AUSTRALASIAN JOURNAL OF COMBINATORICS
Volume **86(2)** (2023), Pages 233–270

Diletter and triletter comma-free codes over finite alphabets

Elena Fimmel

Institute of Mathematical Biology, Faculty for Computer Sciences
Mannheim University of Applied Sciences
68163 Mannheim, Germany
e.fimmel@hs-mannheim.de

Christian J. Michel François Pirot*

Theoretical Bioinformatics, ICube, C.N.R.S., University of Strasbourg
300 Boulevard Sébastien Brant
67400 Illkirch, France
c.michel@unistra.fr francois.pirot@loria.fr

Jean-Sébastien Sereni

Theoretical Bioinformatics, ICube, C.N.R.S., University of Strasbourg
300 Boulevard Sébastien Brant
67400 Illkirch, France
jean-sebastien.sereni@cnrs.fr

Lutz Strüngmann

Institute of Mathematical Biology, Faculty for Computer Sciences
Mannheim University of Applied Sciences
68163 Mannheim, Germany
l.struengmann@hs-mannheim.de

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

* Also at: LORIA (Orpailleur), C.N.R.S., Univ. of Lorraine and INRIA, Vandœuvre-lès-Nancy, France; and Dept. of Mathematics, Radboud University, Nijmegen, Netherlands.

ISSN: 2202-3518

such codes when all words have length 2, or 3. We first explain how different properties combine to obtain a closed-formula. We next develop an approach to tackle well-known sub-families of comma-free codes, such as self-complementary and (generalisations of) non-overlapping codes, for which the aforementioned properties do not hold anymore. 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 (over an alphabet of arbitrary cardinality) and the number of codes reaching the bound. We also provide a characterisation of non-overlapping triletter codes that are inclusion-wise maximal, which allows us to devise the number of such codes. We point out other applications of the method, notably to self-complementary codes, including the recently introduced mixed codes. Our approaches mix combinatorial and graph-theoretical arguments.

1 Introduction

A great number of DNA codes have been proposed in biology, for example the nucleosome positioning code [17], the histone code [37], the splicing code [5], the mRNA degradation site code [10], or the protein folding code [18], to name but a few. These genetic codes are primarily a mapping of a motif, or a set of motifs, with a high occurrence frequency to a biological function. However, there is no underlying mathematical structure, mainly because it does not exist. Our previous work on the genetic code (reviewed in [34, 25]) provides an interesting mathematical structure that allows us to develop not only theoretical results but also probability and statistical studies of genes in this framework, i.e. a return of the theory to the biological reality.

In today's genes, the coding of trinucleotides (3-letter words on the 4-letter alphabet $\mathcal{B} = \{A, C, G, T\}$) into amino acids (1-letter words on a 20-letter alphabet) relies on very complex molecular processes, involving in particular: (i) the messenger RNA by a transcription of DNA leading to a protein coding region bounded by regulatory sequences; (ii) the transfer RNA associated with its protein aminoacyl-tRNA synthetase; (iii) the ribosome with 3 ribosomal RNAs associated with over 50 ribosomal proteins; and (iv) the genetic code $\mathcal{B}^3 = \{A, C, G, T\}^3$ of cardinality 64 coding the amino acids constituting the proteins. In our coding context, the (universal) genetic code has three important properties: 61 trinucleotides code 20 amino acids (surjective map), one trinucleotide ATG among these 61 trinucleotides, called start codon, codes the amino acid methionine and also initiates the reading frame, and three trinucleotides $\{TAA, TAG, TGA\}$, called stop codons, terminate the reading frame without amino acid coding. In other words, the current genetic code has the property of amino acid coding and the property of starting and ending the reading frame. However, this property of reading frame retrieval is very weak, as it is limited to the boundaries of the genes, but it is fully functional because, as mentioned, several classes of RNAs and a large number of proteins are involved. The genetic code has evolved over about 4 billion years as putative fossilized microorganisms that are at least 3.8 billion and possibly 4.3 billion years old have been observed in ferruginous sedimentary rocks [15]. Thus, this weak property of reading frame retrieval in current genes may be the consequence of a very strong property of reading frame retrieval in primitive genes that would have existed with only one class of RNA and in the absence of proteins, or in the presence of only a few primitive proteins. Most likely, such a primitive genetic code would had a stronger mechanism for frame retrieval (called circularity). The trinucleotides in a such circular code allow both the retrieval of the reading frame and the coding of some amino acids, analogous to the ATG start codon in current genes. The stop codons would not have been necessary due to the short length of the primitive genes. Figure 2 in Appendix summarizes this evolutionary hypothesis of the genetic code based on a growing combinatorial hierarchy of trinucleotide circular codes that starts with the (strong) comma-free codes that have the highest ability to retrieve the reading frame and that will be in the focus of this paper.

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 [13] 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 [26] and of Golomb, Welch and Delbrück [27], 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 [14, 16, 29, 30, 33, 38, 39], which led 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, which are referred to as "strong comma-free codes" in the sequel and were also coined "strongly regular codes" or "non overlapping codes" or "cross-bifix-free codes", has also been the focus of several works [2, 9, 31, 32], dating back to 1964, in particular for their interest in automata theory and for frame synchronisation applications [3, 7, 8, 12, 28]. It is no surprise that such natural properties of codes were useful and studied in a variety of contexts, under different names. More information is found in the treatise on codes by Berstel, Perrin and Reutenauer [6, p. 281 and Proposition 7.2.14, p. 285].

While digraphs have been used to study diletter comma-free codes, either implicitly through their adjacency matrix [14] or explicitly [4], appropriate digraphs for comma-free codes with longer words seem less natural and harder to find. This was recently done [20] 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 [21, 22, 19, 20]. We pursue 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 [27] obtained a general upper bound on this size expressed using the Möbius function (see Theorem 3.1). Seven years later, this bound was shown to be attained whenever the length ℓ of the words is odd by Eastman [16], whose construction was subsequently simplified by Scholtz [38]. The situation when ℓ is even is less understood, and although it is known [26] 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 [26] 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 [27] 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 [11] demonstrated an injection of the maximum independent sets of the de Bruijn graph $B(n, 3)$ in 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 $n \geqslant 3$, which yields a closed formula to count the number of maximum comma-free triletter code. We then extend earlier results of Blackburn [9] 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 [9, Conjecture 1]: he provided a construction of strong comma-free ℓ-letter codes and posited that for every integer $\ell \geqslant 2$, there exists an integer n_0 such that the construction yields a maximum strong comma-free code over an alphabet of cardinality n whenever $n \geqslant n_0$.

The recent extension of the graph theoretical approach [20] unveils 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 2, while strong comma-free codes (aka. non-overlapping codes) are those for which the length is 1. In Subsection 4.1, we generalise results obtained by Ball and Cummings [4] 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 strong comma-free 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 since the topic raised interest in different fields.

2 Definitions and Notions

Let Σ be an arbitrary finite alphabet with $n := |\Sigma|$. For an integer $\ell \geqslant 2$, 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, $\cup_{\ell\geqslant 0}\Sigma^\ell$, and we define Σ^+ to be the collection of all finite and non-empty words with letters in Σ, that is, $\cup_{\ell\geqslant 1}\Sigma^\ell$.

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 $m \leqslant k$, 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 $k \in \mathbf{N}$;
- an *expandable (k-)circular ℓ-letter code* if it is 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 $|Y| \leqslant |X|$ whenever Y is an ℓ-letter (k-)circular (comma-free, strong comma-free) code over Σ.

A code that is not expandable is sometimes called *maximal* in the literature, as a short-hand for “inclusion-wise maximal”. Before giving some examples, we also point out that a comma-free code is automatically circular [35, 23].

Example 2.2 Let $\Sigma := \{0, 1, 2\}$ be the ternary alphabet. The set $X_0 := \{012, 220\}$ is a 3-letter code of size 2 over Σ. This code is not strong comma-free since 0 is both prefix and suffix. One sees however that it is comma-free, since the comma-freeness condition is exactly that $X^2 \cap \Sigma^+ X \Sigma^+ = \varnothing$. And indeed, we have $X^2 = \{012012, 012220, 220012, 220220\}$, no element of which can be written as an element in $\Sigma^+ X \Sigma^+$. This implies that X_0 is circular.

On the other hand, the 3-letter code $X_1 := \{012, 110, 122, 211\}$ is not 2-circular as the concatenation 012211 of two words from X_1, when read on a circle, admits a second decomposition in words from X_1, namely $122|110$.

