Circular mixed sets*\textsuperscript{a}
Elena Fimmel\textsuperscript{a}, Christian J. Michel\textsuperscript{b,∗}, Lutz Strüngmann\textsuperscript{a}

\textsuperscript{a} Institute of Mathematical Biology, Faculty for Computer Sciences, Mannheim University of Applied Sciences, 68163 Mannheim, Germany
\textsuperscript{b} Theoretical Bioinformatics, ICube, University of Strasbourg, C.N.R.S., 300 Boulevard Sébastien Brant, 67400 Illkirch, France

**A R T I C L E   I N F O**

Keywords:
Circular mixed set
Circular mixed code
Graph properties
Evolution model of the genetic code

**A B S T R A C T**

In this article, we introduce the new mathematical concept of circular mixed sets of words over an arbitrary finite alphabet. These circular mixed sets may not be codes in the classical sense and hence allow a higher amount of information to be encoded. After describing their basic properties, we generalize a recent graph theoretical approach for circularity and apply it to distinguish codes from sets (i.e. non-codes). Moreover, several methods are given to construct circular mixed sets. Finally, this approach allows us to propose a new evolution model of the present genetic code that could have evolved from a dinucleotide world to a trinucleotide world via circular mixed sets of dinucleotides and trinucleotides.

1. Introduction

A new mathematical concept of circular mixed sets of words over an arbitrary finite alphabet is developed in this article. This approach is a generalization of the circular code theory based on recent results on a finite alphabet with the mixed circular codes (Fimmel et al., 2019) and the relation between \(k\)-circularity and circularity of codes (Fimmel et al., 2020a), and also on different results on the 4-letter genetic alphabet obtained since 1996 (reviewed in Michel, 2008; Fimmel and Strüngmann, 2018). Circularity of a code allows to retrieve the construction (reading) frame unambiguously as the reading of a shifted frame with a circular code is impossible. It is the classical concept of circularity which necessarily implies that the sets of words are codes and which has led to numerous developments in the theory of the genetic code (Michel, 2008; Fimmel and Strüngmann, 2018). Circularity of a set however relaxes the previous unambiguously constraint on the construction frame. Indeed, it allows the construction frame to be read in several different ways, i.e. a higher amount of information can be encoded in the construction frame, but the reading of a shifted frame with a circular set is still impossible. Finally, this work takes up an old issue to determine whether a finite set of variable-length words over a finite alphabet is a code, see for example Berstel et al. (2009). In particular, one algorithm uses the so-called prefix graph of a mixed set of words for determining whether the mixed set is a code (Berstel et al., 2009). We give here a new criterion for this problem using our graph theoretical approach that has been developed to study the circularity of codes (Fimmel et al., 2016), leading to a circularity graph significantly different from this prefix graph.

The paper is structured as follows. In Section 2, we give the definitions of a set, a code, the circularity and comma-freeness (comma-free and strong comma-free) for a set, and the graph associated with a set. These generalized definitions for sets can (obviously) be applied to codes. Then, two examples of the graph associated with a set, one acyclic and the other cyclic, are proposed. Finally, we study the union of circular codes of different word lengths. Examples of the three cases are given: a mixed set that is a code and not circular, a mixed set that is not a code but circular and a mixed set that is not a code and not circular. A theorem states that the only way to construct circular mixed sets is the union of circular \(\ell\)-letter codes. In Section 3, we define a colouring of the associated graph and propose Theorem 5 to identify the codes among the mixed sets. We give two examples from Berstel et al. (2009) to demonstrate our circularity graph approach and to illustrate its differences with the classical graph approach (Berstel et al., 2009). An algorithm is also proposed for determining whether a mixed set is a code. In Section 4, we point out some constructions of circular mixed sets. A proposition proves that the union of two strong comma-free codes is a circular set. Another proposition states that the union of two circular codes with some conditions on the prefix and suffix, is a circular code. Theorem 6 allows to construct a circular mixed set from the union of diletter codes and tetraletter codes. In Section 5, the new mathematical concept of circular mixed sets of words introduced here allows to generalize our previous evolution models of the genetic code in order to explain the transition from a dinucleotide world to a trinucleotide world (codons).

---

\* The authors would like to thank Dr. Martin Starman for many helpful discussions on the manuscript.

∗ Corresponding author.

