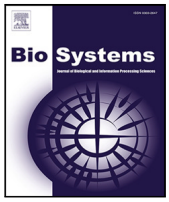
BioSystems 229 (2023) 104906

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems



Circular mixed sets☆

Elena Fimmel [a], Christian J. Michel [b,*], Lutz Strüngmann [a]

[a] Institute of Mathematical Biology, Faculty for Computer Sciences, Mannheim University of Applied Sciences, 68163 Mannheim, Germany
[b] Theoretical bioinformatics, ICube, University of Strasbourg, C.N.R.S., 300 Boulevard Sébastien Brant, 67400 Illkirch, France

ARTICLE INFO

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

ABSTRACT

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 two 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 ℓ-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

2. Definitions and examples

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

Definition 1.

(i) A set $X \subseteq \Sigma^+$ is a *code* if every word $w \in X^+$ has a single decomposition into words of X.
(ii) For $\ell \in \mathbf{N}$ with $\ell \geq 2$, an *ℓ-letter code* is a subset of Σ^ℓ and the elements of Σ^ℓ are called *ℓ-letter words*.

Obviously, any set $X \subseteq \Sigma^\ell$ of ℓ-letter words is a code which justifies (ii) of Definition 1. However, the set $X = \{1011, 10, 11\}$ is not a code over the binary alphabet $\Sigma = \{0, 1\}$ since the word $1011 \in X$ has a second decomposition, namely $1011 = 10 \mid 11$ 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 $X \subseteq \Sigma^+$ be a set. A graph $\mathcal{G} = \mathcal{G}(X) = (V(X), E(X))$ is associated with X in the following way:

(1) The set of *vertices* $V(X)$ of $\mathcal{G}$ is given by

$$V(X) = \{w \in \Sigma^+ \mid v = ww' or v = w'w \ for \ some \ word \ w' \in \Sigma^+ \ and \ v \in X\},$$

i.e. the set of all proper prefixes and proper suffixes of words from X;
(2) The set of *directed edges* $E(X)$ of $\mathcal{G}$ is given by

$$E(X) = \{(w, w') \ where \ ww' \in X \ and \ w, w' \in V(X)\}.$$

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

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

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

Example 2. The mixed set $X = \{b, d, abc, abcd\}$ on the alphabet $\Sigma = \{a, b, c, d\}$ is not a code as the word $abcd$ 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).

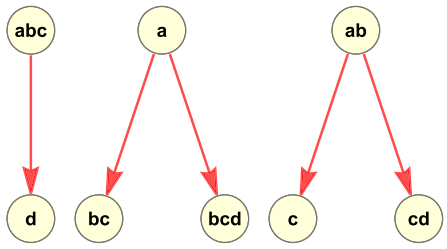


Fig. 1. (Example 2). The acyclic graph of the mixed set $X = \{b, d, abc, abcd\}$ on the alphabet $\Sigma = \{a, b, c, d\}$ that is not a code.

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 $a \to bc \to d \to a$ (Fig. 2).

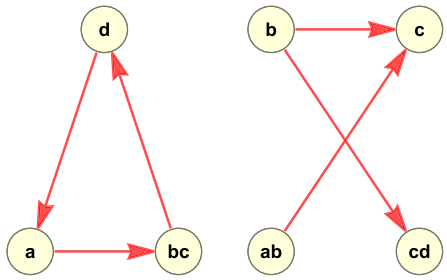


Fig. 2. (Example 3). The cyclic graph of the mixed set $X = \{a, bc, da, abc, bcd\}$ on the alphabet $\Sigma = \{a, b, c, d\}$ that is not a code.

In analogy to Fimmel et al. (2016), we now define comma-freeness and circularity of sets $X \subseteq \Sigma^+$ 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 $\mathcal{G}$ is called *acyclic* if there is no cycle in $\mathcal{G}$ (Diestel, 2005 for more details on graph theory).

Definition 3. A set $X \subseteq \Sigma^+$ 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 $\mathcal{G}(X)$ is at most 2.
(3) *strong comma-free* if X is circular and the maximal length of a path in its associated graph $\mathcal{G}(X)$ is 1.