Consider now the following 2-letter (or diletter) codes over Σ.

$$X_2 := \{01, 12, 20\}, \quad X_3 := \{01, 02, 12\}, \quad X_4 := \{01, 12\}.$$

Both X_2 and X_3 have size 3, and the code X_2 is not 3-circular as the concatenation 012012 of the three words in X_2 admits two different decompositions in words from X_2 when read on a circle, namely 01|20|12 and 12|01|20. The code X_2 is however 2-circular. The code X_3, on the other hand, is circular. Since $X_4 \subset X_3$, the code X_4 must also be circular, and is an expandable circular 3-letter code (over Σ). The code X_3 is a circular 3-letter code of maximal size (over Σ). Indeed $\Sigma^2 \setminus X_3 = \{00, 11, 22, 10, 20, 21\}$, and a circular diletter code cannot have any constant word (which forbids $00, 11, 22$), and the remaining three words all are shifts of a word in X_3, so adding any of them would immediately force the word to have more than one decomposition in words from X, (when read on a circle): for example, in $X_3 \cup \{10\}$ the word 10, on a circle, can also be read as 01, which belongs to $X_3 \cup \{10\}$. As we shall see very soon, this fact is general: an ℓ-letter code containing a word and one of its "cyclic permutations" cannot be 1-circular.

Remark 2.3 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 [31], then rediscovered and called cross-bifix-free codes [3] or strong circular codes [35], and recently redefined using graph theoretical models of the genetic code [22].

We state and prove the following observation, which is now well known, to illustrate the notions just introduced.

Observation 2.4 If X is a strong comma-free ℓ-letter code over a finite alphabet Σ, then X is comma-free.

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.

Observation 2.4 also explains why we keep the name "strong comma-free", which emphasises the relation to comma-free codes. In addition we shall introduce p-comma-free codes, which correspond to strong comma-free codes when $p = 1$ and to comma-free codes when $p = 2$.

Two symmetric groups play an important role in the context of circular codes [25]. The first one acts on the elements of the alphabet Σ and is defined as

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

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 component-wise 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

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

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 3-letter word $b_1b_2b_3$ into the 3-letter 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

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

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

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

Definition 2.5 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 2.6 The Möbius function $\mu\colon \mathbf{N} \to \{-1, 0, 1\}$ is defined as

$$\mu(n) := \begin{cases} 1 & \text{if } n = 1, \\ (-1)^m & \text{if } n = q_1 \cdots q_m \text{ where } q_1,\dots,q_m \text{ are distinct primes,} \\ 0 & \text{otherwise.} \end{cases}$$

Definition 2.7 For every positive integer ℓ and every integer $i \in \{1, \ldots, \ell\}$, we define the *projection on the ith coordinate* $\mathrm{p}_i \colon \Sigma^\ell \to \Sigma$ by $\mathrm{p}_i(x_1 \cdots x_\ell) = x_i$. The projections on two coordinates $\mathrm{p}_{ij} \colon \Sigma^\ell \to \Sigma^2$ 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 [26] 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 [26, 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 $\ell \geqslant 2$. 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 $n \geqslant 2$. The cardinality $S_{\max}(n, \ell)$ of a circular ℓ-letter code of maximal size over Σ and the cardinality $S^1_{\max}(n, \ell)$ of a 1-circular ℓ-letter code of maximal size over Σ are both equal to*

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

Proof: As explained earlier, we have

$$S_{\max}(n, \ell) \leqslant S^1_{\max}(n, \ell) \leqslant \frac{1}{\ell}\sum_{d|\ell} \mu\left(\frac{\ell}{d}\right) n^d.$$

To conclude the proof, it is thus enough to exhibit a circular code over Σ of size $\sum_{d|\ell} \mu\left(\frac{\ell}{d}\right) n^d$.

To this end, let $<$ be a total order on Σ, and consider the ℓ-letter code X over Σ constructed in the following way: X contains the minimum element, according to $<$, of each complete cyclic equivalence class induced by $\mathcal{A}_\ell$. By the definition, $|X| = \sum_{d|\ell} \mu\left(\frac{\ell}{d}\right) n^d$, and it thus suffices to show that X is circular.

Suppose on the contrary that there exist two circular decompositions of some word $x_1 x_2 \dots x_{r\ell} \in \Sigma^{r\ell}$, 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 $k \leqslant \ell - k$. 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

$$a_1 b_1 | \dots | a_r b_r \quad \text{and} \quad b_1 a_2 | \dots | b_r a_1,$$

and hence, setting $a_{r+1} \coloneqq a_1$ for convenience, one sees that $a_i b_i \in X$ and $b_i a_{i+1} \in X$ 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) $a_i b_i < b_i a_i$ since $a_i b_i$ and $b_i a_i$ are in the same equivalence class, $a_i b_i \in X$ 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 $k \leqslant \ell - k$, we know that $|a_i| \leqslant |b_i|$ for every $i \in \{1, \dots, k\}$. 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 $a_i \leqslant b'_i$ by (1), and that $b'_i \leqslant a_{i+1}$ by (2). It follows that $a_i \leqslant a_{i+1}$, and hence $a_i < a_{i+1}$ by (3). Since this is valid for every $i \in \{1, \dots, r+1\}$, 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

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

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 $S^1_{\max}(4,p) = \frac{1}{p}(4^p - 4)$. This 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*

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

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

$k \backslash \ell$	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.

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 of maximal size 20 containing the words AAG, TAA, GGT and GTT. Therefore, the word $TAAGGT$ admits two circular decompositions into words in X, namely $TAA|GGT$ and $T|AAG|GT$.
- The example given above also shows that the construction provided in the proof of Theorem 3.1 does not cover all 1-circular ℓ-letter codes of maximal size: as one can see, there is no order on $\{A, G, T\}$ such that each of the first 3 words above, namely AAG, TAA and GGT, 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 either. 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 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 Comma-Free Di- and Triletter 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 $n \geqslant 3$ 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 [20]. 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 [20, 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 $p \leqslant 2$. 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 a 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, which generalises an earlier result of Ball and Cummings [4], 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 $b > a$.

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 [22, 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,

$$S_1(n,2) = \begin{cases} \frac{1}{2}\left(1-\frac{1}{2}\right) \times n^2 = \frac{1}{4}\cdot n^2 & \text{if } n \text{ is even,} \\ \frac{1}{2}\left(1-\frac{1}{2}\right) \times (n^2-1) = \frac{1}{4}\cdot(n-1)(n+1) & \text{if } n \text{ is odd,} \end{cases}$$

and moreover

$$N_1(n,2) = \begin{cases} \binom{2}{0}\frac{n!}{\left(\frac{n}{2}!\right)^2 \times 1^0} = \binom{n}{n/2} & \text{if } n \text{ is even,} \\ \binom{2}{1}\frac{n!}{\left(\frac{n-1}{2}!\right)^2 \times \left(\frac{n+1}{2}\right)^1} = 2\binom{n}{\frac{n+1}{2}} & \text{if } n \text{ is odd.} \end{cases}$$

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. 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.

Corollary 4.3 *If the length of every directed path in a digraph D is at most p, then (the underlying undirected graph of) D does not contain K_{p+2} 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

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

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 $k+1 \leqslant n$. 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.

Theorem 4.4 (Turán [40]) *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 $G = T_k(n)$.*

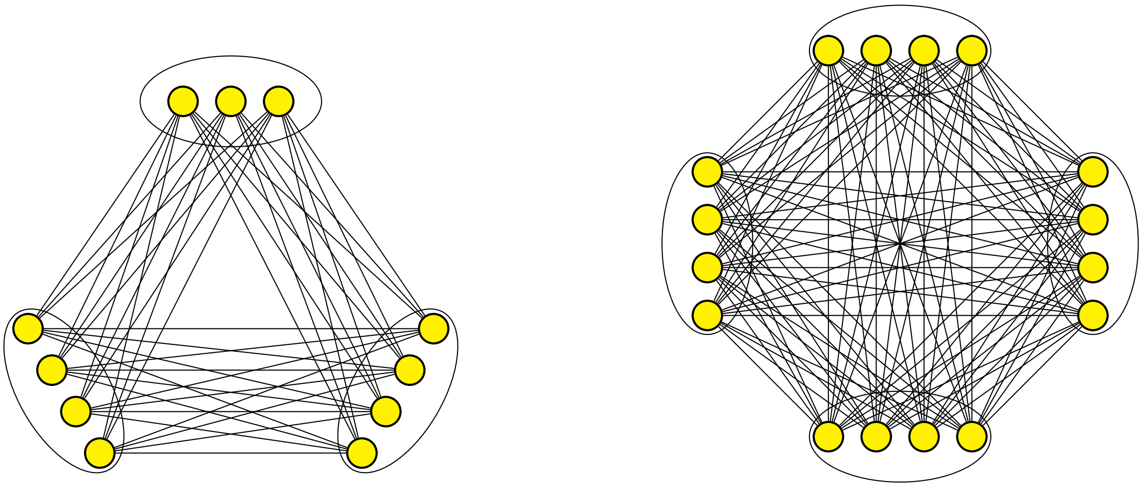

Figure 1: The Turán graph $T_3(11)$ on the left and the Turán graph $T_4(16)$ on the right.

