Comma-free Codes Over Finite Alphabets

Elena Fimmel, Christian Michel, François Pirot, Jean-Sébastien Sereni, Lutz Strüngmann

Comma-free Codes Over Finite Alphabets

ELENA FIMMEL[1], CHRISTIAN J. MICHEL[2,*], FRANÇOIS PIROT[2,3], JEAN-SÉBASTIEN SERENI[2] AND LUTZ STRÜNGMANN[1]

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

[2] *Theoretical Bioinformatics, ICube,*
C.N.R.S., University of Strasbourg,
300 Boulevard Sébastien Brant
67400 Illkirch, France
[*] *Corresponding author*

[3] *LORIA (Orpailleur) and Dept. of Mathematics,*
C.N.R.S., University of Lorraine, INRIA and Radboud University
Vandœuvre-lès-Nancy, France and Nijmegen, Netherlands

Abstract. Comma-free codes have been widely studied in the last sixty years, from points of view as diverse as biology, information theory and combinatorics. We develop new methods to study comma-free codes achieving the maximum size, given the cardinality of the alphabet and the length of the words. Specifically, we are interested in counting the number of such codes. We provide (two different proofs for) a closed-formula. The approach introduced is further developed to tackle well-known sub-families of comma-free codes, such as self-complementary and (generalisations of) non-overlapping codes. We also study codes that are not contained in strictly larger ones. For instance, we determine the maximal size of self-complementary comma-free codes and the number of codes reaching the bound. We provide a characterisation of ℓ-letter non-overlapping codes (over an alphabet of cardinality n), which allows us to devise the number of such codes that are not contained in any strictly larger one. Our approach mixes combinatorial and graph-theoretical arguments.

1. Introduction

A code is *comma-free* if it does not require a distinct symbol to separate code words. Comma-free codes were constructed by Crick, Griffiths and Orgel [**11**] in 1957 as a class of trinucleotide codes to explain how the reading of a sequence of trinucleotides could code for amino acids. Combinatorial properties of comma-free codes were also considered, starting one year later with the seminal works of Golomb, Gordon and Welch [**20**] and of Golomb, Welch and Delbrück [**21**], who addressed the maximal size of a comma-free code with words of an arbitrarily fixed length over an alphabet of arbitrary cardinality. This spawned a number of purely combinatorial works on this topic [**10**, **12**, **25**, **27**, **28**], which led

E-mail address: e.fimmel@hs-mannheim.de, l.struengmann@hs-mannheim.de, c.michel@unistra.fr, sereni@kam.mff.cuni.cz, francois.pirot@loria.fr.

Date: November 22, 2019.

to a number of interesting results and challenging open questions. Biological interest for comma-free codes was increased by the discovery of a symmetry linked to codon frequencies by Arquès and Michel [**1**]. Furthermore, a certain sub-family of comma-free codes, coined "strongly regular codes" or "non overlapping codes", has also been the focus of several works [**7**, **23**, **24**], dating back to 1964, in particular for their interest in automata theory and for frame synchronisation applications [**2**, **5**, **6**, **9**, **22**]. It is no surprise that such natural properties of codes were useful and studied in a variety of contexts, under different names.

While digraphs have been used to study diletter comma-free codes, either implicitly through their adjacency matrix [**10**] or explicitly [**3**], appropriate digraphs for comma-free codes with longer words seem less natural and harder to find. This was recently done [**14**] not only for comma-free codes, but more generally for circular codes, of which comma-free codes form a subfamily. Graph theoretical tools have then been used to extend our understanding and knowledge of such codes [**13**–**16**]. We purse this line of study, providing answers to some open questions raised earlier and a unified approach for studying the structure of various subfamilies of comma-free codes.

One problem of particular interest is to compute the largest possible size $S(n, \ell)$ of a comma-free code with words of length ℓ over an alphabet Σ of cardinality n. Golomb, Welch and Delbrück [**21**] obtained a general upper bound on this size expressed using the Möbius function (see Proposition 3.4). Seven years later, this bound was shown to be attained whenever the length ℓ of the words is odd by Eastman [**12**], whose construction was subsequently simplified by Scholtz [**27**]. The situation when ℓ is even is less understood, and although it is known [**20**] that $S(n, 2\ell)$ is equivalent, as n goes to infinity, to $\alpha_{2\ell} n^{2\ell}$ with $\alpha_{2\ell} \in [1/(2e\ell), 1/(2\ell)]$, the leading coefficient is still to be determined.

We are interested in the number of comma-free codes of size $S(n, \ell)$, which we call *maximum*. Golomb, Gordon and Welch [**20**] proved that $S(n, 2) = \lfloor (n^2/3) \rfloor$ and provided a method to build all diletter codes of this size. Similarly, Golomb, Welch and Delbrück [**21**] proved that $S(n, 3) = \frac{1}{3} n(n^2 - 1)$ and provided, again, a method to build all maximum comma-free triletter codes. They produced a lot of insights into the structure of these codes. Yet, to the best of our knowledge their work does not provide a direct way to count all such codes. Interestingly, Cartwright, Cueto and Tobis [**8**] demonstrated an injection of the maximum independent set in the de Bruijn graph $B(n, 3)$ and the maximum comma-free triletter codes over an alphabet of cardinality n, yielding exponentially (in n) many such codes. However, they noted that the injection is not always a bijection and gave an example showing this when $n = 2$. As we shall see in Subsection 4.2, the injection is actually a bijection as soon as $n \geq 3$, which yields a closed formula to count the number of maximum comma-free triletter code. In addition, we provide a second proof of the formula, which is independent of de Bruijn graphs. The approach allows us to obtain closed formula counting other extremal codes studied in theoretical biology or data communication, specifically inclusion-wise maximal strongly regular triletter codes (Corollary 5.3) and inclusion-wise maximal strongly regular self-complementary triletter codes (Proposition 5.5). In particular, we extend earlier results of Blackburn [**7**] regarding strong comma-free ℓ-letter codes over an alphabet of cardinality n for small values of ℓ. These results are motivated by Blackburn's general conjecture [**7**, Conjecture 1]: he provided a construction of strong comma-free ℓ-letter codes and posited that for every integer $\ell \geq 2$, there exists an integer n_0 such that the construction yields a maximum strong comma-free code over an alphabet of cardinality n whenever $n \geq n_0$.

The recent extension of the graph theoretical approach [**14**] allows new relevant ways to partition all circular ℓ-letter codes over an alphabet of cardinality n. Specifically, one can group those codes according to the maximal length of a path in the digraph associated to the code: comma-free codes are precisely those for which the length is two, while strongly regular codes (aka. non overlapping

codes) are those for which the length is one. In Subsection 4.1, we generalise results obtained by Ball and Cummings [**3**] on $S(n,\ell)$ and the number of comma-free diletter codes of size $S(n,\ell)$ to diletter p-comma-free codes: we provide closed formulas for the maximal size and the number of p-comma-free diletter codes of maximal size for every p, which thus include comma-free codes but also strongly regular codes.
Finally, we include a computer-generated table containing the growth function of all circular triletter codes (of which comma-free codes are a sub-family) over the genetic alphabet $\mathcal{B}$ (of cardinality 4), presented in function of the number of arcs in a longest directed path of the associated graph.
Several examples are provided to illustrate the notions and constructions used, in an effort to increase readability.

2. Definitions and Notions

Let Σ be an arbitrary finite alphabet with $n \coloneqq |\Sigma|$. For an integer $\ell \geq 2$, an *ℓ-letter code* is a set $X \subseteq \Sigma^\ell$. We define Σ^* to be the collection of all finite words with letters in Σ, that is, $\cup_{\ell\geq 0}\Sigma^\ell$, and we define Σ^+ to be the collection of all finite and non-empty words with letters in Σ, that is, $\cup_{\ell\geq 1}\Sigma^\ell$.

Definition 2.1. Let $X \subseteq \Sigma^\ell$ be an *ℓ-letter code* and let $k \in \mathbf{N}$. We say that X is

- a *strong comma-free code* if no element of Σ^+ appears both as a prefix and a suffix in X: in other words, given any two non-necessarily distinct elements $c_1 = x_1 \dots x_\ell$ and $c_2 = y_1 \dots y_\ell$ of X, for every $k \in \{1, \dots, \ell-1\}$ we have
$$x_{\ell+1-k} \dots x_\ell \neq y_1 \dots y_k;$$
- a *comma-free code* if for any two elements $x_1 \dots x_\ell$ and $y_1 \dots y_\ell$ in X, we have
$$\forall i \in \{2, \dots, \ell\}, \quad x_i \dots x_\ell y_1 \dots y_{\ell-i} \notin X;$$
- a *k-circular ℓ-letter code* if for every $m \leq k$, every concatenation $c_1 \dots c_m$ of m elements of X, read on circle, admits exactly one partition (called a *circular decomposition*) into elements from X;
- a *circular ℓ-letter code* if it is a k-circular ℓ-letter code for all $k \in \mathbf{N}$;
- a *maximal (k-)circular ℓ-letter code* if it is not contained in a larger (k-)circular code;
- a *maximum (k-)circular (comma-free, strong comma-free) ℓ-letter code or, equivalently, code of maximal size* if $|Y| \leq |X|$ whenever Y is an ℓ-letter (k-)circular (comma-free, strong comma-free) code over Σ.

Remark 2.2. Strong comma-free codes have been previously defined in a number of contexts and under different names: they were first introduced as non-overlapping codes [**23**], then rediscovered and called cross-bifix-free codes [**2**], and recently redefined using graph theoretical models of the genetic code [**16**]. It is immediately clear that a strong comma-free code is a comma-free one. Indeed, if X is not comma-free, then it contains two elements c_1 and c_2 such that c_1c_2 contains an element $c \in X$ that starts after the first letter and ends before the last one, and hence there is a suffix of c_1 that is also a prefix of c_2, which means that X is not strong comma-free. We keep the name "strong comma-free" to emphasise this relation to comma-free codes, since we shall introduce p-comma-free codes, which correspond to strong comma-free codes when $p = 1$ and to comma-free codes when $p = 2$. We also point out that a comma-free code is automatically circular [**17**].

Two symmetric groups play an important role in the context of circular codes [**19**]. The first one acts on the elements of the alphabet Σ and is defined as

$$S_\Sigma \coloneqq \{\pi\colon \Sigma \to \Sigma \,:\, \pi \text{ is bijective}\}$$

endowed with the usual group operation given by the composition of functions. The group S_Σ has $n!$ elements and for every $\ell \in \mathbf{N}$, any bijective mapping $\pi\colon \Sigma \to \Sigma$ can be applied componentwise to $x \in \Sigma^\ell$ and thus yields a bijective map $\Sigma^\ell \to \Sigma^\ell$, which is also called π. A bijection π of S_Σ is an *involutory function* (or an *involution*) if $\pi \circ \pi(x) = x$ for every $x \in \Sigma$. A *fixed point* of a bijection $\pi \in S_\Sigma$ is an element $x \in \Sigma$ such that $\pi(x) = x$. If $|\Sigma|$ is even, then S_Σ contains involutory bijections without fixed points.
The second relevant symmetric group is S_ℓ, which acts on indices of each element from Σ^ℓ and permutes components in each ℓ-letter word from Σ^ℓ in a prescribed way. Formally we define

$$S_\ell \coloneqq \{\alpha\colon \{1,\dots,\ell\} \to \{1,\dots,\ell\} \,:\, \alpha \text{ is bijective}\}$$

For instance, if $\ell = 3$ then $(132) \in S_3$ is the permutation such that $1 \mapsto 3$, $2 \mapsto 1$ and $3 \mapsto 2$. Any element of S_ℓ induces a mapping on Σ^ℓ by permuting the order of the bases: for instance, if $\ell = 3$ then (132) transforms the triletter word $b_1b_2b_3$ into the triletter word $b_3b_1b_2$. The subgroup $\mathcal{A}_\ell$ of *cyclical permutations* of $(S_\ell, \circ)$ is interesting for us. It is formally defined by

$$\mathcal{A}_\ell \coloneqq \{\alpha_0 = (1)(2)\dots(\ell), \alpha_1 = (23\dots\ell 1), \alpha_2 = (3\dots\ell 12), \dots, \alpha_{\ell-1} = (\ell 12\dots\ell-1)\} \subseteq S_\ell.$$

Following several previous works, we point out that a circular code cannot contain two cyclically equivalent words, *i.e.* two words $w_1, w_2 \in \Sigma^\ell$ for which there exists $\alpha \in \mathcal{A}_\ell$ such that $\alpha(w_1) = w_2$. For example, if $x_1x_2x_3$ and $x_3x_1x_2$ are in the same code X then the word $x_1x_2x_3x_1x_2x_3$ admits two different cyclic decompositions into elements of X, namely

$$x_1x_2x_3|x_1x_2x_3 \quad \text{and} \quad x_1x_2|x_3x_1x_2|x_3.$$

In particular, if $\Sigma = \{x_1,\dots,x_n\}$, then for every $i \in \{1,\dots,n\}$ the trivial ℓ-letter word $x_ix_i\dots x_i$ cannot be a part of a circular code over Σ^ℓ. The classes produced by the cyclic equivalence relation are the *cyclic equivalent classes*. A cyclic equivalence class is *complete* if it has order ℓ, that is, if its representative is not a cyclic permutation of itself.
We now introduce the so-called *reversing permutation*, which inverts the order of letters in any ℓ-letter word over Σ, as

$$\overleftarrow{x_1x_2\dots x_{\ell-1}x_\ell} \coloneqq x_\ell x_{\ell-1}\dots x_2x_1 \in \Sigma^\ell.$$

Definition 2.3. Let Σ be an alphabet and π an involutory bijection of Σ. A code X over Σ is *π-self-complementary* if $\overleftarrow{\pi(x)} \in X$ whenever $x \in X$. A *fixed point* of π is an element $x \in \Sigma$ such that $\pi(x) = x$.