Clearly, the class of circular sets is closed under intersection and subset relation. Moreover, in the case of codes with word length ≥ 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 $X \subseteq \Sigma$ which is a code with trivial associated graph $\mathcal{G}(X)$, 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 $X \subseteq \Sigma^{\geq 2}$ 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 $X \subseteq \Sigma^{\geq 2}$ is*

(1) strong comma-free if no element of Σ^+ appears both as a prefix and a suffix in X;
(2) comma-free if every concatenation c_1c_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 $m \in \mathbf{N}$, 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

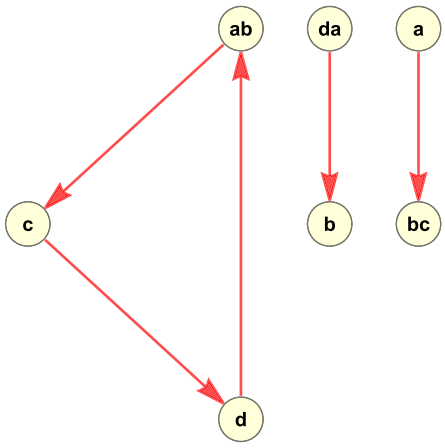


Fig. 3. (Example 4). The cyclic graph of the mixed set $X = \{cd, abc, dab\}$ on the alphabet $\Sigma = \{a, b, c, d\}$ that is a code and not circular.

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 Σ 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 $X_3 = \{abc, dab\}$ 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 $dabcdabc$ admits the 2 circular decompositions $dab|cd|abc$ and $d|abc|dab|c$. The graph associated with X contains the cycle $c \to d \to ab \to c$ (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).

Example 6. Consider the 2 circular codes $X_2 = \{ab\}$ and $X_3 = \{aba, bcd, bdc, cab, cdd, dbd\}$ on the alphabet $\Sigma = \{a, b, c, d\}$. The mixed set $X = \{ab, aba, bcd, bdc, cab, cdd, dbd\}$ is not a code as the word $ababcddbdcab$ has 2 different decompositions into words of X, namely $ab|ab|cdd|bdc|ab$ and $aba|bcd|dbd|cab$. Moreover, the graph associated with X contains the cycle $ab \to a \to b \to cd \to d \to bd \to c \to ab$, thus X is also not circular (Fig. 5).

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 ℓ-letter codes.

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

Proof. Let $X \subseteq \Sigma^+$ be as stated. Put $X_\ell = X \cap \Sigma^\ell$ for any $\ell \in \mathbf{N}$. Then X_ℓ is an ℓ-letter code (since it contains words of length ℓ only). Moreover, X_ℓ is circular as a subset of a circular set. □

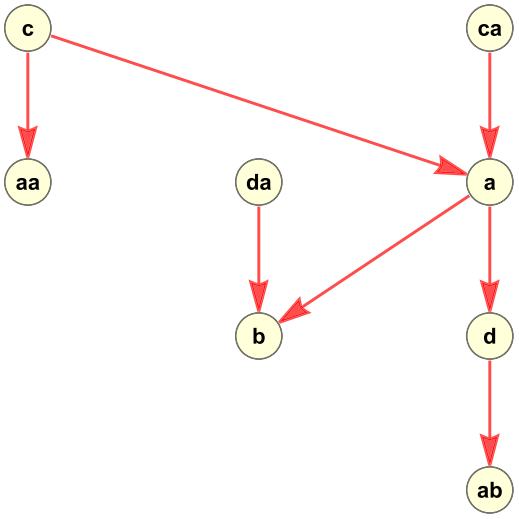


Fig. 4. (Example 5). The acyclic graph of the mixed set $X = \{ab, ad, ca, caa, dab\}$ on the alphabet $\Sigma = \{a, b, c, d\}$ that is not a code but circular.