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 $S_p(n,2) = t_{p+1}(n)$. 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 Corollary 4.3 that (the underlying undirected graph of) G_X does not contain the complete graph on $p+2$ 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 $t_{p+1}(n)$, and hence $S_p(n,2) \leqslant t_{p+1}(n)$.

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

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 of maximal size 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 $(X_1, \ldots, X_{p+1})$ 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 $f\colon \Sigma \to (X_1, \ldots, X_{p+1})$ such that $|f^{-1}(X_i)| = |X_i|$ for every $i \in \{1, \ldots, p+1\}$.

To compute the number of such functions f, recall that $n = (p+1)m + r$

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 $n \in \{3, \dots, 15\}$, as obtained using computers [21] and confirmed by the general formulæ of Theorem 4.1.

where $m = \lfloor n/(p+1) \rfloor$ and $r \in \{0, \dots, p\}$. Set $M := \lceil n/(p+1) \rceil$, so $M = m$ if $r = 0$ and $M = m+1$ otherwise. Using these 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 [24, 20]; we have

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

Remark 4.6 Using computers, all maximum comma-free diletter codes on alphabets of cardinalities $n \in \{2, \dots, 15\}$ have been generated [21]: Theorem 4.1 confirms that the computer programs used were correct, and provides the sought value for every integer $n \geqslant 3$. (See Table 2.)

4.2 Triletter Codes

We now turn our attention to comma-free 3-letter 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 3-letter 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 [26] obtained an upper bound on the maximal size of a 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 [16]. We focus on 3-letter codes: $S(n,3) = \frac{n(n^2-1)}{3}$. Let Σ be an alphabet of cardinality $n \geqslant 3$. While Golomb, Welch and Delbrück [27] described a method to construct all comma-free 3-letter codes of maximal size 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 3-letter 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 [11] 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 [11, Theorem 5.1] that maximum independent sets of $B(n,3)$ inject into the collection of maximum comma-free 3-letter 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 3-letter 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 [27], one realises that $n = 2$ is the only exceptional case: as soon as $n \geqslant 3$, maximum comma-free 3-letter codes over Σ are in bijection with maximum independent sets in the de Bruijn graph $B(n,3)$. We thus obtain the following statement.

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

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

Proof: Let $n \geqslant 3$ 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 [27, 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 [27, 26, 36] numbers $S(4,3) = \frac{4(4^2-1)}{3} = 20$ for the size and $N(4,3) = 4!\left\lceil\frac{(1+\sqrt{2})^4}{2}\right\rceil = 24 \cdot 17 = 408$ for the number of comma-free 3-letter 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 3-letter codes of maximal size, which is 8. Letting Σ be $\{0, 1, 2\}$, these 42 codes are listed below.

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

We now develop an approach 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 [27] do not hold anymore. We consider different families of codes along these lines in the next subsections.

4.3 Self-Complementary Triletter Codes

We consider comma-free 3-letter 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 $N^\pi_{\mathrm{cf}}(n,\ell)$ 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 4.10 *Let Σ be an alphabet of even cardinality 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 $n \geqslant 6$ 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) $N^\pi_{\mathrm{cf}}(2,3) = 2$, $N^\pi_{\mathrm{cf}}(4,3) = 4$, $N^\pi_{\mathrm{cf}}(6,3) = 224$ and if $n \geqslant 8$ 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 4.10, 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 π-self-complementarity, the comma-free 3-letter codes X we consider are also required to satisfy that $\mathrm{p}_{12}(X) \cap \mathrm{p}_{23}(X) = \varnothing$. Notice that this last condition implies that the code is empty if the alphabet has cardinality two.

Proposition 4.11 *Let Σ be an alphabet of even cardinality n and let $\pi \in S_\Sigma$ be an involutory transformation with no fixed point. If X is a π-self-complementary comma-free 3-letter code such that $\mathrm{p}_{12}(X) \cap \mathrm{p}_{23}(X) = \varnothing$, then*

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

Furthermore, there is equality if and only if, setting $n' := n/2 - 1$, one can write

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

with $X_i := \{\alpha_{i+1}, \pi(\alpha_{i+1}), \ldots, \alpha_{n'}, \pi(\alpha_{n'})\}$ and $\Sigma = \{\alpha_0, \pi(\alpha_0), \ldots, \alpha_{n'}, \pi(\alpha_{n'})\}$. In particular, 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 $\Sigma = \{a, \pi(a)\}$. (In this case, only the empty code satisfies all requirements.) Now let $n \geqslant 4$ and let X be a π-self-complementary comma-free 3-letter code of maximal size over an alphabet Σ of cardinality n, such that $\mathrm{p}_{12}(X) \cap \mathrm{p}_{23}(X) = \varnothing$. 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 $\ell_a := |L_a|$ and $r_a := |R_a|$ for each $a \in \Sigma$. It follows that X is contained in $\cup_{a=0}^{n-1} L_a \cdot a \cdot R_a$. Notice that $L_a = \overleftarrow{\pi(R_{n-1-a})}$ and $R_a = \overleftarrow{\pi(L_{n-1-a})}$ 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 $\ell_0 \geqslant \max\{\ell_a, r_a\}$ 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 $\mathrm{p}_{12}(X) \cap \mathrm{p}_{23}(X) = \varnothing$. 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 $\ell_0 \geqslant r_0$, 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 the letter 0. Our strategy is to consider the set W composed of all words in X that contain the letter 0, and show that $|W| \leqslant n^2 - 3n + 2$. It then follows by π-self-complementarity that the number N of words in X that contain 0 or $n - 1$ is at most $2\,|W| \leqslant 2n^2 - 6n + 4$. Now, deleting from X all such words yields a π-self-complementary comma-free 3-letter code X' over the alphabet $\{1, \ldots, n-2\}$ such that $\mathrm{p}_{12}(X') \cap \mathrm{p}_{23}(X') = \varnothing$. We know by induction that $|X'| \leqslant (n-2)^3/3 - (n-2)^2/2 - (n-2)/3$, and we know the shape of X' if there is equality. We would therefore deduce that

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

Furthermore, there would be equality only if X' has maximal size and $N = 2n^2 - 6n + 4$ (in particular, note for (much) later on that then no word in X contains both 0 and $n - 1$, for otherwise $N < 2\,|W|$).

So let us bound from above the size of W, the set composed of all words in X containing the letter 0. Because $\mathrm{p}_{12}(X) \cap \mathrm{p}_{23}(X) = \varnothing$, 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 $W_0 := \{w \in W : \mathrm{p}_2(w) = 0\}$, and $W_1 := W \setminus W_0$. Using the notation previously introduced, we have $|W_0| \leqslant \ell_0 r_0 \leqslant (n-1-m)(n-2-k)$.

To evaluate the size of W_1, set $M_0 := \{1, \dots, n-2\} \setminus L_0$, so $m_0 := |M_0| \in \{m-1, m\}$. Further, if $a \in \{1, \dots, n-2\}$, then $0 \in R_a$ only if $a \in M_0$. So, for $a \in \{1, \dots, n-1\}$, the number of words in W_1 ending by $a0$ is 0 if $a \notin M_0$ or $0 \notin R_a$, and at most ℓ_a otherwise. In particular, the number of words in W_1 ending by 0 cannot exceed $\sum_{a \in M_0} \ell_a$. Similarly, setting $K_0 := \{1, \dots, n-2\} \setminus R_0$, we have $k_0 := |K_0| \in \{k, k+1\}$ and if $a \in \{1, \dots, n-2\}$, then $0 \in L_a$ only if $a \in K_0$. So, for $a \in \{1, \dots, n-1\}$, the number of words in W_1 starting by $0a$ is 0 if $a \notin K_0$ or $0 \notin L_a$, and at most r_a otherwise. In particular, the number of words in W_1 starting by 0 cannot exceed $\sum_{a \in K_0} r_a$. Finally, 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 $m_0 + k_0 \leqslant m + k$.

We now bound the number of words in W_1 by bounding, for each $a \in \{1, \dots, n-2\}$, the number of words w in W_1 with $\mathrm{p}_2(w) = a$. 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,

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

It follows that

$$\begin{aligned} |W| = |W_0| + |W_1| &\leqslant (n-1-m)(n-2-k) + (m+k)(n-1-m) \\ &= (n-1-m)(n-2+m) \\ &= n^2 - 3n + 2 - m(m-1). \end{aligned}$$

As a result, $|W| \leqslant n^2 - 3n + 2$ 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.