E-mail addresses: e.fimmel@hs-mannheim.de (E. Fimmel), c.michel@unistra.fr (C.J. Michel), l.struengmann@hs-mannheim.de (L. Strüngmann).

https://doi.org/10.1016/j.biosystems.2023.104906

Received 22 March 2023; Accepted 29 April 2023
Available online 15 May 2023

0303-2647/© 2023 Elsevier B.V. All rights reserved.
2. Definitions and examples

We first state some definitions and results that are needed in the sequel. Let $\Sigma$ be an arbitrary finite alphabet of cardinality $n \in \mathbb{N}$. As commonly used in the theory of formal languages, $\Sigma^+ = \{l_1 \cdots l_n \mid l_i \in \Sigma, n \in \mathbb{N}\}$ is the set of all words $w$ over $\Sigma$ of finite length excluding the empty word $\epsilon$ while $\Sigma^* = \Sigma^+ \cup \{\epsilon\}$. Moreover, for $n \in \mathbb{N}$, the set $\Sigma^n(\Sigma^m)$ consists of all words $w \in \Sigma^n$ of finite length $l(w) = n \leq m$, respectively, where the length of a word $w = l_1 \cdots l_n \in \Sigma^n$ is defined as $l(w) = n$. By convention, the empty word has length zero.

Definition 1.

(i) A set $X \subset \Sigma^+$ is a code if every word $w \in \Sigma^+$ has a single decomposition into words of $X$.

(ii) For $\ell \in \mathbb{N}$ with $\ell \geq 2$, an $\ell$-letter code is a subset of $\Sigma^\ell$ and the elements of $\Sigma^\ell$ are called $\ell$-letter words.

Obviously, any set $X \subset \Sigma^\ell$ of $\ell$-letter words is a code which justifies (ii) of Definition 1. However, the set $X = \{101,11,10\}$ is not a code over the binary alphabet $\Sigma = \{0,1\}$ since the word $101 \in X$ has a second decomposition, namely $101 = 10 \cdot 11$ of $X$.

We now define circularity and comma-freeness for sets. For this purpose, we need to recall the graph construction from Fimmel et al. (2016). Recall that a word $x \in \Sigma^+$ is called a prefix (suffix, respectively) of a word $w \in \Sigma^+$ if $w = xy$ ($w = yx$, respectively) for some word $y \in \Sigma^*$. A prefix/suffix $x$ of $w$ is called proper if $x \neq w$.

Definition 2. Let $X \subset \Sigma^+$ be a set. A graph $G = G(X) = (V(X), E(X))$ is associated with $X$ in the following way:

1. The set of vertices $V(X)$ of $G$ is given by $V(X) = \{w \in \Sigma^+ \mid v = u w' \text{ or } v = w' u \text{ for some word } w' \in \Sigma^+ \text{ and } v \in X\}$, i.e. the set of all proper prefixes and proper suffixes of words from $X$;

2. The set of directed edges $E(X)$ of $G$ is given by $E(X) = \{(w, w') \mid w w' \in X \text{ and } w, w' \in V(X)\}$.

For the convenience of the reader, we give some examples of the graph associated with a set. Note that by definition $X \subset \Sigma$ is allowed but its graph $G(X)$ is trivial.

Example 1. Let $\Sigma$ be a finite alphabet, e.g. $\Sigma = \{a,b,c,d\}$, and $X \subset \Sigma$ an arbitrary subset, e.g. $X = \{a,b,c\}$. Then $X$ is a code and its associated graph $G(X)$ is empty.

The next example shows that the 1-letter words of a set $X$ are vertices in the associated graph $G(X)$ or not.

Example 2. The mixed set $X = \{h,d,abc,abcd\}$ on the alphabet $\Sigma = \{a,b,c,d\}$ is not a code as the word $abc$ has 2 different decompositions into words of $X$, namely $abcd$ and $abc[d].$ Moreover, the graph associated with $X$ has the vertex $d$ but not the vertex $b$ (Fig. 1).

The next example demonstrates that a graph associated with a set $X$ can be not acyclic, i.e., it may contain a circle. This will be important in the sequel since it is related to the notion of circularity (see Definition 3).

Example 3. The mixed set $X = \{a, bc, da, abc, bcd\}$ on the alphabet $\Sigma = \{a,b,c,d\}$ is not a code as the word $abc$ has two different decompositions into words of $X$, namely $abc$ and $a[bc].$ Moreover, the graph associated with $X$ contains the cycle $a \rightarrow b \rightarrow d \rightarrow a$ (Fig. 2).

In analogy to Fimmel et al. (2016), we now define comma-freeness and circularity of sets $X \subset \Sigma^+$ using associated graphs. Recall from graph theory that a path is a walk through the graph for which all edges are different while a cycle in a (directed) graph is a closed (directed) trail in which only the first and last vertices are the same. A directed graph $G$ is called acyclic if there is no cycle in $G$ (Diestel, 2005 for more details on graph theory).

Definition 3. A set $X \subset \Sigma^+$ is

1. circular if its associated graph is acyclic;
2. comma-free if $X$ is circular and the maximal length of a path in its associated graph $G(X)$ is at most 2;
3. strong comma-free if $X$ is circular and the maximal length of a path in its associated graph $G(X)$ is 1.