Due to the biological origins of our motivations, and also to avoid unessential technicalities, we consider only involutory bijections without fixed points.

Definition 2.4. The Möbius function $\mu\colon \mathbf{N} \to \{-1, 0, 1\}$ is defined as

$$\mu(n) \coloneqq \begin{cases} 0 & \text{if there exists } m \in \mathbf{N}, m > 1 \text{ such that } m^2|n, \\ (-1)^m & \text{if } n \text{ is the product of } m \text{ pairwise distinct prime numbers.} \end{cases}$$

Definition 2.5. For every positive integer ℓ and every integer $i \in \{1,\dots,\ell\}$, we define the *projection on the ith coordinate* $\pi_i\colon \Sigma^\ell \to \Sigma$ by $\pi_i(x_1 \cdots x_\ell) = x_i$. The projections on two coordinates $\pi_{ij}\colon \Sigma^\ell \to \Sigma^2$ are defined in a similar way whenever it makes sense.

3. ℓ-letter Circular Codes

As reported earlier, in the seminal combinatorial work on comma-free codes [**20**] it was observed that a comma-free code cannot contain more than one element in each cyclic equivalence class (and none in non-complete classes). This property actually holds for the larger class of 1-circular codes, which yields an upper bound on the size of such a code: the size cannot exceed the number of complete equivalence classes. This number is straightforward to compute using Möbius's inversion formula [**20**, Theorem 1]: for an ℓ-letter code over an alphabet of cardinality n, it is $\frac{1}{\ell}\sum_{d|\ell}\mu(\ell/d)n^d$. This upper bound is tight, and actually attained by a circular code. Indeed, for any total order on the alphabet, the code composed of the lexicographically smallest element of each complete cyclic equivalence class turns out to be circular. These facts seem to be folklore: we group them in the next theorem and provide a short proof that the bound is attained for completeness. To this end, we introduce the following notation. We define $S^1_{\max}(n,\ell)$ to be the size of a maximum ℓ-letter 1-circular code over an alphabet of cardinality n. Similarly, we define $S_{\max}(n,\ell)$ to be the size of a maximum circular ℓ-letter code over an alphabet of cardinality n. The case where $n = 1$ is trivial: there is no non-empty 1-circular ℓ-letter code if $\ell \geq 2$. Any ordering on a finite alphabet Σ naturally yields an ordering on Σ^ℓ for any positive integer ℓ, using the lexicographical order. We use the same symbol for all these orders.

THEOREM 3.1 (folklore). *Let Σ be an alphabet of cardinality $n \geq 2$ and let $<$ be a total order on Σ. The cardinality $S_{\max}(n,\ell)$ of a circular ℓ-letter code of maximal size over Σ is*

$$\frac{1}{\ell}\sum_{d|\ell}\mu\left(\frac{\ell}{d}\right)n^d,$$

and such a code X can be constructed in the following way: X contains the minimum element, according to $<$, of each complete cyclic equivalence class induced by $\mathcal{A}_\ell$.

PROOF. The upper bound having been explained above, we only show that the code $X \subset \Sigma^\ell$ that contains the minimum element of each cyclic equivalence class that is complete (*i.e.*, of size ℓ) is circular.
Suppose on the contrary that there exist two circular decompositions of some word $x_1x_2\dots x_{r\ell} \in \Sigma^{r\ell}$, the second one obtained by shifting k nucleotides, with $k \in \{1,\dots,\ell-1\}$, that is,

$$x_1\cdots x_\ell|x_{\ell+1}\cdots x_{2\ell}|\dots|x_{(r-1)\ell+1}\cdots x_{r\ell} \in X^+ \quad \text{and}$$
$$x_{k+1}\cdots x_{k+\ell}|x_{k+\ell+1}\cdots x_{k+2\ell}|\dots|x_{k+(r-1)\ell+1}\cdots x_{r\ell}x_1\cdots x_k \in X^+.$$

Note that, up to changing k by $\ell - k$, we may assume that $k \leq \ell - k$. For each $i \in \{1,\dots,r\}$, let us set

$$a_i \coloneqq x_{1+(i-1)\ell}\cdots x_{k+(i-1)\ell} \quad \text{and} \quad b_i \coloneqq x_{k+(i-1)\ell+1}\cdots x_{i\ell},$$

so for every $i \in \{1,\dots,r\}$,

$$|a_i| = k \quad \text{and} \quad |b_i| = \ell - k.$$

The two decompositions can thus be rewritten as

$$a_1b_1|\dots|a_rb_r \quad \text{and} \quad b_1a_2|\dots|b_ra_1,$$

and hence, setting $a_{r+1} \coloneqq a_1$ for convenience, one sees that $a_ib_i \in X$ and $b_ia_{i+1} \in X$ for every $i \in \{1,\dots,r\}$. Therefore, the definition of X implies that $b_i \notin \{a_i, a_{i+1}\}$, since every word in X belongs to a cyclic equivalence class of size ℓ. Furthermore, the definition of X also implies that for every $i \in \{1,\dots,r\}$,

(1) $a_ib_i < b_ia_i$ since a_ib_i and b_ia_i are in the same equivalence class, $a_ib_i \in X$ and $a_i \neq b_i$;

(2) $b_i a_{i+1} < a_{i+1} b_i$ since $b_i a_{i+1}$ and $a_{i+1} b_i$ are in the same equivalence class, $b_i a_{i+1} \in X$ and $a_{i+1} \neq b_i$; and
(3) $a_i \neq a_{i+1}$ since $\{b_i a_{i+1}, a_i b_i\} \subset X$ and X contains only one element in each cyclic equivalence class.

Because $k \leq \ell - k$, we know that $|a_i| \leq |b_i|$ for every $i \in \{1, \ldots, k\}$. Let b'_i be composed of the first k letters of b_i, that is, $b'_i \coloneqq x_{k+1+(i-1)\ell} \cdots x_{2k+(i-1)\ell}$. We know that $a_i \leq b'_i$ by (1), and that $b'_i \leq a_{i+1}$ by (2). It follows that $a_i \leq a_{i+1}$, and hence $a_i < a_{i+1}$ by (3). Since this is valid for every $i \in \{1, \ldots, r+1\}$, we conclude that $a_1 < \cdots < a_r < a_{r+1} = a_1$, a contradiction. □

Example 3.2.

(1) Let us endow $\Sigma = \{0, 1\}$ with the order $0 < 1$. For $\ell = 3$, we have the following complete cyclic equivalence classes:

$$\{100, 010, 001\}, \{110, 011, 101\}.$$

Theorem 3.1 implies that $\{001, 011\}$ is a binary circular 3-letter code of maximal size.

For $\ell = 4$ we have the following complete cyclic equivalence classes:

$$\{1000, 0100, 0010, 0001\}, \{1100, 0110, 0011, 1001\}, \{1110, 0111, 1011, 1101\}.$$

Theorem 3.1 implies that $\{0001, 0011, 0111\}$ is a binary circular 4-letter code of maximal size.

(2) The *genetic alphabet* is $\mathcal{B} = \{A, C, G, T\}$ where A stands for *Adenine*, C for *Cytosine*, G for *Guanine* and T for *Thymine*. Let us endow $\mathcal{B}$ with the order $A < C < G < T$. For $\ell = 2$ we have the following complete cyclic equivalence classes:

$$\{AC, CA\}, \{AG, GA\}, \{AT, TA\}, \{CG, GC\}, \{CT, TC\}, \{GT, TG\}.$$

Theorem 3.1 implies that $\{AC, AG, AT, CG, CT, GT\}$ is a circular dinucleotide code of maximal size.

For $\ell = 3$ we have the following complete cyclic equivalence classes:

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

Theorem 3.1 implies that

$$\{AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGG, AGT, ATC,$$
$$ATG, ATT, CCG, CCT, CGG, CGT, CTG, CTT, GGT, GTT\} \tag{3.1}$$

is a circular trinucleotide code of maximal size.

Let us spell out some important special cases of 1-circular codes.

Special Cases 3.3. *Let n be an integer greater than* 1.

(1) We first consider the cases where ℓ is a prime number p. We have

$$S^1_{\max}(n, p) = \frac{1}{p}\left(n^p - n\right).$$

For $\Sigma = \{0, 1\}$ we obtain $S^1_{\max}(2, p) = \frac{1}{p}\left(2^p - 2\right)$, and hence the following values.

p	2	3	5	7
$S^1_{\max}(2,p)$	1	2	6	18

For $\mathcal{B} = \{A, C, G, T\}$ we obtain $S^1_{\max}(4,p) = \frac{1}{p}(4^p - 4)$. It means that $S^1_{\max}(4,2) = 6$ and $S^1_{\max}(4,3) = 20$.

(2) *More generally, suppose that ℓ is a power of a prime number, that is, $\ell = p^a$ where p is a prime number and a is a positive integer. Then*

$$S^1_{\max}(n, p^a) = \frac{1}{p^a}\left(n^{p^a} - n^{p^{a-1}}\right).$$

In particular, we obtain $S^1_{\max}(2,4) = 3$ and $S^1_{\max}(4,4) = 60$.

(3) *We now consider the case where ℓ is the product of two distinct prime numbers p and q. Then*

$$S^1_{\max}(n, p\cdot q) = \frac{1}{p\cdot q}\left(n^{p\cdot q} - n^p - n^q + n\right).$$

In particular, $S^1_{\max}(2, p\cdot q) = \frac{1}{p\cdot q}(2^{p\cdot q} - 2^p - 2^q + 2)$, and hence $S^1_{\max}(2,6) = 9$.

Theorem 3.1 readily yields an upper bound on the *number* of different circular ℓ-letter codes of a given size.

PROPOSITION 3.4. *Let Σ be an alphabet of cardinality n. For every $k \in \{1, \ldots, S_{\max}(n,\ell)\}$, the number of circular ℓ-letter codes of size k over Σ cannot exceed*

$$N_{\max}(n,\ell,k) \coloneqq \binom{S^1_{max}(n,\ell)}{k}\ell^k.$$

PROOF. There are $\binom{S^1_{max}(n,\ell)}{k}$ possibilities to choose k classes from $S^1_{max}(n,\ell)$ different cyclic equivalence classes. There are ℓ^k ways to choose an arbitrary element in each of the k chosen classes. □

The computer-calculated numbers in Table 1 are the number of circular ℓ-letter codes of size k over a two-letter alphabet, for $\ell \in \{2, \ldots, 6\}$ and all the possible corresponding values of k. As expected, none of these numbers exceeds the upper bound provided by Proposition 3.4. Nevertheless, and not surprisingly, most of them are strictly smaller.

We now point out some facts about Theorem 3.1

REMARK 3.5. Theorem 3.1 means that among all maximum 1-circular ℓ-letter code over a given alphabet, at least one of them is circular. We note that not all of them are, as is seen by considering for instance 3-letter words over the genetic alphabet: since an ℓ-letter code is 1-circular as soon as it contains only words that belong to a complete cyclic equivalence class no two of which being in the same class, there exists a 1-circular 3-letter code X containing the words AAG, TAA, GGT and GTT. Therefore, the word $TAAGGT$ admits two circular decompositions into words in X, namely $TAA|GGT$ and $T|AAG|GT$. (As one can see, there is no order on $\{A, G, T\}$ such that each of the first three words above is the smallest element in its own cyclic equivalence class. Indeed, if AAG is the smallest in its cyclic equivalence class, then $A < G$ (because GAA belongs to this class) and similarly if GGT is minimum in its class then $G < T$ and hence $A < T$. This implies that TAA is not the smallest element in its class.)

Furthermore, Theorem 3.1 does not describe all circular ℓ-letter codes of maximal size. That is, there exist maximum circular ℓ-letter codes for which no order on the alphabet can be found such that every word in the code is the smallest element of its own cyclic equivalence class. For instance, endowing

k \ ℓ	2	3	4	5	6
1	2 (2)	6 (6)	12 (12)	30 (30)	54 (54)
2		8 (9)	48 (48)	374 (375)	1290 (1296)
3			60 (64)	2458 (2500)	17788 (18144)
4				8712 (9375)	154252 (163296)
5				14952 (18750)	857534 (979776)
6				9204 (15625)	2990084 (3919104)
7					6156160 (10077696)
8					6648638 (15116544)
9					2832746 (10077696)

TABLE 1. Growths of circular ℓ-letter codes over an alphabet of cardinality 2 in function of the code size k. In brackets are written the corresponding upper bounds $N_{\max}(2, \ell, k)$ given by Proposition 3.4.

the genetic alphabet with the order $A < C < G < T$ and considering 3-letter words, we have seen that the 20 trinucleotides given in (3.1) form a maximum circular code. However, replacing AAC by ACA also yields a circular code, and yet there is no order on the genetic alphabet such that ACA is the smallest element in its cyclic equivalence class.

4. Characterisation of Di- and Triletter Comma-Free Codes

We define $N(n, \ell)$ to be the number of comma-free ℓ-letter codes of maximal size over an alphabet with n letters, and $S(n, \ell)$ to be their size. We study in this section the case where $\ell \in \{2, 3\}$, starting with $\ell = 2$.