We prove that if X is maximum, then the number of words starting with $0a$, for each $a \in \{1, \dots, n-2\}$ is at most $n-3$. Indeed, either this number is 0, or $a \in K_0$ and $0 \in L_a$. In this case, we know that $0 \notin R_a$. Recall also that R_a never contains $n-1-a$. So if in addition $n-1 \notin R_a$, then $r_a \leqslant n-3$, which implies our statement. In addition, if $n-1 \in R_a$, then the number of words in X starting by $0a$

is at most $r_a - 1$, as otherwise X would contain the word $0a(n-1)$, which prevents X from being maximum, as reported (much) earlier. This proves our statement.

Now, if $m = 0$, then $L_0 = \{0, \ldots, n-2\}$ and hence $0 \notin \bigcup_{a=1}^{n-1} R_a$. Notice that $k_0 = k$ because $m_0 = m$ and $k_0 + m_0 = k + m$. Combining these with our statement above yields that

$$\begin{aligned} n^2 - 3n + 2 = |W| &\leqslant |W_0| + \sum_{\substack{a \in M_0 \\ 0 \in R_a}} \ell_a + \sum_{\substack{a \in K_0 \\ 0 \in L_a}} (n-3) \\ &\leqslant (n-1-m)(n-2-k) + m(n-2-k) + k_0(n-3) \\ &\leqslant (n-1)(n-2-k) + k(n-3) \\ &= n^2 - 3n + 2 - k(n-1) + k(n-3), \end{aligned}$$

which implies that $k = 0$, so $k_0 = 0$ and $K_0 = \varnothing$, and hence $R_0 = \{1, \ldots, n-2\}$, which completes the proof in this case.

We end the proof by showing that if $m = 1$ then X is not of maximal size. Suppose indeed that $m = 1$, that is, $L_0 = \{0, \ldots, n-2\} \setminus \{i\}$ with $i \in \{0, \ldots, n-2\}$. In this case, $|W_0| \leqslant (n-2)(n-2-k)$.

If $i = 0$, then $M_0 = \varnothing$, so no word in X ends by $a0$ with $a \in \{1, \ldots, n-2\}$ (and thus also for $a = n-1$). Recall also that the number of words in X starting with $0a$, for each $a \in K_0$, is at most $n-3$ (and 0 if $a \notin K_0$). It follows that

$$\begin{aligned} n^2 - 3n + 2 = |W| &\leqslant |W_0| + \sum_{\substack{a \in M_0 \\ 0 \in R_a}} \ell_a + k_0(n-3) \\ &\leqslant (n-2)(n-2-k) + (k+1)(n-3) \\ &\leqslant (n-2)^2 + (n-3) \\ &= n^2 - 3n + 1, \end{aligned}$$

a contradiction.

Therefore, $L_0 = \{0, \ldots, n-2\} \setminus \{i\}$ with $i \in \{1, \ldots, n-2\}$. In particular, if $a \in \{1, \ldots, n-2\}$ and $0 \in R_a$, then $a = i$. Since $0 \notin R_{n-1-i}$, the π-self-complementarity of X implies that $n-1 \notin L_i$. Moreover, $0 \in L_0$ and hence $0 \notin R_0$, which implies that $m_0 = m$ and $k_0 = k$. So

$$\begin{aligned} n^2 - 3n + 2 = |W| &\leqslant |W_0| + \sum_{\substack{a \in M_0 \\ 0 \in R_a}} \ell_a + k_0(n-3) \\ &\leqslant (n-2)(n-2-k) + \ell_i + k(n-3) \\ &= (n-2)^2 + \ell_0 - k \\ &= n^2 - 3n + 2 - k, \end{aligned}$$

which implies that $\ell_i = n-2$ and $k = 0$. In particular, recalling that $0 \notin R_0$, we deduce that $R_0 = \{1, \ldots, n-2\}$. This implies that $i \in R_0$ and therefore $0i \in \mathrm{p}_{23}(X)$. Observe now that $n-1 \notin L_i$, because $n-1-i \in \{1, \ldots, n-2\} \subset L_0$, and hence $0 \notin$

R_{n-1-i} (as $\mathrm{p}_{12}(X) \cap \mathrm{p}_{23}(X) = \varnothing$), which implies that $n-1 \notin L_i$ because X is π-self-complementary. Consequently, since $\ell_i = n-2$ we have $L_i = \{0, \ldots, n-2\} \setminus \{n-1-i\}$, and therefore $0i\, \mathrm{p}_{12}(X)$, which now contradicts that $\mathrm{p}_{12}(X) \cap \mathrm{p}_{23}(X) = \varnothing$. This concludes the proof. □

We are now in a position to establish Theorem 4.10.

Proof of Theorem 4.10: Fix an alphabet Σ of even cardinality n. Let X be a maximum π-self-complementary comma-free 3-letter code over Σ. Let (L, M, R) be a partition of Σ such that

- $M \coloneqq \mathrm{p}_1(X) \cap \mathrm{p}_3(X)$;
- $\mathrm{p}_1(X) \setminus M \subseteq L$; and
- $\mathrm{p}_3(X) \setminus M \subseteq R$.

Because X is π-self-complementary, $\overleftarrow{\pi(L)} = R$, and $\overleftarrow{\pi(M)} = M$. Set $X' \coloneqq LM \cup MR \cup LR$. We observe that $\mathrm{p}_{12}(X) \cap \mathrm{p}_{23}(X) \subseteq X' \cup MM$. Indeed, if $w_1w_2w_3 \in X$ then $w_1 \in \mathrm{p}_1(X) \subseteq L \cup M$ and $w_3 \in \mathrm{p}_3(X) \subseteq M \cup R$, and hence

$$\begin{aligned} &\mathrm{p}_{12}(X) \subseteq LL \cup LM \cup LR \cup ML \cup MM \cup MR \quad \text{and} \\ &\mathrm{p}_{23}(X) \subseteq LM \cup MM \cup RM \cup LR \cup MR \cup RR. \end{aligned}$$

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

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

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

$$X \subseteq MMM \cup L(M_2 \cup R_2) \cup (L_2 \cup M_2)R \cup L_2M \cup MR_2. \tag{4.1}$$

To establish (4.1), we shall make several cases so as to ease the checking. Let $w = w_1w_2w_3$ be an element of X. In particular, $w_i \in \mathrm{p}_i(X)$ and hence $w_1 \in L \cup M$ and $w_3 \in M \cup R$. Since $w_1w_2 \in \mathrm{p}_{12}(X)$, we know that if $w_1w_2 \in X'$, then $w_1w_2 \in L_2 \cup M_2$. Similarly, if $w_2w_3 \in X'$, then $w_2w_3 \in M_2 \cup R_2$. 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 $w_2w_3 \in LM \cup LR \subseteq X'$, and hence $w_2w_3 \in M_2 \cup R_2$. Consequently, if $w_1 \in L$ then $w \in L(M_2 \cup R_2)$. 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, $w_1w_2 \in \mathrm{p}_{12}(X)\backslash MM$ and therefore $w_1w_2 \in L_2 \cup M_2$.

   (a) If $w_1w_2 \in L_2$ then $w \in L_2(M \cup R)$.

   (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 (4.1).

We bound the size of X by bounding the size of the right side of (4.1) 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 3-letter code over M, with the additional property that $\mathrm{p}_{12}(X[M]) \cap \mathrm{p}_{23}(X[M]) = \varnothing$. 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, writing $m = |M|$, Proposition 4.11 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 using (4.1) 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

$$\frac{m^3}{3} - \frac{m^2}{2} - \frac{m}{3} + \frac{(n^2-m^2)(n+3m)}{8} + \frac{(n-3m)}{2}|M_2|. \tag{4.2}$$

Let us maximise (4.2). We consider two cases regarding whether $(n-3m) \geqslant 0$, i.e. $m \leqslant n/3$, or $(n-3m) < 0$, i.e. $m > n/3$.