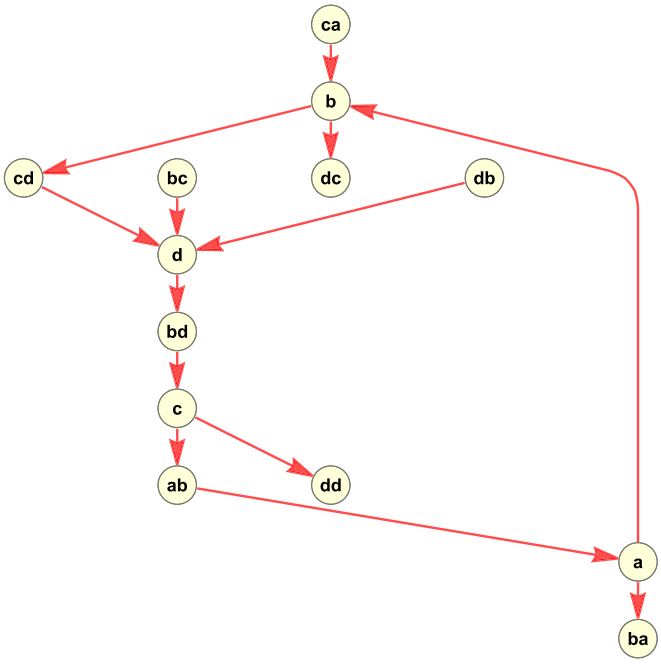


Fig. 5. (Example 6). The cyclic graph of the mixed set $X = \{ab, aba, bcd, bdc, cab, cdd, dbd\}$ on the alphabet $\Sigma = \{a, b, c, d\}$ that is not a code and not circular.

Theorem 2 above already indicates upper bounds for the cardinality of a circular mixed set. In fact, if $s(n, \ell)$ is the maximal size of a circular code $X \subseteq \Sigma^\ell$ where Σ is an alphabet of size n, then a circular set $X \subseteq \Sigma^{\leq n}$ cannot exceed $\sum_{i=1}^{n} s(n, i)$. For instance, a circular mixed set $X \subseteq \Sigma^2 \cup \Sigma^3$ can have cardinality at most 26 if the size of Σ is 4. However, it is still to be shown how circular ℓ-letter codes for various ℓ can be combined to circular mixed sets. We will give some construction methods in Section 4 after recalling some results and construction methods for obtaining circular mixed codes of maximal size obtained by Pirot, Sereni and the authors (Fimmel et al., 2019).

Theorem 3 (*Theorem 3 Fimmel et al., 2019*). *Let $\Sigma = \{a, b, c, d\}$ be a finite alphabet and $X_2 \subseteq \Sigma^2$ be a circular diletter code of maximal size 6. Then X_2 can be embedded into a circular code $X \subseteq \Sigma^2 \cup \Sigma^3$ of maximal*

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* $X_3 = \{aab, aac, aad, bab, bac, bad, bbc, bbd, cab, cac, cad, cbc, cbd, ccd, dab, dac, dad, dbc, dbd, dcd\}$.

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

Theorem 4 (*Theorem 4 Fimmel et al., 2019*)**.** *Let* Σ *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 $X \subseteq \Sigma^2 \cup \Sigma^3 \cup \Sigma^4$. 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 $\mathcal{G}(X)$ the graph associated with X. We define a colouring on $\mathcal{G}(X)$ as follows:

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

$$c : V(\mathcal{G}(X)) \to \{green, yellow\}$$

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

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.