Clearly, the class of circular sets is closed under intersection and subset relation. Moreover, in the case of codes with word length $\geq 2$, the above Definition 3 resembles the original definitions of comma-freeness and circularity from Fimmel et al. (2016). However, note that our Definition 3 also includes $X \subset \Sigma$ which is a code with trivial associated graph $G(X)$, and hence it is circular in our sense but not in the sense of Theorem 1 below. If $X$ contains 1-letter words but also at least one word of length greater than 1, then it is case dependent if the two notions coincide. Therefore, we require the condition $X \subset \Sigma^{\leq 2}$ for simplicity in the next theorem.

Theorem 1 (Theorem 2.6 and Theorem 2.11 Fimmel et al., 2016, and Definition 2.7 Fimmel et al., 2017b). A code $X \subset \Sigma^{\leq 2}$ is

1. strong comma-free if no element of $\Sigma^+$ appears both as a prefix and a suffix in $X$;
2. comma-free if every concatenation $c_1 c_2$ of 2 words from $X$ does not contain as a substring any word from $X$ but $c_1$ and $c_2$ themselves;
3. circular if for any finite concatenation $c_1 \cdots c_m$ of elements from $X$, where $m \in \mathbb{N}$, there is only one partition into elements from $X$ when read on a circle. Any such partition is a circular decomposition of $c_1 \cdots c_m$.

Note that obviously, any set satisfying condition (3) from Theorem 1 has to be a code while this is not implied by condition (1) from
Moreover, $X$. Then $\text{Proof.}$

**Example 6.** Consider the mixed set $X = \{ cd, abc, dab \}$ on the alphabet $\Sigma = \{ a, b, c, d \}$ that is a code and not circular. The cyclic graph of the mixed set $X$ is shown in Fig. 3.

**Example 5.** Consider the mixed set $X = \{ ab, ad, ca, caa, dab \}$ on the alphabet $\Sigma = \{ a, b, c, d \}$ that is not a code but circular. The cyclic graph of the mixed set $X$ is shown in Fig. 4.

**Example 6.** Consider the mixed set $X = \{ ab, ad, ca, caa, dab \}$ on the alphabet $\Sigma = \{ a, b, c, d \}$ that is not a code but circular. The cyclic graph of the mixed set $X$ is shown in Fig. 5.

---

**Definition 3.** Circular codes and (strong) comma-free codes have been studied extensively in the literature, especially in the biological context of frameshift recognition over the genetic code. For the convenience of the reader, we recall that diletter circular codes on an alphabet of size 4 can have a maximum size of at most 6 (Michel and Pirillo, 2013; Fimmel et al., 2015, 2017a) while triletter circular codes could have a cardinality of at most 20 (Arquès and Michel, 1996). Moreover, the structure of circular codes $X \subseteq \Sigma^2$ of maximal size with $\Sigma$ finite is known (Fimmel et al., 2017a).

The following examples show that the union of circular codes of different word lengths can be mixed sets that are not circular and even not a code.

**Example 4.** Consider the 2 circular codes $X_2 = \{ cd \}$ and $X_1 = \{ abc, dab \}$ on the alphabet $\Sigma = \{ a, b, c, d \}$. The mixed set $X = \{ cd, abc, dab \}$ is a code that is not circular. Indeed, the word $dabdcdabe$ admits the 2 circular decompositions $dab|cd|abc$ and $d|abc|dabc$. The graph associated with $X$ contains the cycle $c \rightarrow d \rightarrow ab \rightarrow c$ (Fig. 3).

**Example 5.** Consider the 2 circular codes $X_2 = \{ ab, ad, ca \}$ and $X_3 = \{ caa, dab \}$ on the alphabet $\Sigma = \{ a, b, c, d \}$. The mixed set $X = \{ ab, ad, ca, caa, dab \}$ is not a code as the word $caadab$ has 2 different decompositions into words of $X$, namely $ca|ad|ab$ and $caa|dab$. However, the graph associated with $X$ is acyclic, so $X$ is a circular set (Fig. 4).

The following easy result states that in fact the examples above are the only way to construct circular mixed sets, namely as unions of circular $\ell$-letter codes.

**Theorem 2.** Let $X \subseteq \Sigma^+ \subseteq \Sigma^*$ be a circular set. Then $X$ is the disjoint union $\bigcup_{\ell \in \mathbb{N}}(X \cap \Sigma^\ell \cap \Sigma)$ of circular $l$-letter codes.