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

$$\frac{m^3}{3} - \frac{m^2}{2} - \frac{m}{3} + \frac{(n-m)^2(n+3m)}{4}, \tag{4.3}$$

  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, (4.3) for $m \leqslant n/3$ attains its maximum value when $m = \frac{5n}{13} - \frac{2}{39}\sqrt{27n^2+45n+48} + \frac{2}{13}$, the maximum value being

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

If $n \geqslant 8$, then $\sqrt{27+45/n+48/n^2} \leqslant \sqrt{27+45/8+3/4}$, 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 (4.3) 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 $LM \cup MR = L_2 \cup R_2$. 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 $LM \cup MR$ must be such that $\overleftarrow{\pi(L_2)} = R_2$ and $LR_2 \cap L_2R = \varnothing$. So for each $x \in M$, either $Lx \cup xR \subseteq L_2$, or $Lx \cup xR \subseteq R_2$, or $Lx \subseteq R_2, xR \subseteq L_2$. Note that this determines the choice of $L\pi(x) \cup \pi(x)R$, so there are 3 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 4.11 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 $n \geqslant 8$, and some of the codes when $n = 6$.

- When $m \leqslant n/3$, a code of maximal size is produced only when $m = 0$ and $n \in \{2, 4, 6\}$. Because then $M = \varnothing$, the code is $LLR \cup LRR$ 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 2 codes when $n = 2$, and 4 codes when $n = 4$, and 8 codes when $n = 6$.

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

$$8 + 6^{6/2-1}\binom{6}{2}! = 224.$$

□

Remark 4.12 Theorem 4.10 for $n = 4$ provides a theoretical explanation of the well-known fact [36] that the maximal size of self-complementary comma-free genetic codes is 16.

5 Strong Comma-Free Di- and Triletter Codes

We define $N_{\text{scf}}(n, \ell)$ to be the number of strong comma-free ℓ-letter codes of maximal size over an alphabet Σ with n letters and we let $S_{\text{scf}}(n, \ell)$ be their size. Similarly, we define $N_{\text{nescf}}(n, \ell)$ to be the number of non-expandable strong comma-free ℓ-letter codes over Σ. Our goal in this section is to study these parameters for $\ell \in \{2, 3\}$, maybe laying the foundations leading to general formulæ for arbitrary values of ℓ. The approach also applies to self-complementary codes.

5.1 Diletter Codes

The situation of strong comma-free diletter codes, i.e. $\ell = 2$, of maximal size, is actually a particular case of Theorem 4.1. Indeed, by the definition a diletter code X is comma-free if and only if $\mathrm{p}_1(X) \cap \mathrm{p}_2(X) = \varnothing$. Therefore, a strong comma-free diletter code over an alphabet Σ corresponds exactly to two choices: first that of an ordered pair (P_1, P_2) of disjoint non-empty subsets of Σ, and next that of a subset of P_1P_2, which could be the whole set. It is therefore of maximal size if and only if (as expected because of Theorem 4.1 when $p = 1$) the two sets have size $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$, and the code is the whole set P_1P_2.

We also readily deduce when a strong comma-free diletter codes is not expandable, that is inclusion-wise maximal. Since P_1P_2 is a strong comma-free diletter code for every pair of disjoint subsets P_1 and P_2 of Σ, we infer that non-expandable comma-free diletter codes over Σ are exactly the codes P_1P_2 where (P_1, P_2) is a partition of Σ into non-empty parts. Writing $n = |\Sigma|$, the number of ordered pairs (P_1, P_2) partitioning Σ into non-empty parts is $2^n - 2$, and therefore the number of non-expandable strong comma-free diletter codes over an alphabet of size n is $2^n - 2$. We thus have proved the following statement.

Theorem 5.1 *For each integer $n \geqslant 2$, the set of non-expandable strong comma-free diletter codes over an alphabet Σ of size n corresponds exactly to the set of partitions of Σ into two non-empty parts, and*

$$N_{\text{nescf}}(n, 2) = 2^n - 2.$$

For comparison, Table 3 shows the calculated numbers of strong comma-free diletter codes of maximal size, and of non-expandable strong comma-free diletter codes, over alphabets of cardinalities in $\{2, \ldots, 15\}$, computed using Theorems 4.1 and 5.1. We note that, using the formula provided by Theorem 4.1 (with $p = 1$), one easily sees that $N_{\text{scf}}(2k, 2) = N_{\text{scf}}(2k - 1, 2)$ for every $k \in \mathbf{N}$.

n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$N_{\text{scf}}(n, 2)$	2	6	6	20	20	70	70	252	252	924	924	3432	3432	12870
$N_{\text{nescf}}(n, 2)$	2	6	14	30	62	126	254	510	1022	2046	4094	8190	16382	32766

Table 3: The numbers of strong comma-free diletter codes of maximal size, and of non-expandable strong comma-free diletter codes, over an alphabet of cardinality n for $n \in \{2, \ldots, 15\}$.

We end this section by noticing that all strong comma-free diletter codes are π-self-complementary for a well-chosen bijection $\pi \in S_\Sigma$, specifically any bijection such that $\pi(\mathrm{p}_1(X)) = \mathrm{p}_2(X)$.

5.2 Triletter Codes

We aim now to characterise non-expandable strong comma-free 3-letter codes. It will in particular allow us to count them, and also to count strong comma-free 3-letter codes of maximal size, which form a subset of these.

To this end, we introduce the following notation.

Definition 5.2 *Fix an alphabet Σ. We let $\mathscr{P}_3$ be the set of all 3-letter codes $X_3 = L_1R_2 \cup L_2R_1$, where*

(1) (L_1, R_1) is a partition of Σ into two non-empty *parts; and*

(2) (L_2, R_2) is a partition of $X_2 \coloneqq L_1R_1$.

The sequences from Definition 5.2 allow us to capture the structure of all non-expandable strong comma-free 3-letter code, and to count them. We start with the following statement, recalling the projections defined in Definition 2.7.

Theorem 5.3 *For every alphabet Σ, the set of all non-expandable strong comma-free 3-letter codes is exactly the set $\mathscr{P}_3$.*

Proof: We prove that every element of $\mathscr{P}_3$ is itself a non-expandable strong comma-free 3-letter code. Let $X_3 = L_1R_2 \cup L_2R_1 \in \mathscr{P}_3$. First, we show that X_3 is strong comma-free. Indeed, on the one hand, $\mathrm{p}_1(X_3) \cap \mathrm{p}_3(X_3) = L_1 \cap R_1 = \varnothing$ by the definition of $\mathscr{P}_3$. On the other hand, $\mathrm{p}_{12}(X_3) \subseteq (L_1L_1) \cup L_2$ whereas $\mathrm{p}_{23}(X_3) = R_2 \cup (R_1R_1)$, which shows that $\mathrm{p}_{12}(X_3) \cap \mathrm{p}_{23}(X_3) = \varnothing$ as $L_i \cap R_i = \varnothing$ for $i \in \{1, 2\}$ and $L_2, R_2 \subseteq L_1R_1$ by definition.

Now let $w = w_1w_2w_3 \in \Sigma^3 \backslash X_3$. We want to show that $X_3 \cup \{w\}$ is not strong comma-free. Suppose that $w_1 \in R_1$. We prove that X_3 contains a word ending by w_1, which thus prevents $X_3 \cup \{w\}$ from being strong comma-free. Indeed, if $L_2 \neq \varnothing$, then X_3 contains a word xyw_1 with $xy \in L_2$. We deduce that $L_2 = \varnothing$ and hence $R_2 = X_2 = L_1R_1$. Since $L_1 \neq \varnothing$ by definition, let $z \in L_1$. Then $zzw_1 \in L_1L_1R_1 = L_1R_2 = X_3$. We thus conclude that $w_1 \notin R_1$, and hence $w_1 \in L_1$.

By symmetry of the arguments used, we also infer that $w_3 \in R_1$. Consequently, $w \in L_1(L_1 \cup R_1)R_1$. This implies that $w_2w_3 \notin R_2$ and $w_1w_2 \notin L_2$, because $w \notin X_3 = L_1R_2 \cup L_2R_1$. Now, if $w_2 \in L_1$ then $w_2w_3 \in L_1R_1 \setminus R_2 = L_2$. Since $R_1 \neq \varnothing$, for $x \in R_1$ we have $w_2w_3x \in L_2R_1 \subseteq X_3$, and therefore a prefix of a word of X_3 is a suffix of w, so $X_3 \cup \{w\}$ is not strong comma-free. Similarly, if $w_2 \in R_1$ then we infer that $w_1w_2 \in R_2$, and since L_1 is not empty we know that X_3 contains a word ending by w_1w_2, which prevents $W_3 \cup \{w\}$ from being strong comma-free. We thus have proved that X_3 is a strong comma-free 3-letter code that is not expandable.

Conversely, let X be a non-expandable strong comma-free 3-letter code over Σ. We build a code $X_3 = L_1R_2 \cup L_2R_1 \in \mathscr{P}_3$ and prove that $X = X_3$.