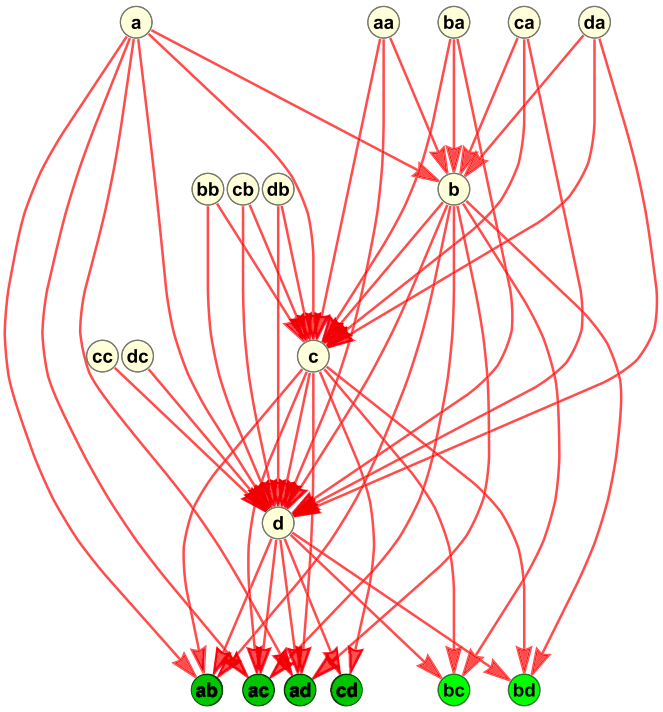


Fig. 6. (Example 7). The acyclic graph of the code constructed in Theorem 3 on the alphabet $\Sigma = \{a, b, c, d\}$.

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

Proof. (1). Let $w_1 \to \cdots \to w_n$ be a path in $\mathcal{G}(X)$ 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 $\mathcal{G}(X)$, $w_1w_2 \in X, \ldots, w_{n-1}w_n \in X$ and $w_2w_3 \in X, \ldots, w_{n-2}w_{n-1} \in X$. Thus the word $w_1 \cdots w_n$ has 2 decompositions over X, namely

  $$w_1w_2 \cdot w_3w_4 \cdot .... \cdot w_{n-1}w_n$$

  and

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

  Hence X is not a code.

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

  $$w_1w_2 \cdot w_3w_4 \cdot .... \cdot w_{n-2}w_{n-1} \cdot w_n$$

  and

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

  Hence X is not a code.

(2). Conversely, now assume that X is not a code, hence there is a (shortest) concatenation of words $w_1, \ldots, w_k \in X$ such that $w_1 \cdot w_2 \cdot .... \cdot w_n$ has a second decomposition over X, let say $v_1 \cdot v_2 \cdot .... \cdot v_m$ 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 $v_1 \cdots v_s$ is shorter than w_1. Without loss of generality, we may assume that $s = 1$, implying $v_1 \in X^+$. Hence $w_1 = v_1 \cdot v_1'$ for some $v_1' \in \Sigma^+$. Consequently, v_1' is now a prefix of v_2 and thus $v_2 = v_1'w_2'$ and, as above without loss of generality, w_2' is a prefix of w_2, i.e. $w_2 = w_2'v_2'$. Again, v_2' is now a prefix of v_3, i.e. $v_3 = v_2'w_3'$. We continue this way and end up with the path

$$v_1 \to v_1' \to w_2' \to v_2' \to w_3' \to \cdots$$

The path ends with either v_m' or w_n' 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* Σ *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* $\tilde{X} := X_3 \cup X_4$.

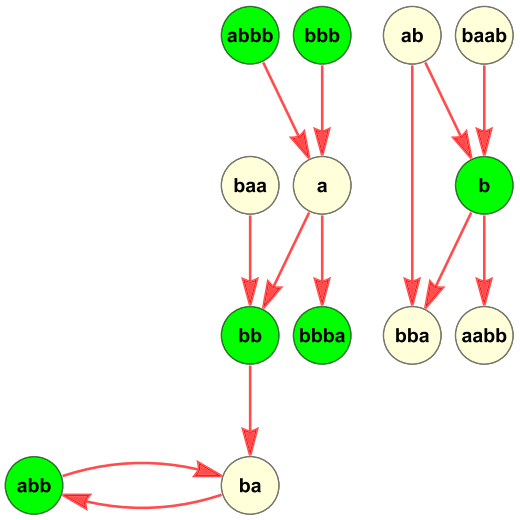


Fig. 7. (Example 8). The graph of the mixed set $X = \{b, abb, bbba, abbba, baabb\}$ on the alphabet $\Sigma = \{a, b\}$ that is not a code since it contains the path, for example, $bb \to ba \to abb$.