**Proof.** Let $X \subseteq \Sigma^+ \subseteq \Sigma^*$ be as stated. Put $X_{\ell} = X \cap \Sigma^\ell$ for any $\ell \in \mathbb{N}$. Then $X_{\ell}$ is an $\ell$-letter code (since it contains words of length $\ell$ only). Moreover, $X_{\ell}$ is circular as a subset of a circular set. □
size 26. Precisely, if $X_2 = \{ab, ac, ad, bc, bd, cd\}$ is a circular diletter code of maximal size, then $X = X_2 \cup X_3$ is a circular code of maximal size 26 where $X_3 = \{aab, aac, aad, bab, bac, bad, bbd, cab, cac, cad, cbc, cdc, cdd, dba, dca, dab, dbd, dcd\}$.

Similarly, circular codes $X \subseteq \Sigma^3 \cup \Sigma^4$ of maximal size 80 were constructed if $\Sigma$ has size 4 (Fimmel et al., 2019).

**Theorem 4** (Theorem 4 Fimmel et al., 2019). Let $\Sigma$ be a finite alphabet of size 4. Then there are circular mixed codes $X \subseteq \Sigma^3 \cup \Sigma^4$ of maximal cardinality 80.

Interestingly, the situation already gets much more complicated when passing to circular codes $X \subseteq \Sigma^3 \cup \Sigma^1 \cup \Sigma^4$. Examples of cardinality 81 are known but in general the maximal cardinality of such a circular mixed code or set is an open problem.

As we are interested in circular mixed sets (mostly not codes), we will show in the next section an efficient way to detect the code property of a set using its associated graph.

### 3. Codes defined by coloured graphs

In this section, we approach the problem of identifying codes among mixed sets by using an appropriate colouring of the associated graph. Thus, let $X \subseteq \Sigma^+$ be a mixed set and $G(X)$ the graph associated with $X$. We define a colouring on $G(X)$ as follows:

**Definition 4.** Let $\Sigma$ be a finite alphabet and $X \subseteq \Sigma^+$ be a mixed set. Moreover, let $G(X)$ be the graph associated with $X$. A vertex colouring $c$ on $G(X)$ is defined by

$$c : V(G(X)) \rightarrow \{\text{green, yellow}\}$$

such that $c(v) = \text{green} \Leftrightarrow v \in X^+$ and $c(v) = \text{yellow} \Leftrightarrow v \notin X^+$.

**Definition 4** proposes a graph colouring to determine if some vertices of the associated graph are themselves concatenations of words from $X$, thus allowing to detect ambiguous words over $X$, i.e. to identify if $X$ is a code or not. We give an example of such an associated coloured graph.

**Example 7.** Let $X$ be the code constructed in Theorem 3 on the alphabet $\Sigma = \{a, b, c, d\}$. Then its associated coloured graph is given in Fig. 6.

**Theorem 5.** Let $\Sigma$ be a finite alphabet and $X \subseteq \Sigma^+$ be a mixed set. Moreover, let $G(X)$ be the graph associated with $X$ with $c$ being the colouring on $G(X)$ defined in **Definition 4**. Then $X$ is a code if and only if there is no path in $G(X)$ starting and ending with a green vertex.

**Proof.** (1). Let $w_1 \rightarrow \ldots \rightarrow w_n$ be a path in $G(X)$ that starts and ends with a green vertex. By **Definition 4**, $w_1 \in X^+$ and $w_n \in X^+$. We now distinguish two cases:

- **Case I:** $n$ is odd
  
  By the definition of $G(X)$, $w_1w_2 \in X$, \ldots, $w_{n-1}w_n \in X$ and $w_2w_3 \in X$, \ldots, $w_{n-2}w_{n-1} \in X$. Thus the word $w_1 \ldots w_n$ has 2 decompositions over $X$, namely
  
  $$w_1w_2 \cdot \ldots \cdot w_{n-1}w_n$$
  
  and
  
  $$w_1 \cdot \ldots \cdot w_{n-2}w_{n-1} \cdot w_n.$$

  Hence $X$ is not a code.

- **Case II:** $n$ is even
  
  By the definition of $G(X)$, $w_1w_2 \in X$, \ldots, $w_{n-2}w_{n-1} \in X$ and $w_2w_3 \in X$, \ldots, $w_{n-2}w_{n-1} \in X$. Thus the word $w_1 \ldots w_n$ has 2 decompositions over $X$, namely
  
  $$w_1w_2 \cdot w_3w_4 \cdot \ldots \cdot w_{n-2}w_{n-1} \cdot w_n$$

  and
  
  $$w_1 \cdot w_2w_3 \cdot \ldots \cdot w_{n-2}w_{n-1}w_n.$$

  Hence $X$ is not a code.