4.1. Diletter Codes. Notice that $S(2, 2) = 1$ and $N(2, 2) = 2$. The case where $n \geq 3$ is more interesting. Let Σ be a finite alphabet and X an ℓ-letter code over Σ. One can associate to any code X a digraph G_X as follows [**14**]. The vertex set of G_X is $\cup_{i=1}^{\ell-1}\Sigma^i$, and there is an arc from w to w' if and only if the concatenation ww' belongs to X. It was proved [**14**, Theorem 2.6] that X is circular if and only if G_X is acyclic. Moreover, if p is the length of a longest (directed) path in G_X, then X is comma-free if and only if $p \leq 2$. Further, X is *strong comma-free* (also known as strongly regular or non-overlapping) if and only if $p = 1$. It thus seems natural to partition all the circular ℓ-letter codes over a given alphabet Σ according to the length of the longest directed path in their associated digraph. A circular code X is *p-comma-free* if no directed path in G_X has length more than p.
For every positive integer p, let $S_p(n, 2)$ be the size of a maximum p-comma-free circular diletter code over an alphabet of cardinality n. Furthermore, let $N_p(n, 2)$ be the number of different such maximum codes. (In particular, $S_2(n, 2) = S(n, 2)$ and $N_2(n, 2) = N(n, 2)$.) Our next result[1] provides formulæ for the size and the number of maximum p-comma-free diletter codes for every integer p. As is usual, $\binom{a}{b}$ stands for the number of ways of choosing b elements of a set of cardinality a; in particular, $\binom{a}{0} = 1$ and $\binom{a}{b} = 0$ if $b > a$.

[1]This seems to generalise an earlier result of Ball and Cummings [**3**], using a similar approach although we could not access the article.

THEOREM 4.1. *Let n be an integer greater than 2 and $p \in \{1, \dots, n-1\}$. We set $m \coloneqq \lfloor \frac{n}{p+1} \rfloor$ and $r \coloneqq n - (p+1)m \in \{0, \dots, p\}$. Then*

$$S_p(n,2) = \frac{1}{2}\left(1 - \frac{1}{p+1}\right)(n^2 - r^2) + \binom{r}{2} = \frac{pn^2 + r(r-p-1)}{2(p+1)}$$

and

$$N_p(n,2) = \binom{p+1}{r} \frac{n!}{m!^{p+1}(m+1)^r}.$$

REMARK 4.2. Applying Theorem 4.1 with $p = 1$ allows us to recover earlier results [**16**, Proposition 3.3(2) and Theorem 3.6(2)], using different arguments. Indeed, in this case if n is even then $m = n/2$ and $r = 0$, while if n is odd then $m = (n-1)/2$ and $r = 1$. Therefore,

$$N_1(n,2) = \begin{cases} \frac{1}{2}\left(1 - \frac{1}{2}\right) \times n^2 = \frac{1}{4} \cdot n^2 & \text{if } n \text{ is even,} \\ \frac{1}{2}\left(1 - \frac{1}{2}\right) \times (n^2 - 1) = \frac{1}{4} \cdot (n-1)(n+1) & \text{if } n \text{ is odd,} \end{cases}$$

and moreover

$$S_1(n,2) = \begin{cases} \binom{2}{0} \frac{n!}{\left(\frac{n}{2}!\right)^2 \times 1^0} = \binom{n}{n/2} & \text{if } n \text{ is even,} \\ \binom{2}{1} \frac{n!}{\left(\frac{n-1}{2}!\right)^2 \times \left(\frac{n+1}{2}\right)^1} = 2\binom{n}{\frac{n+1}{2}} & \text{if } n \text{ is odd.} \end{cases}$$

We use two classical results from graph theory to establish Theorem 4.1, one coming from the study of tournaments and the other from extremal graph theory. To state and use them, we need to introduce some terminology. The *order* of a directed path in a digraph is the number of vertices on the path, that is, its length plus one. A graph is *complete* if every two distinct vertices are adjacent: we let K_n be the complete graph with n vertices. A *tournament* is an orientation of a complete graph, that is, a choice of a direction for each edge of a complete graph. It turns out that every digraph that contains a tournament also contains a directed path going once through each vertex of the tournament. To be more precise, given any digraph D, a *Hamiltonian path* of D is a directed path going once through each vertex of D. A straightforward induction on the number of vertices shows that every tournament admits a Hamiltonian path.
The following immediate corollary is what will be useful to us.

OBSERVATION 4.3. *If the length of a longest directed path in a digraph D is at most p, then (the underlying undirected graph of) D does not contain K_{p+1} as a subgraph.*

The second result we need provides an upper bound on the number of edges in a graph that does not contain a complete subgraph of a certain order. If n and k are two integers, the *Turán graph* $T_k(n)$ is obtained by partitioning n (unlabelled) vertices into k parts with sizes in $\{\lfloor n/k \rfloor, \lceil n/k \rceil\}$ (there is a unique way to do so) and then placing an edge between two vertices if and only if they belong to different parts. The Turán graphs $T_3(11)$ and $T_4(16)$ are depicted in Figure 1. We define $t_k(n)$ to be the number of edges of the Turán graph $T_k(n)$. One can check that

$$t_k(n) = \frac{1}{2}\left(1 - \frac{1}{k}\right)(n^2 - r^2) + \binom{r}{2},$$

where r is the remainder of n divided by k.
If an n-vertex graph G does not contain a complete subgraph on $k+1$ vertices, then how many edges at most can G have? Clearly, if $k+1 > n$ then G can be itself complete and it is the only way to maximise the number of edges in G; hence the question is interesting when $k+1 \leq n$. In 1941, Turán established an upper bound on the number of edges of an n-vertex graph without a complete subgraph on $k+1$ vertices and characterised the graphs attaining this upper bound.

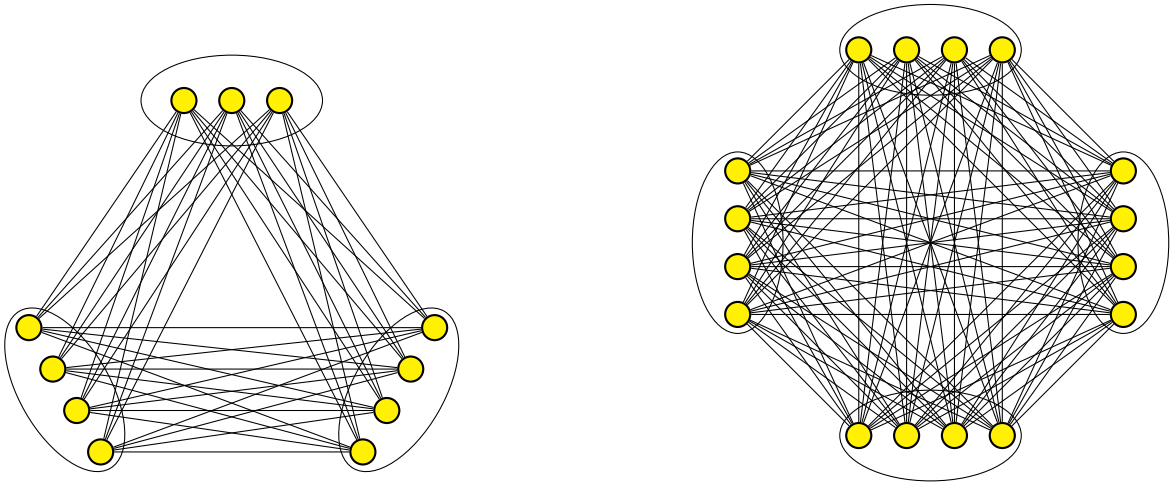

FIGURE 1. The Turán graph $T_3(11)$ on the left and the Turán graph $T_4(16)$ on the right.

THEOREM 4.4 (Turán [**29**]). *Let n be a positive integer and let $k \in \{1, \dots, n-1\}$. Let G be a graph with n vertices that does not contain a complete subgraph on $k+1$ vertices. Then the number of edges of G is at most $t_k(n)$, with equality if and only if $G = T_k(n)$.*

We are now in a position to establish Theorem 4.1.

PROOF OF THEOREM 4.1. Let Σ be an alphabet of cardinality n. We start by establishing that $S_p(n,2) = t_p(n)$. Let X be a p-comma-free diletter code over Σ. Because X is a diletter code, the size of X is precisely the number of arcs in G_X. Since X is p-comma-free, we know that the length of every directed path in G_X is at most p. It follows from Observation 4.3 that (the underlying undirected graph of) G_X does not contain the complete graph on $p+1$ vertices as a subgraph, and hence Theorem 4.4 implies that the number of arcs in G_X or, equivalently, the size of the code X is at most $t_p(n)$, and hence $S_p(n,2) \leq t_p(n)$.

To establish the equality, it remains to exhibit a p-comma-free diletter code of size $t_p(n)$, which can be done as follows. We consider the Turán graph $T_{p+1}(n)$, with vertex partition $(X_1, \dots, X_{p+1})$. We orient all edges between X_i and X_j from the vertices in X_i to those in X_j whenever $1 \leq i < j \leq p+1$. Now, arbitrarily identifying the vertices of G with the elements of Σ yields a p-comma-free diletter code of size $t_p(n)$, and hence $t_p(n) = S_2(n,2)$.

We proceed to calculate the number of maximum p-comma-free diletter codes over Σ. It follows from Theorem 4.4 and our previous considerations that X is a p-comma-free diletter code over Σ if and only if the underlying undirected graph H_X of G_X is $T_{p+1}(n)$. Starting from this graph, every such code is thus created by two choices: first an ordering of the parts of the partition, and next a bijection between the vertices and the alphabet. Note that some bijections yield the same code: we just want to assign the letters to the parts of the partition. Therefore, letting $(X_1, \dots, X_{p+1})$ be the partition of the vertices of $T_{p+1}(n)$, it is more convenient to express this second choice as the choice of a function $f\colon \Sigma \to (X_1, \dots, X_{p+1})$ such that $\left|f^{-1}(X_i)\right| = |X_i|$ for every $i \in \{1, \dots, p+1\}$.

To compute the number of such functions f, recall that $n = (p+1)m + r$ where $m = \lfloor n/(p+1) \rfloor$ and $r \in \{0, \dots, p\}$. Set $M \coloneqq \lceil n/(p+1) \rceil$, so $M = m$ if $r = 0$ and $M = m+1$ otherwise. Using these

n	3	4	5	6	7	8	9	10	11	12	13	14	15
$S(n,2)$	3	5	8	12	16	21	27	33	40	48	56	65	75
$N(n,2)$	6	36	90	90	630	1680	1680	12600	34650	34650	270270	756756	756756

TABLE 2. The sizes and numbers of maximum comma-free diletter codes over an alphabet of cardinality $n \in \{3, \ldots, 15\}$, as obtained using computers [**15**] and confirmed by the general formulæ of Theorem 4.1.

notations, the number of functions f as above is

$$\prod_{i=0}^{r-1} \binom{n-iM}{M} \prod_{i=0}^{p-r} \binom{n-rM-im}{m},$$

noting that if $r = 0$ then the first product is empty — and hence equal to 1.
The number of orderings of the parts is then $\binom{p+1}{r}$, since we only need to choose the r places of the parts of size M. Note that this value is 1 if $r = 0$. We therefore obtain the following formula,

$$\binom{p+1}{r} \cdot \prod_{i=0}^{r-1} \binom{n-iM}{M} \prod_{i=0}^{p-r} \binom{n-rM-im}{m},$$

which is equal to

$$\binom{p+1}{r} \frac{n!}{m!^{p+1}(m+1)^r}.$$

This concludes the proof. □

EXAMPLE 4.5. The genetic alphabet corresponds to the case where $n = 4$, which was previously studied by other means [**14**, **18**]; we have

$$S(4,2) = S_2(4,2) = \frac{1}{3}(4^2 - 1^2) + \binom{1}{2} = 5 \quad \text{and} \quad N(4,2) = N_2(4,2) = \binom{3}{1} \frac{4!}{(1!)^3 2^1} = 36.$$

REMARK 4.6. Using computers, all maximum comma-free diletter codes on alphabets of cardinalities $n \in \{2, \ldots, 15\}$ have been generated [**15**]: Theorem 4.11 confirms that the computer programs used were correct, and provides the sought value for every integer $n \geq 3$. (See Table 2.)

4.2. Triletter Codes. We now turn our attention to comma-free triletter codes of maximal size. Let Σ be an alphabet of cardinality n. The case where $n = 2$ is straightforward. One readily sees that the maximum size of a comma-free triletter code over Σ, that is, $S(2,3)$, is 2 and there are exactly 8 of them: writing $\Sigma = \{0, 1\}$, these eight codes are

$$\{001, 011\}, \{001, 101\}, \{001, 110\}, \{010, 011\}, \{010, 110\}, \{011, 100\}, \{100, 101\}, \{100, 110\}.$$

Indeed, neither 000 nor 111 can be part of a comma-free code and a comma-free code contains at most one element in each complete equivalence class (and none in non-complete equivalence classes). It then only remains to notice that, among the nine possible choices, exactly one is not comma-free: $\{010, 101\}$.
As reported earlier, Golomb, Gordon and Welch [**20**] obtained an upper bound on the size of a maximum comma-free ℓ-letter code over an alphabet of cardinality n, which was shown to be attained for each odd value of ℓ by Eastman [**12**]. We focus on triletter codes: $S(n,3) = \frac{n(n^2-1)}{3}$. Let Σ be an alphabet of cardinality $n \geq 3$. While Golomb, Welch and Delbrück [**21**] described a method

to construct all maximum comma-free triletter codes over Σ, no formula to count them is provided. They showed (cf. note after their Theorem 7) that the product of two groups (that of permutation of the alphabet and that of so-called "reversals") generate all maximum comma-free triletter codes, starting from a number of "basic" codes. However, building the basic codes requires enumerating all integer partitions of $\frac{n(n^2-1)}{3}$ that satisfy a certain property: these do not seem obvious to count. Furthermore, given the number of "basic codes", the cardinality of the group of "reversals" does not seem obvious to find either.