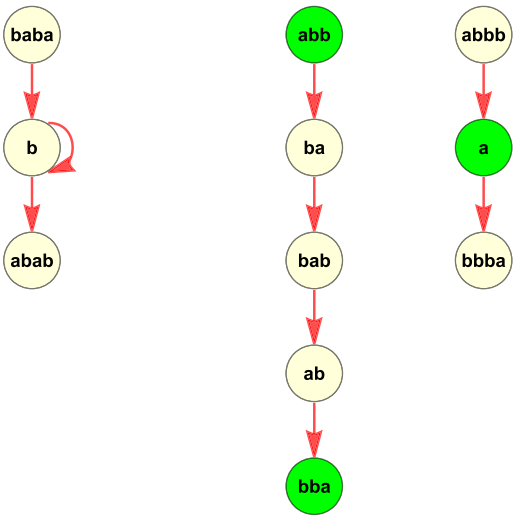


Fig. 8. (Example 9). The graph of the mixed set $X = \{a, bb, abbba, babab\}$ on the alphabet $\Sigma = \{a, b\}$ that is not a code since it contains the path $abb \to ba \to bab \to ab \to bba$.

(i) The mixed set X is a code if and only if there exists no directed path in $\mathcal{G}(X)$ 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 $\mathcal{G}(\tilde{X})$ 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 $\mathcal{G}(X)$ and $\mathcal{G}(\tilde{X})$ are acyclic.

Proof. Clearly, a vertex in $\mathcal{G}(X)$ 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_3. 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 Σ 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 $\mathcal{G}(X)$ associated with X.
- Calculate the set $V'(\mathcal{G}(X))$ with $V'(\mathcal{G}(X)) = \{v \in V(\mathcal{G}(X)) : v \in X^+\}$.
- Colour the graph $\mathcal{G}(X)$ 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 $X = \{b, abb, bbba, abbba, baabb\}$. Then X is not a code as its associated graph $\mathcal{G}(X)$ contains several directed paths that start and end with a green node. For example, the path $bb \to ba \to abb$ gives the word $bbbaabb$ with 2 different decompositions $bbba|abb = b|b|baabb$. Accordingly, X is not a code (Fig. 7).

In fact, all ambiguous words are given by the following regular expression:

$$abbbabbba + bbbabbba + abbbabb(baabb)^* + bbbabb(baabb)^*.$$

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 $X = \{a, bb, abbba, babab\}$. The graph $\mathcal{G}(X)$ 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 $abbbababbbba$ with 2 different decompositions $a|bb|babab|abbba = abbba|babab|bb|a$. 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 $\ell, n \in \mathbf{N}, \ell \neq n$, $X_1 \subseteq \Sigma^\ell$, $X_2 \subseteq \Sigma^n$ 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 $\mathcal{G}(X_1 \cup X_2)$ that can be only of even length and containing alternating edges of $\mathcal{G}(X_1)$ and $\mathcal{G}(X_2)$, 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:

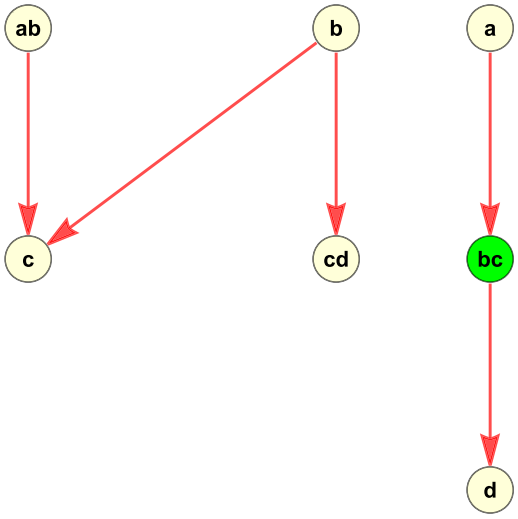


Fig. 9. (Example 10). The graph of the circular mixed code $X = X_1 \cup X_2$ where $X_1 = \{abc, bcd\}$ is not strong comma-free but $X_2 = \{bc\}$ is strong comma-free on the alphabet $\Sigma = \{a, b, c, d\}$.