(2). Conversely, now assume that $X$ is not a code, hence there is a (shortest) concatenation of words $w_1, \ldots, w_n \in X$ such that $w_1w_2 \ldots w_n$ has a second decomposition over $X$, let say $v_1 \cdot v_2 \cdot \ldots \cdot v_m$ for some words $v_1, \ldots, v_m \in X$. Assume without loss of generality that $v_1$ is shorter than $w_1$. Note that if they have the same length, then they are equal, contradicting the shortness of the concatenation. Let $s$ be the largest integer such that $v_1 \ldots v_k$ is shorter than $w_1$. Without loss of generality, we may assume that $s = 1$, implying $v_1 \in X^+$. Hence $w_1 = v_1 \cdot v'_1 \cdot \ldots \cdot v'_s$ for some $v'_1 \in \Sigma^+$. Consequently, $v'_1$ is now a prefix of $v_2$ and thus $v'_2 = v'_1w'_2$ and, as above without loss of generality, $w'_2$ is a prefix of $w_2$, i.e. $w'_2 = w'_1w'_2$. Again, $v'_2$ is now a prefix of $v_3$, i.e. $v_3 = v'_2w'_3$. We continue this way and end up with the path $v_1 \rightarrow v'_1 \rightarrow w'_1 \rightarrow v'_2 \rightarrow w'_2 \rightarrow \ldots$

The path ends with either $v'_1$ or $w'_1$, which must equal to $v_m$ or $w_n$, and hence is an element of $X^+$. Thus the above path starts with a green vertex $v_1 \in X^+$ and ends with a green one.

Thus, determining whether a set is a code or not becomes obvious by analysing its graph. With **Example 7** where the code is constructed using **Theorem 3**, its graph given in Fig. 6 easily shows that it is indeed a code. Moreover, the above **Theorem 5** is a generalization of **Theorem 6.1** in Fimmel et al. (2019).

**Corollary 1** (Theorem 6.1 Fimmel et al., 2019). Let $\Sigma$ be a finite alphabet of size 4. For $i \in \{2, 3, 4\}$, let $X_i \subseteq \Sigma^i$ and set $X := X_2 \cup X_3$ and $\bar{X} := X_3 \cup X_4$. Fig. 6. (Example 7). The acyclic graph of the code constructed in Theorem 3 on the alphabet $\Sigma = \{a, b, c, d\}$. 

- **Case II:** $n$ is even
  
  By the definition of $G(X)$, $w_1w_2 \in X$, \ldots, $w_{n-2}w_{n-1} \in X$ and $w_2w_3 \in X$, \ldots, $w_{n-2}w_{n-1} \in X$. Thus the word $w_1 \ldots w_n$ has 2 decompositions over $X$, namely
  
  $$w_1w_2 \cdot \ldots \cdot w_{n-2}w_{n-1} \cdot w_n$$  
  
  and
  
  $$w_1 \cdot w_2w_3 \cdot \ldots \cdot w_{n-2}w_{n-1}w_n.$$  

  Hence $X$ is not a code.
The mixed set $X$ is a code if and only if there exists no directed path in $G(X)$ between a pair of vertices with labels in $X_2$.

(ii) The mixed set $\tilde{X}$ is a code if and only if there exists no directed path in $G(\tilde{X})$ between a pair of vertices with labels in $X_3$.

(iii) The mixed sets $X$ and $\tilde{X}$ are circular codes if and only if $X$ and $\tilde{X}$ are codes and $G(X)$ and $G(\tilde{X})$ are acyclic.

Proof. Clearly, a vertex in $G(X)$ can only be in $X^+$ if it is in $X_2$ and similarly, a vertex in $\tilde{X}$ can only be in $\tilde{X}^+$ if it is in $X_2$. Thus (i) and (ii) follow from Theorem 5 while (iii) is obvious.

Before we give some examples of sets that are not codes, we briefly describe the algorithm associated with Theorem 5. The colouring is based on the decomposition of words from the set $X$ into concatenations of words from $X$, which is equivalent to the calculation of $X^{\leq n}$ for some natural number $n$. This computation is the most time consuming part of the algorithm.

Algorithm 1. Let $\Sigma$ be a finite alphabet and $X \subseteq \Sigma^*$ be a mixed set. The following algorithm is performed in order to detect if $X$ is a code or not:

- Define the graph $G(X)$ associated with $X$.
- Calculate the set $V'(G(X)) = \{ v \in V(G(X)) : v \in X^+ \}$.
- Colour the graph $G(X)$ as in Definition 4.
- Calculate all paths between green coloured vertices.

Algorithms determining whether a set is a code or not, are of course not new. In fact, in the book (Berstel et al., 2009), a whole chapter is devoted to this question. The algorithms presented by the authors, in particular the so-called prefix graph of a code, are different from our approach. For the convenience of the reader, we pick up the examples from Berstel et al. (2009) to demonstrate our approach.

Example 8 (Example 3.1 from Berstel et al., 2009). Let $\Sigma = \{a,b\}$ and $X = \{b, abb, bbba, abbb, baab\}$. Then $X$ is not a code as its associated graph $G(X)$ contains several directed paths that start and end with a green node. For example, the path $bb \to ba \to ab$ gives the word $bbbaab$ with 2 different decompositions $bbbaab|abb = b|bbab$. Accordingly, $X$ is not a code (Fig. 7).