The *de Bruijn graph* $B(n,3)$ has vertex set Σ^3 and an arc from $N_1N_2N_3$ to $N_4N_5N_6$ if and only if $N_2 = N_4$ and $N_3 = N_5$. (It thus contains $|\Sigma|$ loops.) In 2011, Cartwright, Cueto and Tobis [**8**] counted the number of maximum independent sets in $B(n,3)$ by finding the generating function: solving the corresponding recurrence shows this number to be

$$\left[\frac{(1+\sqrt{2})^n}{2}\right] n!,$$

where $[x]$, for an irrational number x, is the integer closest to x. They moreover observed [**8**, Theorem 5.1] that maximum independent sets of $B(n,3)$ inject into the collection of maximum comma-free triletter codes, thereby obtaining exponentially (in n) many different such codes. They noted that, when $n = 2$, the injection is not surjective. Indeed, there exist precisely two maximum comma-free triletter codes over $\{0,1\}$ that do not correspond to independent sets in $B(2,3)$, namely $\{001, 011\}$ and $\{110, 100\}$.

However, using some of the properties first obtained by Golomb, Welch and Delbrück [**21**], one realises that $n = 2$ is the only exceptional case: as soon as $n \geq 3$, maximum comma-free triletter codes over Σ are in bijection with maximum independent sets in the de Bruijn graph $B(n,3)$. We thus obtain the following statement, of which a second proof is given later in this subsection.

THEOREM 4.7. *If n is an integer greater than 2, then*

$$N(n,3) = \left[\frac{(1+\sqrt{2})^n}{2}\right] n!.$$

FIRST PROOF. Let $n \geq 3$ and suppose that X is a maximum comma-free code that is not a maximum independent set in $B(n,3)$. Consequently, there exist two words w and w' in X of the form $w = N_1N_2N_3$ and $w' = N_2N_3N_4$ where $\{N_1, \ldots, N_4\} \subseteq \Sigma$. As proved by Golomb, Welch and Delbrück [**21**, Note after Theorem 3], as X is maximum every letter in Σ, except possibly one, occurs both as the first letter of a word in X and as the last letter of a word in X. It follows that we may assume, without loss of generality in the sequel, that X contains a word $w'' = N_5N_6N_1$ where $\{N_5, N_6\} \subset \Sigma$. Therefore, the concatenation $w''w'$ contradicts that X is comma-free. The formula follows. □

EXAMPLE 4.8. For the genetic alphabet $\mathcal{B} = \{A, C, G, T\}$, we find back the well-known [**20**, **21**, **26**] numbers $S(4,3) = \frac{4(4^2-1)}{3} = 20$ for the size and $N(4,3) = 4! \left[\frac{(1+\sqrt{2})^4}{2}\right] = 24 \cdot 17 = 408$ for the number of comma-free triletter codes of maximal size over the genetic alphabet.

EXAMPLE 4.9. When the alphabet has size 3, one can check that the formulæ indeed give all the 42 different circular comma-free triletter codes of maximal size, which is 8. Letting Σ be $\{0,1,2\}$, these 42

codes are listed below.

$$\{010,020,021,022,110,120,121,122\},\{001,002,101,102,112,201,202,212\},$$
$$\{001,020,021,022,101,120,121,122\},\{001,020,021,101,120,121,220,221\},$$
$$\{010,012,020,110,112,210,212,220\},\{100,101,102,112,200,201,202,212\},$$
$$\{002,010,012,110,112,202,210,212\},\{001,002,101,102,121,201,202,221\},$$
$$\{001,002,110,112,201,202,210,212\},\{011,020,021,100,120,121,220,221\},$$
$$\{010,012,022,110,112,200,210,212\},\{010,011,020,021,022,120,121,122\},$$
$$\{010,011,012,022,200,210,211,212\},\{010,011,020,021,120,121,220,221\},$$
$$\{100,101,102,112,200,201,202,221\},\{001,020,021,022,110,120,121,122\},$$
$$\{001,002,101,102,122,201,202,211\},\{010,012,110,112,200,202,210,212\},$$
$$\{020,021,022,100,101,120,121,122\},\{011,020,021,022,100,120,121,122\},$$
$$\{010,020,021,110,120,121,220,221\},\{020,021,100,101,120,121,220,221\},$$
$$\{010,011,012,020,022,210,211,212\},\{010,012,020,021,110,112,220,221\},$$
$$\{010,012,020,022,110,112,210,212\},\{100,101,102,200,201,202,211,212\},$$
$$\{010,011,012,200,202,210,211,212\},\{011,012,100,102,200,202,211,212\},$$
$$\{100,101,102,122,200,201,202,211\},\{002,010,011,012,202,210,211,212\},$$
$$\{021,022,100,101,121,122,200,201\},\{100,101,102,121,122,200,201,202\},$$
$$\{002,010,012,110,112,210,212,220\},\{001,002,101,102,112,201,202,221\},$$
$$\{010,011,020,022,120,122,210,211\},\{001,002,101,102,120,121,220,221\},$$
$$\{001,002,101,102,201,202,211,212\},\{002,010,011,012,210,211,212,220\},$$
$$\{001,002,101,102,121,122,201,202\},\{001,020,021,110,120,121,220,221\},$$
$$\{010,011,012,020,210,211,212,220\},\{100,101,102,121,200,201,202,221\}.$$

We now provide a proof of Theorem 4.7 that is independent of the result of Cartwright, Cueto and Tobis [**8**]. It is based on a combinatorial lemma, which allows us to fully characterise the "basic" types of maximum comma-free codes identified by Golomb, Welch and Delbrück [**21**]. This further allows us to find a recursive formula to compute the cardinality of the group of "reversals", thereby recovering Theorem 4.7. Given an ordered family of sets $\mathcal{F} = (L_1, R_1, \ldots, L_m, R_m)$, we define $S(\mathcal{F})$ to be $\sum_{i=1}^{m} |L_i|\,|R_i|$.

LEMMA 4.10. *Let m be a positive integer. Let $\mathcal{F} = (L_1, R_1, \ldots, L_m, R_m)$ be a family of subsets of a set $\mathcal{A} = \{a_1, \ldots, a_m\}$ of cardinality m such that*

(1) $a_j \in L_i$ if and only if $a_i \notin R_j$ for every $(i,j) \in \{1,\ldots,m\}^2$.

Then $S(\mathcal{F}) \leq \frac{m(m^2-1)}{3}$. Moreover, if $S(\mathcal{F}) = \frac{m(m^2-1)}{3}$, then there exists a sequence of the operations (A) and (B) below such that, up to permuting the indices in $\{1,\ldots,m\}$, one has for each $i \in \{1,\ldots,m\}$

$$L_i = \{a_1, \ldots, a_i\} \quad \text{and} \quad R_i = \{a_1, \ldots, a_{i-1}\}.$$

Operation (A) consists in swapping a_i between L_i and R_i for some $i \in \{1,\ldots,m\}$, providing it does not decrease $S(\mathcal{F})$. Operation (B) is as follows: if $1 \leq j < i \leq m$ and $|L_i| = |R_j|$ and $a_j \notin R_i$ then add a_j to R_i and remove a_i from L_j; or similarly, if $|R_i| = |L_j|$ and $a_j \notin L_i$ then add a_j to L_i and remove a_i from R_j.

PROOF. Notice that (1) implies that $a_i \in L_i \cup R_i$ for each $i \in \{1,\ldots,m\}$.

We proceed by induction on the positive integer m, the statement being true if $m = 1$ as then $L_1 = \varnothing$ or $R_1 = \varnothing$. For the induction step, assume that the statement is true for $m - 1 \geq 1$ and let us establish it for m. Let $\mathcal{F} = (L_1, R_1, \ldots, L_m, R_m)$ be a family of subsets that maximises $S(\mathcal{F})$ under the condition (1) stated above. For each $i \in \{1, \ldots, m\}$, we let $\{M_i, S_i\} = \{L_i, R_i\}$ such that $|M_i| \geq |S_i|$. Since we may apply a permutation of $\{1, \ldots, m\}$ to the indices of the elements in $\mathcal{F}$, we may assume that $|M_m| \geq |M_i|$ for every $i \in \{1, \ldots, m\}$. By symmetry of the roles played by L_m and R_m *via* operation (A), we may assume further that $M_m = L_m$ and hence $S_m = R_m$. Let us show how to ensure that $R_m = \mathcal{A}$. First, if there exists $j \in \{1, \ldots, m-1\}$ such that $a_j \notin R_m$, then we set

$$L_i' \coloneqq \begin{cases} L_i & \text{if } i \in \{1, \ldots, m\} \setminus \{j\} \\ L_j \setminus \{a_m\} & \text{if } i = j \end{cases} \quad \text{and} \quad R_i' \coloneqq \begin{cases} R_i & \text{if } i \in \{1, \ldots, m-1\} \\ R_m \cup \{a_j\} & \text{if } i = m, \end{cases}$$

the family $\mathcal{F}'$ just defined satisfies (1). In addition, $S(\mathcal{F}') = S(\mathcal{F}) + |L_m| - |R_j|$. The definition of $\mathcal{F}$ thus implies that $|L_m| = |R_j|$. It follows that (B) can be applied to R_m and L_j, yielding that now $a_j \in R_m$. We conclude that we can ensure that $\{a_1, \ldots, a_{m-1}\} \subseteq R_m$. If $L_m \neq \mathcal{A}$ then $\max\{|L_i|, |R_i|\} \leq m - 1$ for every $i \in \{1, \ldots, m-1\}$ and therefore we now have $|R_m| \geq \max\{|L_i|, |R_i|\}$ for each $i \in \{1, \ldots, m-1\}$. This means that we can proceed similarly with L_m as we just did with R_m and obtain $\mathcal{A} \setminus \{a_m\} \subseteq L_m$. Consequently, up to applying (A), we have $L_m = \mathcal{A}$ and $R_m = \mathcal{A} \setminus \{a_m\}$.

As a result, the family $\mathcal{F}_1 \coloneqq (L_1, R_1, \ldots, L_{m-1}, R_{m-1})$ consists of subsets of $\mathcal{A} \setminus \{a_m\}$, and satisfies the condition (1). In addition, $\frac{1}{3}m(m^2-1) = S(\mathcal{F}) = S(\mathcal{F}_1) + m(m-1)$, which implies that $S(\mathcal{F}_1) = \frac{1}{3}(m-1)((m-1)^2-1)$. Therefore the induction hypothesis yields that up to applying (A) and (B) and re-indexing the elements of $\mathcal{F}_1$, one has $L_i = \{a_1, \ldots, a_i\}$ and $R_i = \{a_1, \ldots, a_{i-1}\}$ for each $i \in \{1, \ldots, m-1\}$, which concludes the proof. □

We shall use Lemma 4.10 to count the number of maximum comma-free triletter codes, and thus recover the statement of Theorem 4.7.

Second proof of Theorem 4.7. Let X be a maximum comma-free triletter code over an alphabet Σ of cardinality $n \geq 3$. Therefore, $|X| = \frac{1}{3} \cdot n(n^2-1)$. Let M be the set of letters appearing at the beginning of a word and at the end of a word, that is, $M \coloneqq \pi_1(X) \cap \pi_3(X)$. Set $X[M] \coloneqq \{N_1N_2N_3 \in X : N_1, N_2, N_3 \in M\}$. It follows from earlier work [**21**] that

$$|X[M]| \leq \sum_{i=2}^{m} i(i-1) = \frac{m(m^2-1)}{3}. \tag{4.1}$$

Indeed we know [**21**] that $|M| \in \{n-1, n\}$ and, in addition, if $|M| = n-1$ then there are precisely $n(n-1)$ words in X that contain the unique letter N in $\Sigma \setminus M$.

We also know [**21**, Theorem 3] that $\pi_{12}(X) \cap \pi_{23}(X) = \varnothing$ and every element in Σ^2 occurs in a word in X, unless $M = \Sigma \setminus \{N\}$ in which case NN does not occur in a word in X. If follows that the subset of Σ^2 appearing in a word in X can be partitioned into the sets L_x, R_x for $x \in \Sigma$, where L_x is the set of letters y such that there is a word in X starting with yx, and R_x is the set of letters y such that there is a word in X ending with xy. (We point out that the sets $L_x \cup R_x$ for $x \in \Sigma$ are the *sections* defined by Golomb, Welch and Delbrück [**21**].) Recall operations (A) and (B) defined in Lemma 4.10: it follows from the definitions that applying any sequence of these operations to X still yield a maximum comma-free triletter code over Σ.