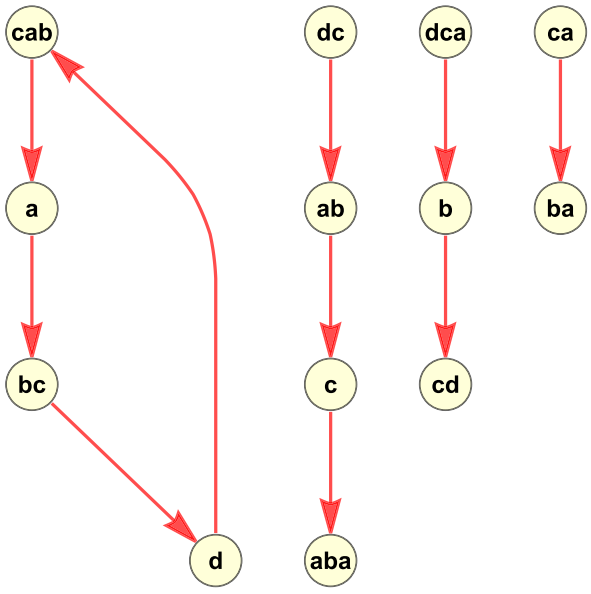


Fig. 10. (Example 11). The graph of the non-circular mixed code $X = X_1 \cup X_2$ where $X_1 = \{abc, bcd\}$ and $X_2 = \{caba, dcab\}$ are both comma-free on the alphabet $\Sigma = \{a, b, c, d\}$.

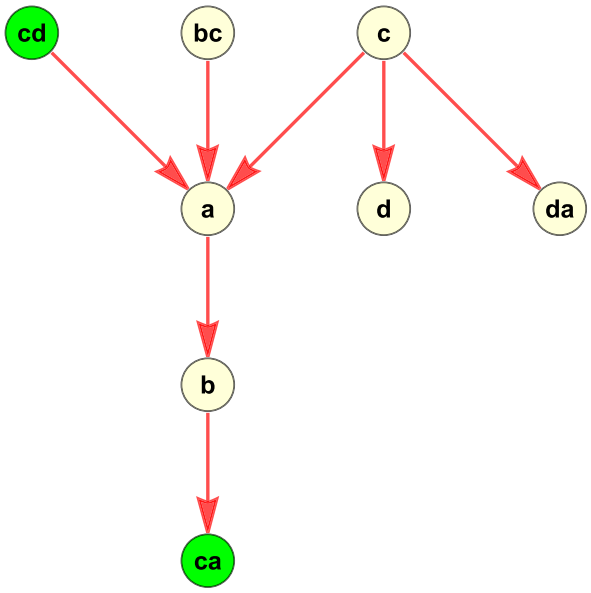


Fig. 11. (Example 12). The graph of the circular mixed set $X = X_1 \cup X_2$ where $X_1 = \{ab, ca, cd\}$ and $X_2 = \{bca, cda\}$ are both strong comma-free codes on the alphabet $\Sigma = \{a, b, c, d\}$. However, X is not a mixed code since it contains the path $cd \to a \to b \to ca$.

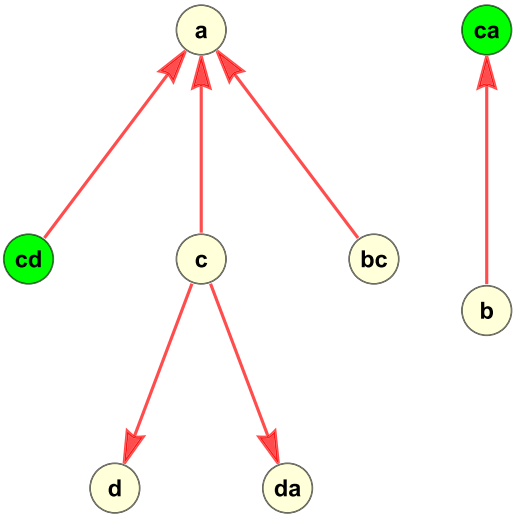


Fig. 12. The graph of the circular mixed code $X = X_1 \cup X_2$ where $X_1 = \{ca, cd\}$ and $X_2 = \{bca, cda\}$ are both strong comma-free codes on the alphabet $\Sigma = \{a, b, c, d\}$.