- We let (L_1, R_1) be a partition of Σ such that $\mathrm{p}_1(X) \subseteq L_1$ and $\mathrm{p}_2(X) \subseteq R_1$. We set $X_2 = L_1R_1$.
- We let (L_2, R_2) be a partition of X_2 such that $L_1 = X_2 \cap \mathrm{p}_{12}(X)$.

Observe that this construction is possible because the code X itself is strong comma-free, and this also implies that R_1 contains $X_2 \cap \mathrm{p}_{23}(X)$. Furthermore, by the definition, the code $X_3 = L_1R_2 \cup L_2R_1$ belongs to $\mathscr{P}_3$, and it follows from our previous considerations that X_3 is in particular strong comma-free. Therefore, to prove that $X = X_3$, it suffices to prove that $X \subseteq X_3$, since X is non-expandable.

Let $w = w_1w_2w_3 \in X$. It follows from the definitions that $w_1 \in L_1$ and $w_3 \in R_1$. If $w_1w_2 \in X_2 = L_1R_1$, then $w_1w_2 \in X_2 \cap \mathrm{p}_{12}(X)$ and hence $w_1w_2 \in L_2$. Consequently, $w_1w_2w_3 \in L_2R_1 \subseteq X_3$. If $w_1w_2 \notin X_2$, then $w_2 \in \Sigma \backslash R_1 = L_1$. It follows that $w_2w_3 \in L_1R_1 \cap \mathrm{p}_{23}(X) = X_2 \cap \mathrm{p}_{23}(X)$, so $w_2w_3 \in R_2$. Therefore, $w_1w_2w_3 \in L_1R_2 \subseteq X_3$, which concludes the proof. □

We now provide an example, and also illustrate the fact that Theorem 5.3 does not readily generalise to non-expandable strong comma-free ℓ-letter codes with $\ell \geqslant 4$.

Example 5.4 Let us consider the ternary alphabet $\Sigma = \{0, 1, 2\}$, and the circular strong comma-free code $X = \{001, 021, 201, 221\}$, which is indeed of maximal size. Following the notation in the proof of Theorem 5.3, one has $L_1 = \{0, 2\}$ and $R_1 = \{1\}$. By the 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 $L_2 = \varnothing$ and hence $R_2 = \{01, 21\}$. As expected, $X_3 = L_1R_2 \cup L_2R_1 = X$.

Example 5.5 A straightforward generalisation of Theorem 5.3 fails for all alphabet cardinalities larger than 3. Let us consider for instance the binary alphabet $\Sigma =$

$\{0,1\}$, and set $(L_1, R_1) = (\{0\},\{1\})$. We define the partition (L_2, R_2) of $X_2 = L_1R_1 = \{01\}$ by $L_2 = \varnothing$ and $R_2 = X_2$. Next, $X_3 = L_1R_2 \cup L_2R_1 = \{001\}$ is partitioned by $L_3 = \varnothing$ and $R_3 = X_3$. So $X_4 = L_1R_3 \cup L_2R_2 \cup L_3R_1 = \{0001\}$. Choosing now $L_4 = X_4$ and $R_4 = \varnothing$, we obtain $X_5 = L_1R_4 \cup L_2R_3 \cup L_3R_2 \cup L_4R_1 = \{00011\}$, which is an expandable strong comma-free code, because $\{00011, 01011\}$ is strong comma-free.

Regarding 4-letter codes, let us first note that the maximal size of a strong-comma free 4-code over the binary alphabet being 1, any binary 4-letter code either is not strong comma-free, or is maximum, and hence not expandable. The situation changes already over the ternary alphabet $\Sigma = \{0,1,2\}$. Let $L_1 = \{0,1\}$ and $R_1 = \{2\}$. For $i \geqslant 2$, let $L_i = X_i$ (and thus $R_i = \varnothing$) if i is even, whereas $R_i = X_i$ (and thus $L_i = \varnothing$) if i is odd. Then $L_2 = X_2 = \{02, 12\}$, next $R_3 = X_3 = \{022, 122\}$ and then $X_4 = \{0022, 0122, 1022, 1122\}$. However, X_4 is expandable, since $X_4 \cup \{0212\}$ is strong comma-free.

From Theorem 5.3 it follows a closed formula for the number of different non-expandable strong comma-free 3-letter codes over an alphabet of cardinality n.

Corollary 5.6 *Let n be an integer greater than 1. The number of different non-expandable strong comma-free 3-letter 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 $n \in \mathbf{N}$. By Theorem 5.3, every non-expandable 3-letter code over Σ corresponds to two choices: that of a partition (L_1, R_1) of Σ into two non-empty parts, and next that 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 $m := |L_1|$, 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 5.7 We provide examples for the formula (5.1) given by Corollary 5.6 for alphabets of cardinality at most 4. We also illustrate the way to build all the corresponding 3-letter codes offered by Theorem 5.3.

(1) Corollary 5.6 ensures that there are 4 different non-expandable 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 non-expandable strong comma-free 3-letter 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 non-expandable 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 non-expandable strong comma-free codes over $\{0,1,2\}$ are listed below.

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

(3) Over the genetic alphabet $\mathcal{B} = \{A,C,G,T\}$, there are 160 different non-expandable strong comma-free 3-letter 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 non-expandable 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 non-expandable strong comma-free codes. For instance, if $L_1 = \{A\}$ and $R_1 = \{C,G,T\}$, then the 8 such codes are

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

We continue by showing how to derive formulæ for $N_{\text{scf}}(n,3)$ and $S_{\text{scf}}(n,3)$ from Theorem 5.3. Our next statement generalises Theorem 3.11 of a previous work [22], giving the size, the number and the shape of strong comma-free 3-letter codes over the genetic alphabet ($n = 4$).

Corollary 5.8 *Let n be an integer greater than 1. The number of different strong comma-free 3-letter codes over an alphabet of cardinality n is*

$$S_{\text{scf}}(n,3) = \begin{cases} \frac{4n^3}{27} & \text{if } n \equiv 0 \pmod 3, \\ \left\lceil \frac{2n}{3} \right\rceil^2 \cdot \left\lfloor \frac{n}{3} \right\rfloor & \text{if } n \equiv 1 \pmod 3, \\ \left\lfloor \frac{2n}{3} \right\rfloor^2 \cdot \left\lceil \frac{n}{3} \right\rceil & \text{if } n \equiv 2 \pmod 3. \end{cases} \tag{5.2}$$

Furthermore, every such code is of the form $L_1L_1(\Sigma\backslash L_1)$ or $(\Sigma\backslash L_1)L_1L_1$, where L_1 is a subset of Σ of size $2n/3$, or $\lceil 2n/3 \rceil$ or $\lfloor 2n/3 \rfloor$ (regarding the value of n modulo 3). In particular,

$$N_{\text{scf}}(n,3) = \begin{cases} 2 \cdot \binom{n}{2n/3} & \text{if } n \equiv 0 \pmod 3, \\ 2 \cdot \binom{n}{\lceil 2n/3 \rceil} & \text{if } n \equiv 1 \pmod 3, \\ 2 \cdot \binom{n}{\lfloor 2n/3 \rfloor} & \text{if } n \equiv 2 \pmod 3. \end{cases}$$

Proof: Let X be a strong comma-free 3-letter code of maximal size over an alphabet Σ of cardinality n, so $|X| = S_{\text{scf}}(n, 3)$. By Theorem 5.3, we know that $X = L_1R_2 \cup L_2R_1$ for some choice of a partition (L_1, R_1) of Σ into non-empty parts, and a partition (L_2, R_2) of $X_2 = L_1R_1$. Setting $\ell_i = |L_i| \in \{1, \ldots, n-1\}$ for $i \in \{1, 2\}$, we have $|R_1| = n - \ell_1$ and $|R_2| = \ell_1(n - \ell_1) - \ell_2$. Consequently, as L_1R_2 and L_2R_1 are disjoint,

$$|X| = \ell_1(\ell_1(n - \ell_1) - \ell_2) + (n - \ell_1)\ell_2. \tag{5.3}$$

Note also that $\ell_2 \in \{0, \ldots, \ell_1(n - \ell_1)\}$. Let $f(\ell_1, \ell_2)$ be the right side of (5.3). Optimising $\ell_2 \mapsto f(\ell_1, \ell_2)$ for every fixed value of ℓ_1 yields that the maximum is attained when $\ell_2 = 0$ if $n < 2\ell_1$ and when $\ell_2 = \ell_1(n - \ell_1)$ otherwise. Note that, up to reversing the code X, i.e. reading each of its words from right to left, we may assume without loss of generality that $\ell_1 \geqslant \left\lfloor \frac{n}{2} \right\rfloor$. In this case, the obtained maximum is $\ell_1^2(n - \ell_1)$, which is maximised when $\ell_1 \in \left\{\left\lfloor \frac{2n}{3} \right\rfloor, \left\lceil \frac{2n}{3} \right\rceil\right\}$, which implies to take $\ell_2 = 0$. Therefore,

$$|X| \in \left\{ \left\lfloor \frac{2n}{3} \right\rfloor \left(\left\lfloor \frac{2n}{3} \right\rfloor \left\lceil \frac{n}{3} \right\rceil \right), \left\lceil \frac{2n}{3} \right\rceil \left(\left\lceil \frac{2n}{3} \right\rceil \left\lfloor \frac{n}{3} \right\rfloor \right) \right\},$$

and comparing the two possibilities yields (5.2).