For every order $\sigma \colon \Sigma \to \{1, \ldots, n\}$ on Σ, set

$$X_\sigma \coloneqq \left\{N_1N_2N_3 \in \Sigma^3 : \sigma(N_1) \leq \sigma(N_2) \text{ and } \sigma(N_2) > \sigma(N_3)\right\}.$$

Given an order $\sigma\colon \Sigma \to \{1,\dots,n\}$ of Σ, a maximum comma-free code Y is *σ-canonical* if there exists a sequence of operations (A) and (B), as defined in Lemma 4.10, that transform Y into X_σ.
By Lemma 4.10, if Y is a maximum comma-free code on Σ, and if $\sigma_Y : \Sigma \to \{1,\dots,m\}$ is such that

$$|L_{\sigma_Y{}^{-1}(\{i+1\})}| \times |R_{\sigma_Y{}^{-1}(\{i+1\})}| \geq |L_{\sigma_Y{}^{-1}(\{i\})}| \times |R_{\sigma_Y{}^{-1}(\{i\})}|$$

for each $i \in \{1,\dots,m-1\}$, then Y is σ_Y-canonical. However, the order σ_Y may not be unique: this is why the following counting needs to be done in an amortized way.
We fix an order $\sigma\colon \Sigma \to \{1,\dots,n\}$ on Σ. For X_σ, setting $x_0 \coloneqq \sigma^{-1}(\{1\})$, we have $L_{x_0} = \varnothing = R_{x_0}$ and, for convenience, we redefine L_{x_0} to $\{x_0x_0\}$, which leaves $\sum_{x\in\mathcal{A}} |L_x|\,|R_x|$ unchanged. Observe that now $|L_x| = \sigma(x)$ and $|R_x| = \sigma(x)-1$ for every $x \in \Sigma$. (This would not have been true for x_0 had we not changed the definition of L_{x_0}.) Let (B') be the reverse of operation (B), that is, applying (B') to a code Y yields a code Y' if and only if one can obtain Y' by applying (B) to Y. By applying all sequences of operations (A) and (B') to X_σ that preserve that the order σ_Y for the obtained maximum code Y can still be chosen to be equal to σ, we obtain all σ-canonical maximum comma-free codes.
We observe that operation (B') to x and y yields that $|L_x| = |R_x| = |L_y| = |R_y| = \sigma(x)$, thereby creating an ambiguity on the new order induced on Σ. Since there are two possibilities to order x and y, we shall count such changes with an amortized factor of $\frac{1}{2}$.
We count the number x_n of possibilities inductively, noticing that $x_0 = 1 = x_1$. We distinguish cases regarding whether the operation (B') is performed on the two largest elements according to σ. If this is not the case, then we remove the maximum element $x \coloneqq \sigma^{-1}(\{m\})$ of Σ: considering $X_{\sigma'}$ where $\sigma' \coloneqq \sigma|(\Sigma \setminus \{x\})$, one obtains by induction x_{n-1} different maximum codes on $\Sigma \setminus \{x\}$, each of which can be extended into a maximum code on Σ in two ways, regarding whether the operation (A) is applied to x. Next, if the operation (B) is performed on the two largest elements x and y, then one obtains $|L_x| = |R_x| = |L_y| = |R_y| = n-1$. Removing x and y from Σ, and considering $X_{\sigma''}$ where $\sigma'' \coloneqq \sigma|(\Sigma \setminus \{x,y\})$, one obtains x_{n-2} different codes. There are two ways in which the operation (B') could have been performed on the two largest elements: either directly or after having applied the operation (A) to both of them, which yields $2x_{n-2}$ different codes. However, as hinted at earlier, each of these codes is σ'-canonical for two different orders, namely σ and the order obtained from σ by inverting the two largest elements. Consequently, an amortizing factor of $\frac{1}{2}$ is applied in this case. It follows that, $x_n = 2x_{n-1} + x_{n-2}$ and we infer that

$$x_n = \sum_{i=0}^{\lfloor n/2\rfloor} \binom{n}{2i} 2^i = \frac{(1+\sqrt{2})^n - (1-\sqrt{2})^n}{2} = \left[\frac{(1+\sqrt{2})^n}{2}\right],$$

which concludes the proof. □

The approach developed to establish Theorem 4.7 allows us to perform a similar study also for codes that need not intersect every complete cyclic equivalence class, and hence for which the properties unveiled by Golomb, Welch and Delbrück [**21**] do not hold anymore. We consider different families of codes along these lines in the next subsections.

4.3. Maximum Self-Complementary Comma-Free Triletter Codes. We consider comma-free triletter codes that are π-self-complementary for an involutory transformation $\pi \in S_\Sigma$ with no fixed point. (Involutory transformations with fixed points could be dealt with, at the expense of more tedious notation and analysis; however, given our original biological motivations and the fact that adding fixed points does not change the essence of the argumentation, we omit this case entirely.) If $\pi \in S_\Sigma$ is involutory with no fixed point, then we define $S^\pi_{\mathrm{cf}}(n,\ell)$ to be the number of comma-free

π-self-complementary ℓ-letter codes of maximal size over an alphabet with n letters and $S^{\pi}_{\mathrm{cf}}(n,\ell)$ to be their size — as we shall see, these numbers do not depend on the choice of π.

THEOREM 4.11. *Let Σ be an alphabet of even cardinality $n \in 2\mathbf{N}$ and let $\pi \in S_\Sigma$ be an involutory transformation with no fixed point.*

(1) $S^{\pi}_{\mathrm{cf}}(2,3) = 2$, $S^{\pi}_{\mathrm{cf}}(4,3) = 16$ and if $n \geq 6$ then

$$S^{\pi}_{\mathrm{cf}}(n,3) = \frac{n(n^2-1)}{3} - \frac{n^2}{2} + 2 = \frac{n(2n+1)(n-2)}{6} + 2.$$

(2) $N^{\pi}_{\mathrm{cf}}(2,3) = 2$, $N^{\pi}_{\mathrm{cf}}(4,3) = 4$, $N^{\pi}_{\mathrm{cf}}(6,3) = 54$ and if $n \geq 8$ then

$$N^{\pi}_{\mathrm{cf}}(n,3) = 6^{n/2-1}\left(\frac{n}{2}\right)!.$$

We first prove a characterisation of a slightly more constrained family of comma-free codes: this family will be useful to be able to apply induction to establish Theorem 4.11, and it also seems to be a legitimate family to study on its own. The additional restrictions allow us to use a more direct approach than that used to establish Theorem 4.7. In addition to the self-complementarity, the comma-free triletter codes X we consider are also required to satisfy that $\pi_{12}(X) \cap \pi_{23}(X) = \varnothing$. Notice that this last condition implies that the code is empty if the alphabet has cardinality two.

PROPOSITION 4.12. *Let Σ be an alphabet of even cardinality $n \in 2\mathbf{N}$ and let $\pi \in S_\Sigma$ be an involutory transformation with no fixed point. If X is a π-self-complementary comma-free triletter code such that $\pi_{12}(X) \cap \pi_{23}(X) = \varnothing$, then*

$$|X| \leq \frac{n(n^2-1)}{3} - \frac{n^2}{2}.$$

Furthermore, there is equality if and only if, setting $n' \coloneqq n/2 - 1$, there is an enumeration of the alphabet $\Sigma = \{\alpha_0, \pi(\alpha_0), \ldots, \alpha_{n'}, \pi(\alpha_{n'})\}$ such that

$$X = \sum_{i=0}^{n'-1} \left(X_i \cup \{\alpha_i\}\right) \alpha_i X_i + \sum_{i=0}^{n'-1} X_i \pi(\alpha_i) \left(X_i \cup \{\pi(\alpha_i)\}\right),$$

where $X_i \coloneqq \{\alpha_{i+1}, \pi(\alpha_{i+1}), \ldots, \alpha_{n'}, \pi(\alpha_{n'})\}$. It follows that the number of such codes X of maximal size is precisely

$$\left(\frac{n}{2}\right)! \cdot 2^{n/2-1}.$$

PROOF. We proceed by induction on the even cardinality n of the alphabet Σ, the statement being trivially true if $n = 2$, that is, if $\Sigma = \{a, \pi(a)\}$. Now let $n \geq 4$ and let X be a π-self-complementary comma-free triletter code of maximal size over an alphabet Σ of cardinality n. To make the notation lighter, we assume without loss of generality that $\Sigma = \{0, \ldots, n-1\}$ and $\pi(a) = n - 1 - a$ for each $a \in \{0, \ldots, n/2 - 1\}$.

Similarly as in the proof of Theorem 4.7, we define for every $a \in \Sigma$ the set L_a of all letters $x \in A$ such that X contains a word starting with xa, and the set R_a of all letters $y \in A$ such that X contains a word ending with ay. We set $\ell_a \coloneqq |L_a|$ and $r_a \coloneqq |R_a|$ for each $a \in \Sigma$. It follows that X is contained in $\sum_{a=0}^{n-1} L_a \cdot a \cdot R_a$. Notice that $L_a = R_{n-1-a}$ and $R_a = L_{n-1-a}$ for each $a \in \{0, \ldots, n/2 - 1\}$ due to the π-self-complementarity of X.

Up to permuting $\{0, \ldots, n/2-1\}$ and replacing X with $\pi(X)$, we can suppose without loss of generality that $\ell_0 \geq \max\{\ell_a, r_a\}$ for each $a \in \{0, \ldots, n-1\}$. Notice that $n - 1 - a \notin L_a \cup R_a$ for every $a \in \{0, \ldots, n-1\}$ because X is π-self-complementary and $\pi_{12}(X) \cap \pi_{23}(X) = \varnothing$. Consequently, there exists a non-negative integer m such that $\ell_0 = n - 1 - m$. Moreover, because $0 \notin L_0 \cap R_0$ and $\ell_0 \geq r_0$, we can write $r_0 = n - 2 - k$ for some non-negative integer k.

We want to count the number of words in X that contain 0. Our strategy is to remove from X the set W composed of all words in X that contain the letter 0, and show that $|W| \leq n^2 - 3n + 2$. It then follows by π-self-complementarity that the number x of words in X that contain 0 or $n-1$ is at most $2|W| \leq 2n^2 - 6n + 4$. Now, deleting from X all such words yields a π-self-complementary comma-free triletter code X' over the alphabet $\{1, \dots, n-2\}$ such that $\pi_{12}(X') \cap \pi_{23}(X') = \varnothing$. We know by induction that $|X'| \leq (n-2)^3/3 - (n-2)^2/2 - (n-2)/3$, and we know the shape of X' if there is equality. We would therefore deduce that

$$\begin{aligned} |X| &\leq 2n^2 - 6n + 4 + \frac{(n-2)^3}{3} - \frac{(n-2)^2}{2} - \frac{n-2}{3} \\ &= \frac{n^3}{3} - \frac{n^2}{2} - \frac{n}{3}. \end{aligned}$$

Furthermore, there would be equality only if X' has maximal size and $x = 2n^2 - 6n + 4$ (in particular, note for (much) later on that then no word in X contains both 0 and $n-1$, for otherwise $x < 2|W|$). So let us bound the size of $W = \{w \in X : 0 \in w\}$ from above. Because $\pi_{12}(X) \cap \pi_{23}(X) = \varnothing$, we know that for all letters a and b, if $a \in L_b$ then $b \notin R_a$. In symbols, $\{a \in \Sigma : b \in R_a\} \cap L_b = \varnothing$. Similarly, if $a \in R_b$ then $b \notin L_a$. Let us write $W = W_0 \cup W_1$ where $W_0 \coloneqq \{w \in W : \pi_2(w) = 0\}$, and $W_1 \coloneqq W \setminus W_0$. Using the notation previously introduced, we have $|W_0| \leq (n-1-m)(n-2-k)$. To evaluate the size of W_1, set $M_0 \coloneqq \{1, \dots, n-2\} \setminus L_0$, so $m_0 \coloneqq |M_0| \in \{m-1, m\}$. Further, if $a \in \{1, \dots, n-2\}$, then $0 \in R_a$ only if $a \in M_0$. Similarly, setting $K_0 \coloneqq \{1, \dots, n-2\} \setminus R_0$, we have $k_0 \coloneqq |K_0| \in \{k, k+1\}$ and if $a \in \{1, \dots, n-2\}$, then $0 \in L_a$ only if $a \in K_0$. Since $m_0 = m$ if and only if $0 \in L_0$ and $k_0 = k+1$ if and only if $0 \in R_0$, the fact that $0 \notin L_0 \cap R_0$ implies that $m_0 + k_0 \leq m + k$. We bound the number of words in W_1 by bounding, for each $a \in \{1, \dots, n-2\}$, the number x_a of words w in W_1 with $\pi_2(w) = a$.