Example 11. Let $\Sigma = \{a, b, c, d\}$ and $X_1 = \{abc, bcd\}$, $X_2 = \{caba, dcab\}$. 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 $w = abcdcab$ has 2 decompositions into words of X on the circle: $abc|dcab = bcd|caba$ (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 $\mathcal{G}(X_1 \cup X_2)$ contains a directed path $cd \to a \to b \to ca$ which starts and ends with a green vertex. In other words, the word $cdabca$ has 2 decompositions $cda|bca = cd|ab|ca$ 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 $\ell, n \in \mathbf{N}, \ell \neq n$, $X_1 \subseteq \Sigma^\ell$, $X_2 \subseteq \Sigma^n$ 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 $\mathcal{G}(X_1 \cup X_2)$. The circle C must contain edges from both $\mathcal{G}(X_1)$ and $\mathcal{G}(X_2)$ 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.

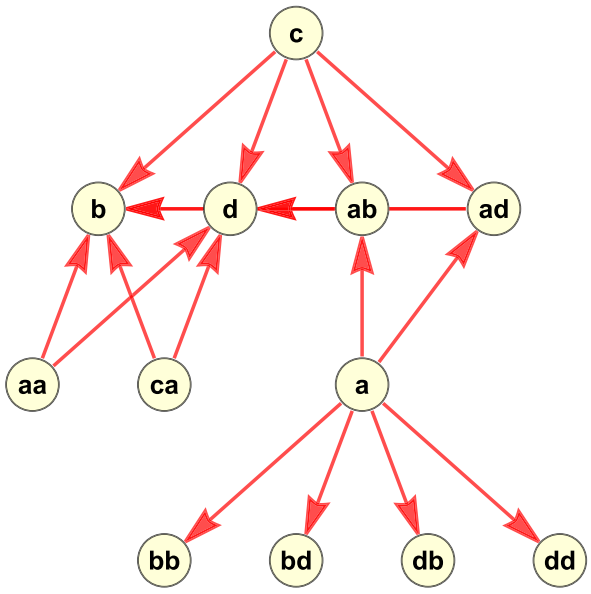


Fig. 13. (Example 13). The graph of the circular mixed code $X = X_1 \cup X_2$ where $X_1 = \{cb, cd\}$ and $X_2 = \{aab, aad, abb, abd, adb, add, cab, cad\}$ on the alphabet $\Sigma = \{a, b, c, d\}$.

(2). Let us show that $X_1 \cup X_2$ is a code. Let us assume the opposite. Without lost of generality, let $\ell < n$ also hold. Then according to Theorem 5, there is a directed path in $\mathcal{G}(X_1 \cup X_2)$ which begins and ends with a 'green' vertex: $w_1 \to \ldots \to w_k$, with $w_1, w_k \in (X_1 \cup X_2)^+$. Since we only have words of two different lengths ℓ and n in $X_1 \cup X_2$, w_k can only be a concatenation of the words from X_1: $w_k = v_1 \ldots v_m$, $v_i \in X_1$. Then v_m is both suffix of a word from X_2 and prefix of a word from X_1 (itself). Contradiction to the assumption. □

The above condition is sufficient but not necessary, as the above Example 10 shows: bc is at the same time a prefix (suffix) of $bcd(abc) \in X_1$ and a prefix (suffix) of $bc \in X_2$ but $X_1 \cup X_2$ is a circular mixed code.

For the convenience of the reader, we give an example for Proposition 2 that has a biological motivation.