It also follows that every strong comma free 3-letter code X of maximal size corresponds uniquely to the choice of a subset L_1 of Σ of cardinality either $\left\lfloor \frac{2n}{3} \right\rfloor$ (if n is 0 or 1 modulo 3) or $\left\lceil \frac{2n}{3} \right\rceil$ (if n is 2 modulo 3), and then to the choice of a "reading direction", i.e. $X = L_1(L_1R_1)$ or $X = (L_1R_1)L_1$. (Note that there is no over counting here as $|L_1|$ is never $\frac{n}{2}$.) We thus infer that

$$N_{\text{scf}}(n, 3) = \begin{cases} 2 \cdot \binom{n}{2n/3} & \text{if } n \equiv 0 \pmod 3, \\ 2 \cdot \binom{n}{\lceil 2n/3 \rceil} & \text{if } n \equiv 1 \pmod 3, \\ 2 \cdot \binom{n}{\lfloor 2n/3 \rfloor} & \text{if } n \equiv 2 \pmod 3, \end{cases}$$

which concludes the proof. □

5.3 Self-Complementary Triletter Codes

The framework we developed also allows us to study non-expandable self-complementary strong comma-free 3-letter codes. We in particular prove that, in this case, all non-expandable codes are actually of maximal size.

For every alphabet Σ of cardinality $n \geqslant 2$ and every $\pi \in S_\Sigma$, we define $N^{\pi}_{\text{scf}}(n, \ell)$ to be the number of strong comma-free ℓ-letter codes of maximal size over Σ that are π-self-complementary, and $S^{\pi}_{\text{scf}}(n, \ell)$ 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 5.9 *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 non-expandable different π-self-complementary strong comma-free 3-letter 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

$$N^{\pi}_{\mathrm{scf}}(n,3) = 2^{\binom{n/2+1}{2}} = 2^{n(n+2)/8}. \tag{5.4}$$

Proof: The argument is similar to that establishing Theorem 5.3, with the additional requirement that both partitions must satisfy the π-self-complementary conditions, which are $\overleftarrow{\pi(L_1)} = R_1$ and $\overleftarrow{\pi(L_2)} = R_2$. In particular, $|L_1| = |R_1| = n/2$. The number of such ordered pairs (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 $n^2/4$. Consequently, the number of partitions (L_2, R_2) of $L_1R_1 \setminus S$ with $\overleftarrow{\pi(L_2)} = R_2$ is

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

and we thus obtain (5.4). Moreover, because every non-expandable π-self-complementary strong comma-free 3-letter code X is of the form $L_1R_2 + L_2R_1$ for such partitions (L_1, R_1) and (L_2, R_2), it follows that $|X| = n\binom{n/2}{2}$. □

Example 5.10 We consider the genetic alphabet $\mathcal{B} = \{A, C, G, T\}$ along with the involutory bijection c from Definition 2.5. Every strong comma-free 3-letter 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.

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

We end this section by exhibiting another result that is obtained thanks to the approach we used in this work. Recently, the notion of "mixed codes", mixing dinucleotides, trinucleotides and tetranucleotides over the genetic alphabet $\mathcal{B}$, have been introduced [19]: circular mixed codes have been constructed and biologically-inspired properties of such codes studied. In particular it was shown [19, 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 5.11 *Let Σ be an alphabet of even cardinality n and $\pi \colon \Sigma \to \Sigma$ an involution with no fixed point.*

(1) If $n \geqslant 8$ 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.

6 Future Work

We conclude with some remarks about possibilities of future work on ℓ-letter codes (for $\ell \geqslant 4$) raised by the approach used in Subsection 5.2.

Indeed, it appears to us that the set $\mathscr{P}_\ell$, the straightforward generalisation of the definition of the set $\mathscr{P}_3$ to ℓ-letter words, can be proved to be composed of strong comma-free codes only, and to contain all the non-expandable ones. Characterising the choices of partitions leading precisely to the non-expandable strong comma-free codes seems a challenging and stimulating problem, which could lead to finding better asymptotics, when n goes to infinity, for the number of non-expandable strong comma-free ℓ-letter codes.

In addition, since every maximum code is in particular not expandable, the aforementioned injection and the fact that every code arisen from an element in $\mathscr{P}_\ell$ is strong comma-free could lead to a generalisation of Corollary 5.8 to arbitrary world lengths ℓ, by solving an optimisation problem in the variables $|L_1|, \ldots, |L_{\ell-1}|$. Our preliminary investigations lead, for instance, to a statement similar to that of Corollary 5.8 for 4-letter codes, in particular implying that $S_{\mathrm{scf}}(n,4)$ is essentially $\frac{3n^4}{16}$. The method is to adapat the proofs in Subsection 5.2, the analysis being only a little more tedious. We hope it would be possible to generalise this to all values of ℓ. This seems an interesting topic for future work.

Acknowledgements

The authors would like to thank the anonymous referees for their careful readings of preliminary versions of this article, which helped to improve it.

Appendix. 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.

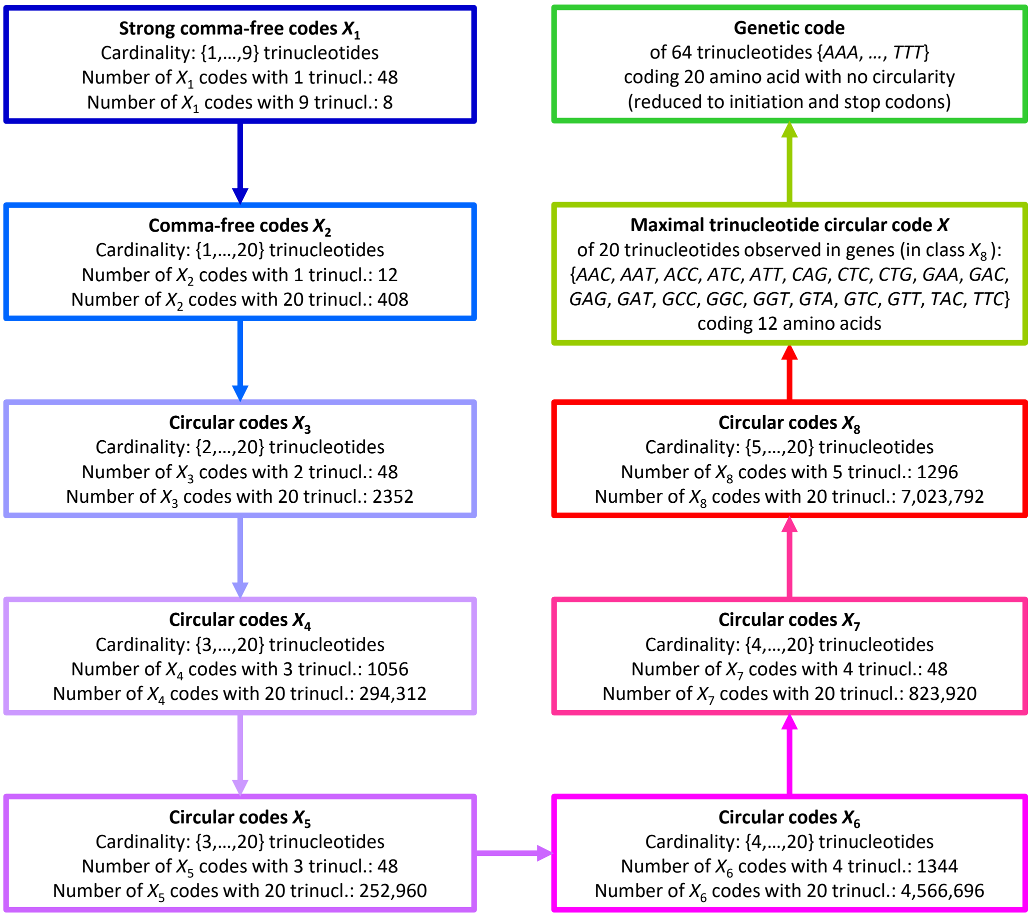
Strong comma-free codes X_1
Cardinality: {1,…,9} trinucleotides
Number of X_1 codes with 1 trinucl.: 48
Number of X_1 codes with 9 trinucl.: 8
Comma-free codes X_2
Cardinality: {1,…,20} trinucleotides
Number of X_2 codes with 1 trinucl.: 12
Number of X_2 codes with 20 trinucl.: 408
Circular codes X_3
Cardinality: {2,…,20} trinucleotides
Number of X_3 codes with 2 trinucl.: 48
Number of X_3 codes with 20 trinucl.: 2352
Circular codes X_4
Cardinality: {3,…,20} trinucleotides
Number of X_4 codes with 3 trinucl.: 1056
Number of X_4 codes with 20 trinucl.: 294,312
Circular codes X_5
Cardinality: {3,…,20} trinucleotides
Number of X_5 codes with 3 trinucl.: 48
Number of X_5 codes with 20 trinucl.: 252,960
Circular codes X_6
Cardinality: {4,…,20} trinucleotides
Number of X_6 codes with 4 trinucl.: 1344
Number of X_6 codes with 20 trinucl.: 4,566,696
Circular codes X_7
Cardinality: {4,…,20} trinucleotides
Number of X_7 codes with 4 trinucl.: 48
Number of X_7 codes with 20 trinucl.: 823,920
Circular codes X_8
Cardinality: {5,…,20} trinucleotides
Number of X_8 codes with 5 trinucl.: 1296
Number of X_8 codes with 20 trinucl.: 7,023,792
Maximal trinucleotide circular code X
of 20 trinucleotides observed in genes (in class X_8):
{*AAC*, *AAT*, *ACC*, *ATC*, *ATT*, *CAG*, *CTC*, *CTG*, *GAA*, *GAC*, *GAG*, *GAT*, *GCC*, *GGC*, *GGT*, *GTA*, *GTC*, *GTT*, *TAC*, *TTC*}
coding 12 amino acids
Genetic code
of 64 trinucleotides {*AAA*, …, *TTT*}
coding 20 amino acid with no circularity
(reduced to initiation and stop codons)