Fix $a \in \{1, \dots, n-2\}$. If $0 \in R_a$, then the number of words in X ending with $a0$ is at most ℓ_a, which is at most $n-1-m$. Similarly, if $0 \in L_a$ then the number of words in X starting with $0a$ is at most r_a, which is at most $n-1-m$. Consequently,

$$\begin{aligned} |W_1| &\leq \sum_{a \in M_0 \cup K_0} (n-1-m) \\ &\leq (m_0 + k_0)(n-1-m) \\ &\leq (m+k)(n-1-m). \end{aligned}$$

It follows that

$$\begin{aligned} |W| &\leq (n-1-m)(n-2-k) + (m+k)(n-1-m) \\ &= (n-1-m)(n-2+m) \\ &= n^2 - 3n + 2 - m(m-1). \end{aligned}$$

As a result, $|W| \leq n^2 - 3n + 2$ with equality if and only if $m \in \{0, 1\}$ and $m_0 + k_0 = m + k$. Therefore, the size of X is indeed at most $n^3/3 - n^2/2 - n/3$, with equality only if all inequalities written so far are equalities: in particular, $\ell_a = n-1-m$ if $a \in M_0$ and $r_a = n-1-m$ if $a \in K_0$; $m_0 + k_0 = m + k$; $m \in \{0, 1\}$ and X' is of maximal size. It only remains to prove that if X is of maximal size then it has the announced form. So assume that X has maximal size. Then X' is of maximal size over the alphabet $\Sigma \setminus \{0, n-1\} = \{1, \dots, n-2\}$, and hence the induction hypothesis tells us that up to

permuting $\{1,\dots,n/2-2\}$,

$$X' = \sum_{i=1}^{n/2-2} \{i, i+1, \dots, n-2-i\} i \{i+1, \dots, n-2-i\} + \sum_{i=1}^{n/2-2} \{i+1, \dots, n-2-i\}(n-1-i)\{i+1, \dots, n-1-i\}.$$

If $m = 0$, then $L_0 = \{0, \dots, n-2\}$ and hence $0 \notin \bigcup_{a=1}^{n-1} R_a$. This implies that $|R'_a| \geq |R_a| - 1$ for each $a \in \{1, \dots, n-2\}$. Further, notice that $k_0 = k$ because $m_0 = m$ and $k_0 + m_0 = k + m$. Since

$$\begin{aligned} n^2 - 3n + 2 = |W| &\leq (n-1)(n-2-k) + \sum_{a \in K_0} r_a \\ &\leq n^2 - 3n + 2 - k(n-1) + k(n-1), \end{aligned}$$

we deduce that either $k = 0$ or $r_a = n-1$ for each $a \in K_0 \subseteq \{1, \dots, n-2\}$. However, we know by the shape of X' that $|R'_a| \leq n-3$, and hence $r_a = |R_a| \leq n-2$ for each $a \in \{1, \dots, n-2\}$. It thus follows that $K_0 = \varnothing$, and hence $R_0 = \{1, \dots, n-2\}$, which completes the proof in this case.
We end the proof by showing that if $m = 1$ then X is not maximum. Suppose indeed that $m = 1$, that is, $L_0 = \{0, \dots, n-2\} \setminus \{i\}$ with $i \in \{0, \dots, n-2\}$. If $i = 0$, then $0 \notin \cup_{a=1}^{n-2} R_a$, and hence

$$\begin{aligned} |W| &\leq (n-2)(n-2-k) + k_0(n-2) \\ &\leq (n-2)^2 + (n-2) \\ &= n^2 - 3n + 2, \end{aligned}$$

with equality only if $k_0 = k+1$, so $K_0 \neq \varnothing$ and $0 \in R_0$. Further, $0 \in L_a$ and $r_a = n-2$ for each $a \in K_0$. Letting $a \in K_0$, since $0 \notin R_a$ we must have $n-1 \in R_a$. Therefore, a word in X contains both 0 and $n-1$, which, as reported (much) earlier in the proof is not possible.
Therefore, there exists a unique $i \in \{1, \dots, n-2\}$ such that $0 \in R_i$, since $\ell_0 = n-2$. In particular, $0 \in L_0$ and hence $0 \notin R_0$, which implies that $m_0 = m$ and $k_0 = k$. Since $0 \notin R_{n-1-i}$, the π-self-complementarity of X implies that $n-1 \notin L_i$. As

$$\begin{aligned} |W| &\leq (n-2)(n-2-k) + k_0(n-2) + \ell_i \\ &= (n-2)^2 + \ell_i \\ &= n^2 - 4n + 4 + \ell_i, \end{aligned}$$

we deduce that $\ell_i = n-2$ and $0 \in L_i$, so $i \notin R_0$ and thus $r_0 \leq n-3$ as none of 0, i and $n-1$ belongs to R_0. Still since $n-1 \notin L_i$, we know that $|L'_i| \geq |L_i| - 1 = \ell_i - 1 = n-3$. It thus follows that $i = 1$ and hence $L_1 = \{0, \dots, n-3\}$. Since $0 \in L_1 \cap R_1$, the number of words w in X with $\pi_2(w) = 2$ that contain 0 is at most $\ell_1 + r_1 - 1$ (as otherwise the word 010 would be counted twice). Consequently,

$$\begin{aligned} |W| &\leq (n-2)(n-2-k) + \ell_1 + r_1 - 1 + (k-1)(n-2) \\ &\leq (n-2)^2 + \ell_1 - 1 \\ &= n^2 - 3n + 1, \end{aligned}$$

which concludes the proof. □

We are now in a position to establish Theorem 4.11.

Proof of Theorem 4.11. Fix an alphabet Σ of cardinality $n \in 2\mathbf{N}$. Let X be a maximum π-self-complementary comma-free triletter code over Σ. Let (L, M, R) be a partition of Σ such that

- $M \coloneqq \pi_1(X) \cap \pi_3(X)$;

- $\pi_1(X) \setminus M \subseteq L$; and
- $\pi_3(X) \setminus M \subseteq R$.

Because X is π-self-complementary, $\overleftarrow{\pi(L)} = R$, and $\overleftarrow{\pi(M)} = M$. Set $X' \coloneqq LM + MR + LR$. We observe that $\pi_{12}(X) \cap \pi_{23}(X) \subseteq X' \cup MM$. Indeed, if $w_1w_2w_3 \in X$ then $w_1 \in \pi_1(X) \subseteq L \cup M$ and $w_3 \in \pi_3(X) \subseteq M \cup R$, and hence

$$\pi_{12}(X) \subseteq LL + LM + LR + ML + MM + MR \quad \text{and}$$
$$\pi_{23}(X) \subseteq LM + MM + RM + LR + MR + RR.$$

We define (L_2, M_2, R_2) to be a partition of X' such that

- $M_2 \coloneqq X' \cap \pi_{12}(X) \cap \pi_{23}(X)$;
- $L_2 \coloneqq X' \cap \pi_{12}(X) \setminus \pi_{23}(X)$; and
- $R_2 \coloneqq X' \cap \pi_{23}(X) \setminus \pi_{12}(X)$.

Again because X is self-complementary, $\overleftarrow{\pi(L_2)} = R_2$ and $\overleftarrow{\pi(M_2)} = M_2$. In addition, M_2 contains all the π-self-complementary diletter words occurring in words in X'. We assert that

$$X \subseteq MMM + L(M_2 + R_2) + (L_2 + M_2)R + L_2M + MR_2. \tag{4.2}$$

To establish (4.2), we shall make several cases so as to ease the checking. Let $w = w_1w_2w_3$ be an element of X. In particular, $w_i \in \pi_i(X)$ and hence $w_1 \in L \cup M$ and $w_3 \in M \cup R$. Since $w_1w_2 \in \pi_{12}(X)$, we know that if $w_1w_2 \in X'$, then $w_1w_2 \in L_2 + M_2$. Similarly, if $w_2w_3 \in X'$, then $w_2w_3 \in M_2 + R_2$. By the symmetry of the roles played by L and R, we may assume that $w_2 \in L \cup M$.

(1) If $w_2 \in L$ then $w_2w_3 \in LM + LR \subseteq X'$, and hence $w_2w_3 \in M_2 + R_2$. Consequently, if $w_1 \in L$ then $w \in L(M_2 + R_2)$. Otherwise, $w_1 \in M$ and hence either $w \in MR_2$ or $w_2w_3 \in M_2$. In the latter case, however, there would exist $w_4 \in \Sigma$ such that $w_2w_3w_4 \in X$ by the definition of M_2 and, as $w_1 \in M$, there would exist $w_5w_6 \in \Sigma^2$ such that $w_5w_6w_1 \in X$, which contradicts that X is comma-free: the concatenation of $w_5w_6w_1$ and $w_2w_3w_4$ contains the word $w = w_1w_2w_3$.

(2) If $w_2 \in M$, then either $w \in MMM$ or $w_1 \in L$ or $w_3 \in R$. By symmetry, we may assume that the former holds, *i.e.*, $w_1 \in L$. Consequently, $w_1w_2 \in \pi_{12}(X) \setminus MM$ and therefore $w_1w_2 \in L_2 + M_2$.
   (a) If $w_1w_2 \in L_2$ then $w \in L_2(M + R)$.
   (b) If $w_1w_2 \in M_2$, then we deduce similarly as in case (1) that $w_3 \notin M$. It follows that $w_3 \in R$ and consequently $w \in M_2R$, which concludes the proof of (4.2).

We bound the size of X by bounding the size of the right side of (4.2) We start by computing the maximal size of $X[M] \coloneqq X \cap MMM$. Notice that $X[M]$ is a π-self-complementary comma-free triletter code over M, with the additional property that $\pi_{12}(X[M]) \cap \pi_{23}(X[M]) = \varnothing$. Indeed, if there exist a word $w_1w_2w_3$ in $X[M]$ and a letter $w_4 \in M$ such that $w_4w_1w_2 \in X[M]$, then X cannot be comma-free: the definition of M ensures that there exist w_5 and w_6 in Σ such that $w_5w_6w_4 \in X$, and therefore the concatenation of the two words $w_5w_6w_4$ and $w_1w_2w_3$ (which are both in X) contains the word $w_4w_1w_2$, which also belongs to X. Consequently, Proposition 4.12 ensures that the size of $X[M]$ is at most

$$\frac{m(m^2-1)}{3} - \frac{m^2}{2}.$$

Noticing that $|L| = \frac{n-m}{2} = |R|$ and $|L_2| = \frac{|X'|-|M_2|}{2} = |R_2|$, one sees that the size of X is at most

$$\begin{aligned} & |X[M]| + 2|L|(|M_2| + |R_2|) + 2|L_2| \cdot |M| \\ = \; & |X[M]| + (n-m)\frac{|X'| + |M_2|}{2} + m(|X'| - |M_2|), \end{aligned}$$

which, since $|X'| = \frac{1}{4}(n-m)(n+3m)$, is at most

$$\frac{m^3}{3} - \frac{m^2}{2} - \frac{m}{3} + \frac{(n^2-m^2)(n+3m)}{8} + \frac{(n-3m)}{2}|M_2|. \tag{4.3}$$

Let us maximise (4.3). We consider two cases regarding whether $m \leq n/3$.

- If $m \leq n/3$, then the maximum is attained only if $|M_2| = |X'|$. In this case, the function becomes

$$\frac{m^3}{3} - \frac{m^2}{2} - \frac{m}{3} + \frac{(n-m)^2(n+3m)}{4}, \tag{4.4}$$

  which, given than m is an even integer, attains a maximum value that is at most $\frac{n^3}{3} - \frac{n^2}{2} - \frac{n}{3} + 2$. Indeed, (4.4) for $m \leq n/3$ attains its maximum value when $m = \frac{5n}{13} - \frac{2}{39}\sqrt{27n^2+45n+48} + \frac{2}{13}$, the maximum value being

$$\begin{aligned} n^3\left(\frac{113}{507} + \frac{4}{507}\sqrt{27+45/n+48/n^2}\right) - n^2\left(\frac{37}{338} - \frac{20}{1521}\sqrt{27+45/n+48/n^2}\right) \\ -n\left(\frac{95}{507} - \frac{64}{4563}\sqrt{27+45/n+48/n^2}\right) - \frac{10}{169}. \end{aligned}$$

  If $n \geq 8$, then $\sqrt{27+45/n+48/n^2} \leq \sqrt{27+45/8+3/4}$, and substituting one readily checks that the obtained value is less than $\frac{n^3}{3} - \frac{n^2}{2} - \frac{n}{3} + 2$. If $n = 6$, then (4.4) becomes

$$\frac{13}{12}m^3 - 8m^2 + \frac{26}{3}m + 54,$$

  which for $m \in \{0, 2\}$ is maximised when $m = 0$, reaching the value 54 so exactly $\frac{6^3}{3} - \frac{6^2}{2} - \frac{6}{3} + 2$. If $n = 2$ or $n = 4$, then since m is even and no more than $n/3$, we have $m = 0$ and the obtained functions have respective values 2 and 16.
- If $m > n/3$, then the maximum is attained only if M_2 is minimised. This means that M_2 is equal to the set of the π-self-complementary dinucleotides in X'. It follows that $M_2 = \{x\pi(x) : x \in L\}$ and thus $|M_2| = |L| = \frac{n-m}{2}$. In this case, the function becomes

$$\frac{m^3}{3} - \frac{m^2}{2} - \frac{m}{3} + \frac{(n^2-m^2)(n+3m)}{8} + \frac{(n-3m)(n-m)}{4},$$

  which, given that m is an even integer in $[0, n]$ and n is even, is maximised only if $m \in \{n-2, n\}$. When $n = 2$, since $n - 2 < n/3$ we deduce that $m = 2$, which yields the value 0. When $n \neq 2$, calculations show the maximum to be attained only when $m = n - 2$, and it is then $\frac{n^3}{3} - \frac{n^2}{2} - \frac{n}{3} + 2$. We point out that this value is less than 16 if $n = 4$, and equal to 54 if $n = 6$.

There remains to count the number of possible codes of maximal size.