In fact, all ambiguous words are given by the following regular expression:
\[ abbbabba + bbbabba + abbbab(baab)^* + bbbab(baab)^* \]

A second example from Berstel et al. (2009) where there is a unique path starting and ending with a green vertex.

Example 9 (Example 8.1 from Berstel et al., 2009). Let $\Sigma = \{a,b\}$ and $X = \{a, bb, abbb, baab\}$. The graph $G(X)$ shows a (unique) directed path that starts and ends with a green node: the path $abb \to ba \to bab \to ab \to bba$ gives the word $abbbababbbab$ with 2 different decompositions $abbbabab|bb = abba|babbbb|bba$. Accordingly, $X$ is not a code (Fig. 8).

Let us note that the graph theoretical approach in Berstel et al. (2009) relies on Theorem 8.2, whose formulation sounds very similar to that of our Theorem 5. However, it is based on a different and more elaborate definition of the graph than our approach.

4. Constructing circular mixed sets

In this section, we will give some properties and criteria when two graphs of circular codes can be joint to an acyclic graph, i.e. their union is a circular mixed set.

Proposition 1. Let $\ell, n \in \mathbb{N}, \ell \neq n$, $X_1 \subseteq \Sigma^\ell, X_2 \subseteq \Sigma^n$ be strong comma-free codes. Then $X_1 \cup X_2$ is a circular mixed set.

Proof. Assume that $X_1 \cup X_2$ is not circular. Then there is an oriented circle in $G(X_1 \cup X_2)$ that can be only of even length and containing alternating edges of $G(X_1)$ and $G(X_2)$, since $X_1$ and $X_2$ are strong comma-free. Thus, $\ell = n$ which contradicts the assumption.

The above condition is sufficient but not necessary, as the following example shows:

Example 10. Let $\Sigma = \{a, b, c, d\}$ and $X_1 = \{abc, bcd\}$, $X_2 = \{bc\}$. The code $X_1$ is not strong comma-free and the code $X_2$ is strong comma-free but $X_1 \cup X_2$ is a circular mixed code (Fig. 9).

It should also be noted that it would not be enough if $X_1$ and $X_2$ were ‘simply’ comma-free in the above proposition, as the next example shows:

Example 11.
Example 11. Let $\Sigma = \{a, b, c, d\}$ and $X_1 = \{abc, bcd\}$, $X_2 = \{cba, dcb\}$. The codes $X_1$ and $X_2$ are both comma-free but $X = X_1 \cup X_2$ is a mixed code that is not circular as the word $w = abedcba$ has 2 decompositions into words of $X$ on the circle: $abc|dcb = bcd|cba$ (Fig. 10).

Let us finally note that Proposition 1 does not guarantee that $X_1 \cup X_2$ is a mixed code.

Example 12. Let $\Sigma = \{a, b, c, d\}$ and $X_1 = \{ab, ca, cd\}$, $X_2 = \{bca, cda\}$. The codes $X_1$ and $X_2$ are both strong comma-free but the acyclic graph $G(X_1 \cup X_2)$ contains a directed path $cd \rightarrow a \rightarrow b \rightarrow ca$ which starts and ends with a green vertex. In other words, the word $cdabca$ has 2 decompositions $cd|ab = ba|bc$ into words of $X_1 \cup X_2$ and thus $X_1 \cup X_2$ is not a mixed code (Fig. 11).

However, if we consider the following codes $X_1 = \{ca, cd\}$ and $X_2 = \{bca, cda\}$ where $X_1$ and $X_2$ are both strong comma-free then $X_1 \cup X_2$ is a circular mixed code (see Fig. 12).

Proposition 2. Let $\ell, n \in \mathbb{N}$, $\ell \neq n$, $X_1 \subseteq \Sigma^\ell$, $X_2 \subseteq \Sigma^n$ be circular codes. If no prefix of a word from $X_1$ is at the same time a suffix of a word from $X_2$, then $X_1 \cup X_2$ is a circular mixed code.

Proof. (1). Suppose we have a directed circle $C$ in $G(X_1 \cup X_2)$. The circle $C$ must contain edges from both $G(X_1)$ and $G(X_2)$ since the two sets are circular and the associated graphs cannot contain circles. Consider the longest segment on $C$ which consists only of edges from $X_2$, and its last directed edge $(v, w)$. Then $w$ is concurrently suffix of a word from $X_2$ and prefix of a word from $X_1$. Contradiction to the assumption.
we only have words of two different lengths
Theorem 5, there is a directed path in
Without lost of generality, let
\{T\} and \{a, b, c, d\}.

