATT, (GTT)^2, ATC$ of $l = 54$ nucleotides in *S. cerevisiae*,
- $GTC, (ATC)^9, ACC, (ATC)^2, (ATT)^3, GGT$ of $l = 51$ nucleotides in *S. cerevisiae*,
- $CAG, GTC, (TTC)^{21}, (CTC)^{11}, CTG$ of $l = 105$ nucleotides in *M. musculus*,
- $TTC, CAG, GGC, (ATC)^5, (ATT)^{14}$ of $l = 66$ nucleotides in *M. musculus*,
- $GCC, GTC, ACC, GTC, ACC, GTC, GCC, ACC, (GTC)^8, CTC, ATC, CTC, GTC, GCC, (GTC)^2$ of $l = 69$ nucleotides in *Z. mays*,
- $GAC, GGC, AAC, GAG, GAC, GAG, (GAC)^5, GGC, (GAC)^4, GGT, GGC, GAC, GGT, GAC, GGC$ of $l = 66$ nucleotides in *Z. mays*, etc.

The X motifs of minimal or low cardinality, i.e., pure and mixed repeated trinucleotides of length $\geq$30 nucleotides, are very rare in genes (eukaryotes, bacteria, archaea, plasmids, viruses). The X motifs in genes have preferentially increased cardinalities and decreased lengths (El Soufi and Michel 2017) and can be associated with a genetic information of high complexity.

3.7 Reading Frame Retrieval in Current Genes

The reading frame retrieval of circular codes, which is at most 13 nucleotides, is true for sequences composed of at most 20 trinucleotides belonging to the circular codes, e.g., the circular code X identified in genes. A similar reasoning can be applied to the comma-free codes. For the strong comma-free codes, the case is even worse as the maximal cardinality is nine trinucleotides (Fig. 8).

However, current genes use 61 trinucleotides (without the three stop trinucleotides $\{TAA, TAG, TGA\}$), and not 20 trinucleotides. Hence, the frequency $P(X)$ of the

Fig. 9 A model of reading frame retrieval in genes using the X circular code motifs (Color figure online)

circular code X in genes is obviously less than 100%. Precisely, $P(X) = 46.6\%$ in genes of bacteria (7,851,762 genes, 2,481,566,882 trinucleotides), $P(X) = 41.5\%$ in genes of eukaryotes (1,662,579 genes, 824,825,761 trinucleotides), $P(X) = 47.5\%$ in genes of plasmids (237,486 genes, 68,244,356 trinucleotides) and $P(X) = 43.4\%$ in genes of viruses (184,344 genes, 45,688,798 trinucleotides) (Table 5b in Michel 2015). For an order of magnitude and by considering, for example, the bacterial genes, a trinucleotide of X has an average frequency equal to 2.33% while a trinucleotide of the 41 remaining trinucleotides has an average frequency equal to 1.30%.

The problem of reading frame retrieval in current genes has been initially addressed according to a statistical approach. The method developed was to use a nucleotide window, as with the circular code but longer, containing the circular code information, precisely the code X in the start frame f of the window, the permuted code $\mathcal{P}(X) = X_1$ in the shifted frame f_1 of the window and the second permuted code $\mathcal{P}^2(X) = X_2$ in the shifted frame f_2 of the window (details in Sect. 3.4. in Frey and Michel 2006). In the bacterial genes studied in 2006 (483926 genes, 523375 kb), the reading frame is retrieved with a frequency of 48.0% (probability of 1/3 in the random case) with a window of 5 nucleotides, 58.6% with a window of 13 nucleotides (100% with a circular code) and 81.0% with a window of 50 nucleotides (Fig. 3 in Frey and Michel 2006). However, it is difficult to propose a biological process for the ribosome for maintaining the reading frame in a gene by considering several tens of nucleotides.

Recent results have shown that the X circular code motifs occur preferentially in genes compared to genomes (noncoding regions of eukaryotes) with a factor of about 8 (Tables 4 and 5, Figs. 7 and 8 in El Soufi and Michel 2016). Thus, a model of frame retrieval can be proposed here where the ribosome pairs with the X motifs located at different positions in the genes (Fig. 9). Each pairing needs at most 13 nucleotides.

The ribosome contains the circular code information for pairing with the X motifs in genes. Indeed, the X motifs are 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; El Soufi and Michel 2014, 2015). However, the experimental biological mechanism by which the circular code maintains the translation frame is unknown.