- When $m > n/3$, we have shown that the size is maximal if and only if $|M| = n - 2$, $|L| = 1 = |R|$, $M_2 = LR$ and $LM + MR = L_2 + R_2$. There are n possible choices for the partition (L, M, R) of Σ satisfying $\pi(L) = R$ and $\pi(M) = M$. The choice of the partition (L_2, R_2) of $LM + MR$ must be such that $\overleftarrow{\pi(L_2)} = R_2$ and $LR_2 \cap L_2R = \varnothing$. So for each $x \in M$, either $Lx + xR \subseteq L_2$, or $Lx + xR \subseteq R_2$, or $Lx \subseteq R_2, xR \subseteq L_2$. Note that this determines the choice of $L\pi(x) + \pi(x)R$, so there are three possible outcomes for each pair

of complementary elements in M. The number of choices for the partition (L_2, R_2) is therefore $3^{m/2} = 3^{n/2-1}$. Finally, Proposition 4.12 ensures that there are precisely $2^{m/2-1}\left(\frac{m}{2}\right)!$ possibilities for $X[M]$, yielding a total of

$$6^{n/2-1}\left(\frac{n}{2}\right)!$$

different maximum π-self-complementary comma-free codes. These count all such codes for every $n \geq 8$, and some of the codes when $n = 6$.

- When $m \leq n/3$, a code of maximal size is produced only when $m = 0$ and $n \in \{2, 4, 6\}$. Because then $M = \varnothing$, the code is $LLR + LRR$ with $|L| = n/2 = |R|$. There are $2^{n/2}$ choices for a partition (L, R) of Σ such that $\overleftarrow{\pi(L)} = R$, yielding two codes when $n = 2$, four codes when $n = 4$ and eight codes when $n = 6$.

The total number of codes when $n = 6$ is thus

$$8 + 6^{6/2-1}\left(\frac{6}{2}\right)! = 224.$$

□

REMARK 4.13. Theorem 4.11 for $n = 4$ provides a theoretical explanation of the well-known fact [**26**] that the maximal size of self-complementary comma-free genetic codes is 16.

5. Characterisation of ℓ-letter Strong Comma-Free Codes

We define $N_{\mathrm{scf}}(n, \ell)$ to be the number of strong comma-free ℓ-letter codes of maximal size over an alphabet Σ with n letters and we let $S_{\mathrm{scf}}(n, \ell)$ be their size.

For comparison, Table 3 shows the calculated numbers of strong comma-free diletter codes of maximal size over alphabets of cardinalities in $\{2, \ldots, 15\}$, computed using Theorem 4.1. Using the formula provided by Theorem 4.1 (with $p = 1$), one easily sees that $N_{\mathrm{scf}}(2k, 2) = N_{\mathrm{scf}}(2k-1, 2)$ for every $k \in \mathbf{N}$.

n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$N_{\mathrm{scf}}(n, 2)$	2	6	6	20	20	70	70	252	252	924	924	3432	3432	12870

TABLE 3. The number of strong comma-free diletter codes of maximal size over an alphabet of cardinality n for $n \in \{2, \ldots, 15\}$.

We aim now to characterise maximal strong comma-free ℓ-letter codes for $\ell \geq 3$. To this end, we introduce the following notation. Given an alphabet Σ and a positive integer ℓ, let $\mathscr{M}_\ell$ be the collection of all maximal strong comma-free circular ℓ-letter codes over Σ. (In particular, $\mathscr{M}_1 = \Sigma$.) Further, let $\mathscr{P}_\ell$ be the collection of all sequences $((L_i, R_i))_{1 \leq i \leq \ell}$ where

(1) (L_1, R_1) is a partition of Σ into two **non-empty** parts; and
(2) (L_i, R_i) is a partition of $\sum_{j=1}^{i-1} L_j R_{i-j}$ for every $i \in \{2, \ldots, \ell\}$.

We prove the following theorem and, actually, a stronger statement describing the structure of all maximal strong comma-free ℓ-letter codes in terms of the maximal strong comma-free $(\ell - 1)$-letter codes — the stronger statement, being slightly technical, is given at the beginning of the proof.

THEOREM 5.1. *For every integer $\ell \geq 3$, there exists a bijection f_ℓ between $\mathscr{M}_\ell$ and $\mathscr{P}_\ell$.*

PROOF. Fix an integer $\ell \geq 3$. In addition to the statement of the theorem, we prove that for every $X \in \mathscr{M}_\ell$ if $f_\ell(X) = ((L_i, R_i))_{1\leq i\leq \ell}$ then $X = \sum_{i=1}^{\ell-1} L_i R_{\ell-i}$. Further, for each $j \in \{1, \dots, \ell\}$ the j-letter code $X_j \coloneqq \sum_{i=1}^{j-1} L_i R_{j-i}$ is a maximal strong comma-free code over Σ (and hence $f_j(X_j) = ((L_i, R_i))_{1\leq i\leq j}$).

Let X be a maximal strong comma-free code over Σ^ℓ. First note that since X is strong comma-free, no word can appear both as a prefix and as a suffix in X. (In particular, a letter that starts a word in X cannot end a word in X.) What follows might be better digested by looking simultaneously at Example 5.2.

Let us construct the sets (L_i, R_i) corresponding to X, for $i < \ell$.

- We let (L_1, R_1) be a partition of Σ such that every letter beginning a word in X belongs to L_1 while every letter ending a word in X belongs to R_1.
- Once (L_j, R_j) is constructed for every $j < i$, we set $X_i \coloneqq \sum_{j=1}^{i-1} L_j R_{i-j}$. We then let (L_i, R_i) be a partition of X_i such that L_i consists of all the words in X_i that appear as a prefix of a word in X, and hence R_i contains, in particular[2], all words in X_i that appear as a suffix of a word in X.

Now set $X_\ell \coloneqq \sum_{i=1}^{\ell-1} L_i R_{\ell-i}$. We show that $X \subseteq X_\ell$. Let $w \in X$. We set $L \coloneqq \sum_{i<\ell} L_i$ and $R \coloneqq \sum_{i<\ell} R_i$. We build a sequence of decompositions of w in $L(L+R)^*R$ of decreasing lengths with a last term in LR, which ensures that $w \in X_\ell$.

To this end, we encode each decomposition by a binary word over the alphabet $\{l, r\}$. First, we define $p_0(w)$ as follows: $p_0(w) \in l(l+r)^{\ell-2}r$ with $p_0(w)_i = l$ if $w_i \in L_1$, and $p_0(w)_i = r$ if $w_i \in R_1$. Now, notice that there must be at least one occurrence of lr in every word in $l(l+r)^*r$ and hence in particular in $p_0(w)$. For every non-negative integer k, if $p_k(w) \in l(l+r)^+r$ then we define $p_{k+1}(w) \in l(l+r)^*r$ as follows: every occurrence of lr in $p_k(w)$ corresponds to a subword w' of w in $L_iR_j \subseteq X_{i+j}$ for some positive integers i and j; we obtain $p_{k+1}(w)$ by replacing each of these occurrences by l if $w' \in L_{i+j}$ and by r if $w' \in R_{i+j}$. Thus $p_{k+1}(w) \in l(l+r)^*r$ (and hence lr occurs in $p_{k+1}(w)$). Notice also that the length of $p_{k+1}(w)$ is positive and less than that of $p_k(w)$. Therefore the procedure stops at some step k_0, and then $p_{k_0}(w) = lr$, which guarantees that $w \in LR \subseteq X_\ell$.

We now show that $X_\ell \subseteq X$. To this end, it is enough to show that X_ℓ is itself a strong comma-free code. Let us assume for the sake of contradiction that there exist two words $w \in X_i$ and $w' \in X_j$ such that a suffix s of w is a prefix p' of w', and let us choose them in such a way that $i+j$ is minimised. Let $w = w_0w_1$ and $w' = w'_0w'_1$ be the respective decompositions in LR of w and w'. Because $L \cap R = \varnothing$, we know that $w_1 \neq w'_0$ and hence either $s \neq w_1$ or $p' \neq w'_0$. By symmetry of the following argument, we may assume that the former is true. Consequently, either $|s| > |w_1|$ and then w_0 has a suffix that is a prefix of w_1, or $|s| < |w_1|$ and then w_1 has a suffix that is a prefix of w_1. Either way, this contradicts the minimality of $i+j$. We conclude that X_ℓ is indeed a strong comma-free code, and hence $X = X_\ell$. The statement follows. □

EXAMPLE 5.2. Let us consider the alphabet $\Sigma = \{0, 1, 2\}$ of cardinality 3, and the circular strong comma-free code $X = \{001, 021, 201, 221\}$, which is indeed of maximal size. Following the notation in the proof, one has $L_1 = \{0, 2\}$ and $R_1 = \{1\}$. By definition, $X_2 = L_1R_1 = \{01, 21\}$. Since L_2 is composed of all elements in X_2 that appear as a prefix in a word in X, we have $L_2 = \varnothing$. Similarly, R_2 is composed of all elements in X_2 that appear as a suffix in a word in X, and hence $R_2 = \{21, 01\}$. As expected, $X_3 = L_1R_2 + L_2R_1 = X$. Following the argument showing that $X \subseteq X_3$, consider now

[2]As it will follow from the forthcoming arguments, R_i consists precisely of the words in X_i that appear as a suffix of a word in X, that is, every word in X_i is either the prefix or the suffix of a word in X. This fact is, however, not needed at this point.

the word $w \coloneqq 021 \in X$. The first decomposition is llr. The occurrence of lr corresponds to the word 21, which belongs to $L_1R_1 \subseteq X_2$ and, further, to R_2. Consequently we replace this occurrence by r, thereby obtaining lr and thus confirming that $w \in X_3$.

Theorem 5.1 provides a way to construct maximal strong comma-free codes. For instance, for triletter words (*i.e.* $\ell = 3$) over an alphabet Σ of cardinality $n \geq 2$, constructing a maximal strong comma-free code amounts to choosing

(1) a partition (L_1, R_1) of Σ into two non-empty parts; and
(2) a partition (L_2, R_2) of L_1R_1 into two parts.

One then obtains the maximal strong comma-free code X defined by

$$X \coloneqq \{w_1w_2w_3 \,:\, w_1w_2 \in L_2 \text{ and } w_2w_3 \in R_2\}\,.$$

From Theorem 5.1 follows a closed formula for the number of different maximal strong comma-free ℓ-letter codes over an alphabet of cardinality ℓ.

Corollary 5.3. *Let n be an integer greater than* 1*. The number of different maximal strong comma-free triletter codes over an alphabet of cardinality n is*

$$\sum_{m=1}^{n-1} \binom{n}{m} 2^{m(n-m)}. \tag{5.1}$$

Proof. Let Σ be an alphabet of cardinality $n \in \mathbf{N}$. As reported before Corollary 5.3, every maximal triletter code over Σ corresponds to a choice of a partition (L_1, R_1) of Σ into two non-empty parts and, for each such choice, to a choice of a partition (L_2, R_2) of L_1R_1 into two parts. To count this, fix an arbitrary choice for (L_1, R_1). Set $m \coloneqq |L_1|$, hence $|L_1R_1| = m(n-m)$. It follows that there are $2^{m(n-m)}$ different choices for (L_2, R_2). Now it remains to notice that there are $\binom{n}{m}$ partitions (L_1, R_1) of Σ such that $|L_1| = m$. The announced formula follows since we impose that $L_1 \neq \varnothing \neq R_1$. □

Example 5.4. We provide examples for the formula (5.1) given by Corollary 5.3 for alphabets of cardinality at most 4. We also illustrate the way to build them all offered by Theorem 5.1.

(1) Corollary 5.3 ensures that there are 4 different maximal strong comma-free binary codes. Indeed, if $\Sigma = \{0, 1\}$ then there are exactly two choices for (L_1, R_1), each of which yields two different codes of size 1. These four codes are

$$\{001\}, \{011\},$$
$$\{110\}, \{100\}.$$

(2) If $\Sigma = \{0, 1, 2\}$, then the number of different maximal strong comma-free triletter codes over an alphabet Σ of cardinality 3 is 24. There are indeed six choices for (L_1, R_1). For each of them, one has four choices for (L_2, R_2), yielding one maximal code of size 2, two of size 3 and one of size 4 for a total of $6 \cdot (1 + 2 + 1) = 24$. The 24 maximal strong comma-free codes over $\{0, 1, 2\}$ are listed below.

$$\{001, 002\}, \{002, 011, 012\}, \{001, 021, 022\}, \{011, 012, 021, 022\},$$
$$\{002, 012, 102, 112\}, \{012, 112, 022\}, \{002, 102, 122\}, \{022, 122\},$$
$$\{001, 201, 021, 221\}, \{021, 221, 011\}, \{001, 201, 211\}, \{011, 211\},$$
$$\{010, 112\}, \{112, 100, 102\}, \{110, 120, 122\}, \{010, 012, 120, 122\},$$
$$\{110, 120, 210, 220\}, \{120, 220, 100\}, \{110, 210, 200\}, \{100, 200\},$$
$$\{220, 221\}, \{221, 200, 201\}, \{220, 210, 211\}, \{200, 201, 210, 211\}.$$

(3) Over the genetic alphabet $\mathcal{B} = \{A, C, G, T\}$, there are 160 different maximal strong comma-free triletter codes. Indeed, first there are 6 partitions (L_1, R_1) of $\mathcal{B}$ such that $|L_1| = 2 = |R_1|$, each of which yielding a set L_1R_1 of size 4 and therefore $2^4 = 16$ different maximal strong comma-free codes over $\mathcal{B}$. Second, there are 8 partitions (L_1, R_1) of $\mathcal{B}$ such that $|L_1| = 1$ or $|R_1| = 1$, each yielding a set L_1R_1 of size 3 and therefore $2^3 = 8$ different maximal strong comma-free codes. For instance, if $L_1 = \{A\}$ and $R_1 = \{C, G, T\}$, then the 8 such codes are

$$\{AAC, AAG, AAT\}, \{ACC, ACG, ACT, AGC, AGG, AGT, ATC, ATG, ATT\},$$
$$\{AAG, AAT, ACC, ACG, ACT\}, \{AAC, AGC, AGG, AGT, ATC, ATG, ATT\},$$
$$\{AAC, AAT, AGC, AGG, AGT\}, \{AAG, ACC, ACG, ACT, ATC, ATG, ATT\},$$
$$\{AAC, AAG, ATC, ATG, ATT\}, \{AAT, ACC, ACG, ACT, AGC, AGG, AGT\}.$$

5.1. The Number of Strong Comma-Free Self-Complementary Triletter Codes. The developed framework also allows us to study maximal self-complementary strong comma-free triletter codes. For every alphabet Σ and every $\pi \in S_\Sigma$, we define $N^\pi_{\mathrm{scf}}(n, \ell)$ to be the number of strong comma-free ℓ-letter codes of maximal size that are π-self-complementary, and $S^\pi_{\mathrm{scf}}(n, 3)$ their size. Due to our original motivation arising from biology, we restrict to involutory transformations without fixed points, although the following result could be extended to any involutory transformation $\pi \in S_\Sigma$ with k fixed points such that $|\Sigma| - k$ is even.