\(\Sigma = \{a, b, c, d\}\).

\(\{aabb, aacc, addd, aedc, bbdd, dabc\} \subseteq T_{\text{ess}}(\Sigma)\)
be a circular code of maximal size 12. Then \(D \cup X\) is a circular mixed set of size 18 (Fig. 14).

5. A new evolution model

The new mathematical concept of circular mixed sets of words introduced here allows to generalize our previous evolution models of the genetic code, in particular the model based on circular mixed codes (see Figure 8 in Fimmel et al., 2019). Indeed, the property of circular mixed sets allows to read DNA words of different lengths by increasing the combinatorial decoding in “reading frame”, which is impossible with the circular codes, while excluding the decoding in frameshifts, as with the circular codes. Circular sets are less constrained than circular codes from a coding point of view. Furthermore, while the genetic code for coding amino acids is based on DNA words of length 3, i.e. trinucleotides, it is unrealistic, from our point of view, that primitive life began directly with words of length 3, and only with words of length 3. It is more likely that primitive soup began with words of length 2, i.e. dinucleotides, then a mixing of dinucleotides and trinucleotides when the coding required the introduction of longer DNA words to code a greater number of amino acids, to finally arrive at the current coding of 20 amino acids with a genetic code based on 64 trinucleotides and no dinucleotides. Several biological observations are consistent with such a mixing model, to name a few. It was proposed that dinucleotides bound the alpha-keto acid precursor of the amino acid coded and catalyzed its transformation to the cognate amino acid (Copley et al., 2005; Yaman and Harvey, 2021). Identical dinucleotides in the 1st and 2nd codon positions code for the same or closely related amino acids. For example, the current codons GGN all code for glycine, suggested that the GG dinucleotide played a role in the synthesis of glycine. Correlations between the anticond dinucleotides and their amino acids are observed (Jungck, 1978). In the evolution of the genetic code, self-complementary dinucleotides are also involved in the 1st and 2nd codon positions, namely: CGN for Arg (N being any nucleotide), ATT, ATC, and ATA for Ile, and TAT and TAC for Tyr (Rodin et al., 2011). Dinucleotides defined by the 3rd nucleotide of a codon \(c_1\) and the 1st nucleotide of a next codon \(c_2\) has a role in the codon concatenation \(c_1 \cdot c_2\) (Parvathy et al., 2022).

From a coding point of view, the crossover from a dinucleotide world to a trinucleotide world during evolution can be modelled with the following evolutionary steps:
The circular mixed set \( \{ ba, ca, cb, cd, da, db, aab, acc, add, acb, bbd, bda, ccd, cdb, dba, dbe, dca \} \) of diletter and tetraletter words of size 18 on the alphabet \( \Sigma = \{ a, b, c, d \} \).

(i) a dinucleotide word, thus under a theory of circular dinucleotide codes;
(ii) a mixing of dinucleotides and trinucleotides such that dinucleotides and trinucleotides can be read simultaneously and in several ways (so is not a code) by avoiding the frameshifts (so circularity is preserved to some extent), thus under a theory of circular mixed sets;
(iii) a mixing of dinucleotides and trinucleotides such that dinucleotides and trinucleotides can be read simultaneously and unambiguously in all frames, thus under a theory of circular mixed codes;
(iv) a trinucleotide word, thus under a theory of circular trinucleotide codes;
(v) the genetic code.

Two evolutionary steps were already identified. The evolutionary step (iv) is observed in nature as a maximal (20 trinucleotides) self-complementary circular trinucleotide code is identified in genes of bacteria, archaea, eukaryotes, plasmids and viruses (Arquès and Michel, 1996; Michel, 2015, 2017):

\[
X = \{ AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GTG, GTA, GTT, TTC \}. \tag{1}
\]

This circular code \( X \) seems to play an important role in the frameshift recognition of the translational process in the ribosome (Thompson et al., 2021; Michel and Thompson, 2020; Dila et al., 2019). The evolutionary step (iii) has also been investigated as Proposition 5 in Fimmel et al. (2019) has proved that at most two dinucleotides, namely \{ AT, GC \}, can be added to \( X \) in order to obtain a self-complementary circular mixed code.

This opens the question of the existence of larger circular mixed sets when adding dinucleotides to the code \( X \). Surprisingly this is the case. We start with a simple observation that reflects the above result.

**Lemma 1.** Let \( X \) (1) be the maximal self-complementary circular trinucleotide code found in genes. Then

- \( X \cup \{ CG \} \) is a non-circular code since the word \( CG|GCC|GCG \) has a second decomposition \( GGC|CG|GCC \) on the circle.
- \( X \cup \{ TA \} \) is a non-circular code since the word \( TA|ATT|AAT \) has a second decomposition \( AAT|TA|ATT \) on the circle.