4 Conclusion

Circular codes have two important proper subsets: the well-known comma-free codes and the strong comma-free codes identified here. These three classes of codes are important for investigating and identifying new properties in the genetic code. In particular, error detection and error correction during the translation process can be explained by circularity and the stronger versions of (strong) comma-freeness.

In the present article, the class of strong comma-free codes has been investigated for the first time. We have demonstrated that this class is a proper subset of the very important class of so-called CF^3-codes which allow an immediate frameshift error detection in each of the three frames. Additionally, some useful and easy-to-handle conditions for screening codes with regard to their strong comma-freeness were found. If a trinucleotide in a given code has repetitive nucleotides in the first and the last positions, then the code cannot be strong comma-free. If a given code has no repetitive nucleotides in at least two positions, i.e., in the first and third positions, or in the first and second positions, or in the second and third positions, the code is automatically CF^3, i.e., it has the property of the immediate error detection in the three frames. A complete description of strong comma-free codes of maximal size and self-complementary strong comma-free codes is also given.

Some new properties of the genetic code and its evolution are also identified. Each amino acid in the standard genetic code is coded by at least one strong comma-free code of size 1. As the reading window of such strong codes is at most two nucleotides, the amino acid decoding at the codon-anticodon process could be done with the first and second pairing positions, the second and third pairing positions or the third pairing position and a next position between tRNA and mRNA. In the last case, a tetranucleotide might be involved in amino acid decoding. There are 9 amino acids $S = \{Asn, Asp, Gln, Gly, Lys, Met, Phe, Pro, Trp\}$ among 20 such that for each amino acid from S, its synonymous trinucleotide set (excluding $\{AAA, CCC, GGG, TTT\}$) is a strong comma-free code. The primeval RNY code is a self-complementary CF^3-code of size 16, which is in addition to the union of two strong comma-free codes of size 8, complementary to each other. This result might indicate that during the evolutionary process the RNY code was created from a more ancient strong comma-free code by complementarity, in agreement with the proposed model of evolution of circular codes (Fig. 8).

References

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

Canapa A, Cerioni PN, Barucca M, Olmo E, Caputo V (2002) A centromeric satellite DNA may be involved in heterochromatin compactness in gobiid fishes. Chromosom Res 10:297–304

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

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

Crick F, Griffith JS, Orgel LE (1957) Codes without commas. Proceedings of the National Academy of Sciences, vol 43. U.S.A, pp 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

El Soufi K, Michel CJ (2017) Unitary circular code motifs in genomes of eukaryotes. Biosystems 153:45–62

Fimmel E, Giannerini S, Gonzalez D, Strüngmann L (2014) Circular codes, symmetries and transformations. J Math Biol. doi:10.1007/s00285-014-0806-7

Fimmel E, Strüngmann L (2015) On the hierarchy of trinucleotide n-circular codes and their corresponding amino acids. J Theor Biol 364:113–120
Fimmel E, Strüngmann L (2016) Maximal dinucleotide comma-free codes. J Theor Biol 389:206–213
Fimmel E, Michel CJ, Strüngmann L (2016) n-Nucleotide circular codes in graph theory. Philos Trans R Soc A 374:20150058
Frey G, Michel CJ (2006) Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. Comput Biol Chem 30:87–101
Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Ann Rev Genet 44:445–477
Golomb SW, Delbruck M, Welch LR (1958a) Construction and properties of comma-free codes. Biologiske Meddelelser, Kongelige Danske Videnskabernes Selskab 23:1–34
Golomb SW, Gordon B, Welch LR (1958b) Comma-free codes. Can J Math 10:202–209
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, Pirillo G (2011) Strong trinucleotide circular codes. Int J Comb 2011:659567. doi:10.1155/2011/659567
Michel CJ, Pirillo G, Pirillo MA (2008a) Varieties of comma free codes. Comput Math Appl 55:989–996
Michel CJ, Pirillo G, Pirillo MA (2008b) A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theor Comput Sci 401:17–26
Michel CJ, Pirillo G, Pirillo MA (2012) A classification of 20-trinucleotide circular codes. Inf Comput 212:55–63
Nirenberg MW, Matthaei JH (1961) The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proceedings of the National Academy of Sciences, vol 47. U.S.A., pp 1588–1602
Shepherd JCW (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proceedings of the National Academy of Sciences, vol 78. U.S.A., pp 1596–1600