PROPOSITION 5.5. *Let n be a positive and even integer and Σ an alphabet of cardinality n. If $\pi \in S_\Sigma$ is an involutory transformation with no fixed points, then all maximal different π-self-complementary strong comma-free triletter codes over Σ have size $n\binom{n/2}{2}$ and hence all of them are actually of maximal size, so*

$$S^\pi_{\mathrm{scf}}(n, 3) = n\binom{n/2}{2} = \frac{n^2(n-2)}{8}.$$

It follows that

$$N^\pi_{\mathrm{scf}}(n, 3) = 2^{\binom{n/2+1}{2}} = 2^{n(n+2)/8}. \tag{5.2}$$

PROOF. The argument is similar to that establishing Theorem 5.1, with the additional requirement that both partitions must satisfy the π-self-complementary conditions, which are $\overleftarrow{c(L_1)} = R_1$ and $\overleftarrow{c(L_2)} = R_2$. The number of such partitions (L_1, R_1) is $2^{n/2}$. Fix such a partition and let S be the set of π-self-complementary dinucleotides in L_1R_1. The size of S is $\frac{n}{2}$ and that of L_1R_1 is $n^2/4$ since L_1 and R_1 have the same size. Consequently, the number of partitions (L_2, R_2) of $L_1R_1 \setminus S$ with $\overleftarrow{c(L_2)} = R_2$ is

$$2^{\binom{n/2}{2}}.$$

Again, X is $L_1R_2 + L_2R_1$, which shows that the size of X is $n\binom{n/2}{2}$. □

EXAMPLE 5.6. We consider the genetic alphabet $\mathcal{B} = \{A, C, G, T\}$ along with the involutory bijection c (Definition 2.3). Every strong comma-free triletter code of maximal size that is self-complementary corresponds to one choice for (L_1, R_1) and a subsequent choice for (L_2, R_2). There are four valid choices for L_1, namely $\{A, C\}$, $\{A, G\}$, $\{T, C\}$ and $\{T, G\}$. If we choose $L_1 = \{A, C\}$ (and hence $R_1 = \{T, G\}$) then $L_1R_1 = \{AT, AG, CT, CG\}$. We now choose (L_2, R_2), which is a partition of L_1R_1 deprived of all the self-complementary dinucleotides it may contain, that is deprived of AT and CG in our case. Thus there are exactly two choices for (L_2, R_2), *i.e.*, $(\{AG\}, \{CT\})$ and $(\{CT\}, \{AG\})$. Each choice

yields a code of size 4. In total there are thus $4 \cdot 2 = 8$ different codes of size 4, listed below.

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

We end with two examples showing that the techniques we introduced in this work can be applied to other settings. Recently, the notion of "mixed codes", mixing dinucleotides, trinucleotides and tetranucleotides over the genetic alphabet $\mathcal{B}$, have been introduced [**13**]: circular mixed codes have been constructed and biologically-inspired properties of such codes studied. In particular it was shown [**13**, Proposition 7] that the maximal size of a self-complementary mixed comma-free code in $\mathcal{B}^2 \cup \mathcal{B}^3$ is 20, and there are precisely 4 such codes. The techniques developed here allow us to generalise this result to any alphabet Σ of even cardinality and any involution $\pi\colon \Sigma \to \Sigma$. The proof of the following statement uses an approach similar to those presented, and we omit it.

THEOREM 5.7. *Let Σ be an alphabet of even cardinality n and $\pi\colon \Sigma \to \Sigma$ an involution with no fixed point.*

(1) If $n \geq 8$ then the maximum π-self-complementary comma-free mixed circular codes in $\Sigma^2 \cup \Sigma^3$ have size $\frac{n(n^2+5)}{3} - \frac{n^2}{2} - 1$ and there are exactly $6^{n/2-1}\left(\frac{n}{2}\right)!$ such codes.

(2) If $n = 4$ or $n = 6$, then the maximum π-self-complementary comma-free mixed circular codes in $\Sigma^2 \cup \Sigma^3$ are of sizes 20 or 63, and there are exactly 4 or 224 such codes, respectively.

We end by pointing out that the approach used to study strong comma-free ℓ-letter codes (Theorem 5.1) can be extended to mixed codes: using analogous notations, one can obtain a maximal mixed strong comma-free code by taking $X = \bigcup_{i=(\ell+1)/2}^{\ell} X_i$ when ℓ is odd, and either $X = L_{\ell/2} \cup \bigcup_{i=\ell/2+1}^{\ell} X_i$ or $X = R_{\ell/2} \cup \bigcup_{i=\ell/2+1}^{\ell} X_i$ when ℓ is even. In a stronger statement, any *maximum* mixed strong comma-free code is of this form.

References

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

[2] D. Bajić and J. Stojanović, *Distributed sequences and search process*, 2004 IEEE International Conference on Communications (Paris, France, 2004), IEEE, 2004, pp. 514-518.

[3] A. H. Ball and L. J. Cummings, *Extremal digraphs and comma-free codes*, Ars Combinatoria **1** (1976), no. 1, 239–251.

[4] ———, *The comma-free codes with words of length two*, Bull. Austral. Math. Soc. **14** (1976), no. 2, 249–258.

[5] S. Bilotta, E. Grazzini, E. Pergola, and R. Pinzani, *Avoiding cross-bifix-free binary words*, Acta Inform. **50** (2013), no. 3, 157–173.

[6] S. Bilotta, E. Pergola, and R. Pinzani, *A new approach to cross-bifix-free sets*, IEEE Trans. Inform. Theory **58** (2012), no. 6, 4058–4063. MR2924424

[7] S. R. Blackburn, *Non-overlapping codes*, IEEE Transactions on Information Theory **61** (2015), no. 9, 4890–4894.

[8] D. A. Cartwright, M. A. Cueto, and E. A. Tobis, *The maximum independent sets of de Bruijn graphs of diameter 3*, Electron. J. Combin. **18** (2011), no. 1, Paper 194, 18.

[9] Y. M. Chee, H. M. Kiah, P. Purkayastha, and C. Wang, *Cross-bifix-free codes within a constant factor of optimality*, IEEE Trans. Inform. Theory **59** (2013), no. 7, 4668–4674. MR3071351

[10] L. J. Cummings, *Comma-free codes and incidence algebras*, Combinatorial mathematics, IV (Proc. Fourth Australian Conf., Univ. Adelaide, Adelaide, 1975), Springer, Berlin, 1976, pp. 1–6, Lecture Notes in Math., Vol. 560.

[11] F. H. C. Crick, J. S. Griffith, and L. E. Orgel, *Codes without commas*, Proceedings of the National Academy of Sciences of the United States of America **43** (1957), no. 5, 416–421.

[12] W. Eastman, *On the construction of comma-free codes*, IEEE Transactions on Information Theory **11** (1965), no. 2, 263–267.

[13] E. Fimmel, C. J. Michel, F. Pirot, J.-S. Sereni, and L. Strüngmann, *Circular Mixed Codes*, Mathematical Biosciences **317** (2019), 108231.

[14] E. Fimmel, C. J. Michel, and L. Strüngmann, *n-nucleotide circular codes in graph theory*, Philosophical Transactions of the Royal Society A **374** (2016), no. 2063, 20150058.

[15] ———, *Diletter circular codes over finite alphabets*, Mathematical Biosciences **294** (2017), 120–129.

[16] ———, *Strong Comma-Free Codes in Genetic Information*, Bulletin of Mathematical Biology **79** (2017), no. 8, 1796–1819.

[17] E. Fimmel and L. Strüngmann, *On the hierarchy of trinucleotide n-circular codes and their corresponding amino acids*, Journal of Theoretical Biology **364** (2015), 113–120.

[18] ———, *Maximal dinucleotide comma-free codes*, Journal of Theoretical Biology **389** (2016), 206–213.

[19] ———, *Mathematical Fundamentals for the noise immunity of the genetic code*, BioSystems **164** (2018), 186–198.

[20] S. W. Golomb, B. Gordon, and L. R. Welch, *Comma-free codes*, Canadian Journal of Mathematics **10** (1958), 202–209.

[21] S. W. Golomb, L. R. Welch, and M. Delbrück, *Construction and properties of comma-free codes*, Biologiske Meddelelser, Kongelige Danske Videnskabernes Selskab **23**, 1–34.

[22] L. J. Guibas and A. M. Odlyzko, *Maximal prefix-synchronized codes*, SIAM J. Appl. Math. **35** (1978), no. 2, 401–418.

[23] V. I. Levenšteĭn, *Decoding automata which are invariant with respect to the initial state*, Problemy Kibernet. No. **12** (1964), 125–136 (Russian).

[24] V. N. Levenšteĭn, *The maximal number of words in codes without overlap*, Problemy Peredači Informacii **6** (1970), no. 4, 88–90 (Russian).

[25] V. I. Levenshtein, *Combinatorial problems motivated by comma-free codes*, J. Combin. Des. **12** (2004), no. 3, 184–196.

[26] C. J. Michel, G. Pirillo, and M. A. Pirillo, *Varieties of comma-free codes*, Comput. Math. Appl. **55** (2008), no. 5, 989–996.

[27] R. Scholtz, *Maximal and variable word-length comma-free codes*, IEEE Transactions on Information Theory **15** (1969), no. 2, 300–306.

[28] B. Tang, S. W. Golomb, and R. L. Graham, *A new result on comma-free codes of even word-length*, Canad. J. Math. **39** (1987), no. 3, 513–526.

[29] P. Turán, *Eine Extremalaufgabe aus der Graphentheorie*, Mat. Fiz. Lapok **48** (1941), 436–452 (Hungarian, with German summary).

Appendix A. Growth Function of Genetic Trinucleotide Circular Codes

We present the growth function of circular 3-letter codes over the alphabet $\mathcal{B}$ (of cardinality 4), presented in function of the number a of arcs in a longest directed path of the associated graph. The codes were generated using a computer. It seems exciting to obtain a mathematical explanation of the number of 3-letter circular genetic code of maximal size, which is 12 964 440.

TABLE 4. Growth function of 3-letter circular codes $X \subseteq \mathcal{B}^3$ (cardinality between 1 and 20) as a function of the maximal path length a (from 1 to 8) in their associated graph G_X, defined in Subsection 4.1.

$\lvert X\rvert$ \ a	1	2	3	4	5	6	7	8	**Total**
1	48	12	0	0	0	0	0	0	60
2	564	1092	48	0	0	0	0	0	1704
3	2432	23176	3720	1056	48	0	0	0	30432
4	4968	239040	82488	50196	4080	1344	48	0	382164
5	5424	1524636	894912	958344	109248	70560	3792	1296	3568212
6	3288	6635052	5711520	10066008	1455408	1477332	93840	65064	25507512
7	1080	20707380	23608200	66358032	11578248	16920696	1184928	1281216	141639780
8	168	47742486	67286520	295339356	60415008	120991116	9070416	13723032	614568102
9	8	82816624	138365616	929260512	218650464	580183752	45957504	91507728	2086742208
10	0	109358220	212231640	2131173360	569191680	1949610312	162487776	408593256	5542646244
11	0	110895036	248599344	3635098536	1092252720	4724611056	414758832	1276845600	11503061124
12	0	87031844	225759720	4668405744	1569961080	8412344832	781162896	2871001008	18615667124
13	0	53227980	160087992	4539916512	1705224984	11124273000	1099164288	4721590800	23403485556
14	0	25473732	88569264	3341064744	1402203888	10963159272	1160318208	5719845816	22700634924
15	0	9519912	37872240	1846581744	867844824	8016801504	914981088	5093921760	16787523072
16	0	2743080	12273168	753781272	397991256	4288163160	531158208	3292912176	9279022320
17	0	591864	2914992	220449432	131222040	1630269696	220422672	1502846352	3708717048
18	0	90420	479256	43730412	29429376	417392700	61906128	459071448	1012099740
19	0	8760	48912	5281272	4022160	64576488	10548336	84240864	168726792
20	0	408	2352	294312	252960	4566696	823920	7023792	12964440