Forced by Lemma 1, we are left with the two dinucleotides \( GC \) and \( AT \) to be added to \( X \). In order to preserve self-complementarity and to obtain a maximal circular set, there are exactly 4 possible maximal self-complementary dinucleotide codes containing \{ AT, GC \}:

\[
D_1 = \{ AT, GC, AC, GT, AG, CT \}, \quad D_2 = \{ AT, GC, AC, GT, GA, TC \}, \quad D_3 = \{ AT, GC, AG, CT, CA, TG \}, \quad D_4 = \{ AT, GC, CA, TG, GA, TC \}.
\]

The codes \( D_1 \) and \( D_2 \) are non-circular, thus \( X \cup D_1 \) and \( X \cup D_4 \) are non-circular. However, using the dinucleotide circular codes \( D_1 \) and \( D_2 \) we can now state the main theorem.

**Theorem 7.** Let \( X \) (1) be the maximal self-complementary circular trinucleotide code found in genes. Then

\[
(1) \quad X \cup D_1 \text{ is a non-circular set of maximal size 26 that is not a code. In fact, there are 90 paths starting and ending with a green vertex and there are 2 cycles.}
\]
Fig. 15. The circular mixed set $X \cup D_2$ where $X$ (1) is the maximal self-complementary circular trinucleotide code found in genes and $D_2 = \{AT, GC, GT, GA, TC\}$ is a maximal self-complementary circular dinucleotide code.

(2) $X \cup D_2$ is a circular set of maximal size 26 that is not a code. In fact, there are 106 paths starting and ending with a green vertex and no cycle (Fig. 15).

The circular mixed set $X \cup D_2$ has interesting biological properties that should be investigated in the future. In particular, the two dinucleotides $AT$ and $GA$ of $X \cup D_2$ allow to generate the start codon $ATG$ and the stop codon $TGA$ that do not belong to $X$, but to the current genetic code. It is important to stress that it is a specific property of the circular mixed set $X \cup D_2$ and not of the circular mixed code $X \cup \{AT, GC\}$ that does not contain $GA$.

**Remark 1.** The circular mixed set $X \cup D_2$ can generate the word $AT|GA|AT$ that can be read by the two trinucleotides $ATG|AAT$ where the start trinucleotide $ATG \notin X$ and $AAT \in X$. The circular mixed set $X \cup D_2$ can generate the word $GA|AT|GA$, a replacement of $AT$ by $GA$ and conversely, that can be read by the two trinucleotides $GAA|TGA$ where $GAA \in X$ and the stop trinucleotide $TGA \notin X$.

At this point we would like to remark that for both sets $X \cup D_1$ and $X \cup D_2$ in Theorem 7, infinitely many ambiguous words can be constructed although there are only finitely many paths starting and ending with a green vertex. One can simply concatenate words that have two different decompositions of the given code. However, there is still an important difference: the non-circular set $X \cup D_2$ has a green vertex that is part of one of the two cycles. This allows to concatenate words that are unambiguous but its concatenation is ambiguous. A theoretical property that is worth to be investigated in future work.

Based on these results, in particular Theorem 7, a new evolution model of the genetic code can be proposed that can explain the transition from a dinucleotide world to a trinucleotide world (Fig. 16). As a biological summary, a circular mixed set allows several decompositions in the reading frame, but without decomposition in the shifted frame, contrary to a circular mixed code where the decomposition in the reading frame is unique, and obviously without decomposition in the shifted frame. From an evolutionary point of view, these two mathematical concepts may lead to a model where mixed sets are older and flexible compared to mixed codes that are more recent and constrained (Fig. 16). For example, the word $ATGTACTCGC$ has two decompositions in the reading frame and no decomposition in the shifted frame with the circular mixed set $X \cup D_2$: $AT|GT|AC|TC|GC$ and $AT|GA|TC|GC$, while the circular mixed code $X \cup \{AT, GC\}$ has a unique decomposition in the reading frame, here the second one.

6. Conclusion

We have developed a new mathematical concept of circular mixed sets of words over an arbitrary finite alphabet, which to our knowledge has never been proposed. Circular mixed sets allow a higher amount of information to be encoded in the construction frame compared to the classical circular mixed codes. By generalizing a recent graph theoretical approach for circularity with a colouring, Theorem 5 allows to identify the codes among the mixed sets. An algorithm is proposed for determining whether a mixed set is a code. We also describe some constructions of circular mixed sets. Throughout the article, several
examples of graphs are given to illustrate the concept and the different propositions. Finally, circular mixed sets of words can provide the basis for a mathematical structure for a new evolution model of the present genetic code that could have evolved from a dinucleotide world to a trinucleotide word.

Declaration of competing interest

The authors report no conflict of interest.

Data availability

No data was used for the research described in the article.

References