Figure 2: A combinatorial hierarchy of the trinucleotide circular codes leading to the genetic code. The trinucleotide circular codes in X_p is given as a function of the maximal path length p in the associated graph (numbers from Table 4 in Appendix).

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 p (from 1 to 8) in their associated graph G_X, defined in Subsection 4.1.

p $\|X\|$	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

References

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

[2] D. Bajić and T. Loncar-Turukalo, A simple suboptimal construction of cross-bifix-free codes, *Cryptogr. Comm.* 6 (2014), 27–37.

[3] D. Bajić and J. Stojanović, Distributed sequences and search process, *IEEE Int. Conf. Comm.* Vol. 1 (2004), 514–518.

[4] A. H. Ball and L. J. Cummings, Extremal digraphs and comma-free codes, *Ars Combin.* 1(1) (1976), 239–251.

[5] M. Baralle and F. Baralle, The splicing code, *BioSystems* 164 (2018), 39–48.

[6] J. Berstel, D. Perrin and C. Reutenauer, *Codes and Automata*, in: *Encyclopedia of Mathematics and its Applications* Vol. 29, Cambridge University Press, Cambridge, 2010.

[7] S. Bilotta, E. Grazzini, E. Pergola and R. Pinzani, Avoiding cross-bifix-free binary words, *Acta Inform.* 50(3) (2013), 157–173.

[8] S. Bilotta, E. Pergola and R. Pinzani, A new approach to cross-bifix-free sets, *IEEE Trans. Inform. Theory* 58(6) (2012), 4058–4063.

[9] S. R. Blackburn, Non-overlapping codes, *IEEE Trans. Inform. Theory* 61(9) (2015), 4890–4894.

[10] S. A. Cakiroglu, J. B. Zaugg and N. M. Luscombe, Backmasking in the yeast genome: encoding overlapping information for protein-coding and rna degradation, *Nucleic Acids Res.* 44 (2016), 8065–8072.

[11] 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(1) (2011), #P194, 18pp.

[12] 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(7) (2013), 4668–4674.

[13] F. H. C. Crick, J. S. Griffith and L. E. Orgel, Codes without commas, *Proc. Nat. Acad. Sci. U.S.A.* 43(5) (1957), 416–421.

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

[15] M. S. Dodd, D. Papineau, T. Grenne, J.F. Slack, M. Rittner, F. Pirajno, J. O'Neil and C. T. S. Little, Evidence for early life in earth's oldest hydrothermal vent precipitates, *Nature* 543 (2017), 60–64.

[16] W. L. Eastman, On the construction of comma-free codes, *IEEE Trans. Inform. Theory* IT-11 (1965), 263–267.

[17] B. Eslami-Mossallam, R. D. Schram, M. Tompitak, J. van Noort and H. Schiessel, Multiplexing genetic and nucleosome positioning codes: a computational approach, *PLoS One* 11(6) (2016), e0156905.

[18] G. Faure, A. Ogurtsov, S. A. Shabalina and E. V. Koonin, Adaptation of mrna structure to control protein folding, *RNA Biol.* 14 (2017), 1649–1654.

[19] E. Fimmel, C. J. Michel, F. Pirot, J.-S. Sereni and L. Strüngmann, Mixed circular codes, *Math. Biosci.* 317 (2019), 108231, 14 pp.

[20] E. Fimmel, C. J. Michel, and L. Strüngmann, n-nucleotide circular codes in graph theory, *Phil. Trans. Roy. Soc. A* 374(2063) (2016), 20150058, 19 pp.

[21] E. Fimmel, C. J. Michel and L. Strüngmann, Diletter circular codes over finite alphabets, *Math. Biosci.* 294 (2017), 120–129.

[22] E. Fimmel, C. J. Michel and L. Strüngmann, Strong comma-free codes in genetic information, *Bull. Math. Biol.* 79(8) (2017), 1796–1819.

[23] E. Fimmel and L. Strüngmann, On the hierarchy of trinucleotide n-circular codes and their corresponding amino acids, *J. Theoret. Biol.* 364 (2015), 113–120.

[24] E. Fimmel and L. Strüngmann, Maximal dinucleotide comma-free codes, *J. Theoret. Biol.* 389 (2016), 206–213.

[25] E. Fimmel and L. Strüngmann, Mathematical fundamentals for the noise immunity of the genetic code, *BioSystems* 164 (2018), 186–198.

[26] S. W. Golomb, B. Gordon and L. R. Welch, Comma-free codes, *Canadian J. Math.* 10 (1958), 202–209.

[27] S. W. Golomb, L. R. Welch and M. Delbrück, Construction and properties of comma-free codes, *Biologiske Meddelelser, Kongelige Danske Videnskabernes Selskab* 23 (1958), 1–34.

[28] L. J. Guibas and A. M. Odlyzko, Maximal prefix-synchronized codes, *SIAM J. Appl. Math.* 35(2) (1978), 401–418.

[29] N. H. Lam, Finite completion of comma-free codes I, *Theor. Inform. Appl.* 38(2) (2004), 91–115.

[30] N. H. Lam, Finite completion of comma-free codes II, *Theor. Inform. Appl.* 38(2) (2004), 117–136.

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

[32] V. I. Levenšteĭn, The maximal number of words in codes without overlap, *Problemy Peredači Informacii* 6(4) (1970), 88–90.

[33] V. I. Levenšteĭn, Combinatorial problems motivated by comma-free codes, *J. Combin. Des.* 12(3) (2004), 184–196.

[34] C. J. Michel, A 2006 review of circular codes in genes, *Comput. Math. Appl.* 55(5) (2008), 984–988.

[35] C. J. Michel and G. Pirillo, Strong trinucleotide circular codes, *Int. J. Comb.* (2011), Art. ID 659567, 14 pp.

[36] C. J. Michel, G. Pirillo and M. A. Pirillo, Varieties of comma-free codes, *Comput. Math. Appl.* 55(5) (2008), 989–996.

[37] K. Prakash and D. Fournier, Evidence for the implication of the histone code in building the genome structure, *BioSystems* 164 (2018), 49–59.

[38] R. A. Scholtz, Maximal and variable word-length comma-free codes, *IEEE Trans. Inform. Theory* IT-15 (1969), 300–306.

[39] B. Tang, S. W. Golomb and R. L. Graham, A new result on comma-free codes of even word-length, *Canad. J. Math.* 39(3) (1987), 513–526.

[40] P. Turán, Eine Extremalaufgabe aus der Graphentheorie, *Mat. Fiz. Lapok* 48 (1941), 436–452.

(Received 6 Aug 2020; revised 6 Feb 2023, 25 Apr 2023)