Example 13. Let $\Sigma = \{a, b, c, d\}$ and $X_1 = \{cb, cd\}$, $X_2 = \{aab, aad, abb, abd, adb, add, cab, cad\}$. According to Proposition 2, $X_1 \cup X_2$ is a circular mixed code[1] (Fig. 13).

We close this section with a result that allows to combine diletter codes and tetraletter codes over an alphabet Σ of even size. Recall that the *symmetric group* S_Σ is defined as the group of all bijective mappings (permutations) $\pi : \Sigma \to \Sigma$. Let V_Σ be the subset of the symmetric group S_Σ consisting of all permutations of order 2 without fixpoints. Let $Tess(\Sigma) = \{N_1 N_2 \alpha(N_1)\alpha(N_2) : N_1, N_2 \in \Sigma, \alpha \in V_\Sigma\}$. In a biological context this set $Tess(\Sigma)$ has been studied extensively to model parts of the evolution of the genetic code and it is known that the maximal size of a circular code in $Tess(\Sigma)$ is 12 (Fimmel et al., 2020b for details).

Theorem 6. *Let Σ be a finite alphabet of even size and let $D \subseteq \Sigma^2$ be a circular code. Moreover, let $X' \subseteq Tess(\Sigma)$ be any circular code. Then the union $X = D \cup X'$ is a circular mixed set.*

[1] If we replace the letters a and c with A and G (purines) and b and d with C and T (pyrimidines), the code X_2 represents one half of the so-called RNY-code (purine as first base and pyrimidine as third base). The RNY-code is considered as a possible predecessor of the modern genetic code. It is self-complementary and comma-free, even in all three frames (e.g. Eigen and Schuster, 1978; Shepherd, 1981; Arquès and Michel, 1996; Fimmel et al., 2017b; Fimmel and Strüngmann, 2016).

Proof. Let D and X' be given. We claim that $X = D \cup X'$ is a circular mixed set, i.e. the associated graph $\mathcal{G}(D \cup X') = \mathcal{G}(X) = \mathcal{G}(D) \cup \mathcal{G}(X')$ is acyclic. Certainly, it is not a code if D contains two diletter words of the form $N_1 N_2$ and $\alpha(N_1 N_2)$ for some $\alpha \in V_\Sigma$ and $N_1 N_2 \alpha(N_1 N_2) \in X'$. Assume that $\mathcal{G}(X)$ contains a cycle $v_1 \to \cdots \to v_n \to v_1$. This can only happen if either all v_i are diletters or if all $v_i \in \Sigma \cup \Sigma^3$, i.e. they are either letters or triletters. In the first case, we obtain a contradiction since then the cycle would belong to $\mathcal{G}(X')$ but X' is circular. Hence assume all $v_i \in \Sigma \cup \Sigma^3$. If all $v_i \in \Sigma$ then there is a contradiction with the circularity of D since then the cycle belongs to $\mathcal{G}(D)$. Thus there is a subpath of the form $N \to N_1 N_2 N_3 \to N'$. However, this means that $N N_1 N_2 N_3$ and $N_1 N_2 N_3 N'$ are in X'. Therefore $N_2 N_3 = \alpha(N N_1)$ for some $\alpha \in V_\Sigma$ and $N_3 N' = \alpha'(N_1 N_2)$ for some $\alpha' \in V_\Sigma$. Consequently, $\alpha(N_1) = N_3 = \alpha'(N_1)$ which means that $\alpha = \alpha'$ and so $N = \alpha(N_2) = \alpha'(N_2) = N'$, in contradiction with the circularity of X'. □

Again, we close with an example.

Example 14. Let $\Sigma = \{a, b, c, d\}$ and $D = \{ba, ca, cb, cd, da, db\}$ be a circular diletter code of maximal size 6,

$$X = \{aabb, aacc, aadd, adcb, bbdd, bcad, ccbb, ccdd, cdba, dabc, dacb, dcab\} \subseteq Tess(\Sigma)$$

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 anticodon 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:

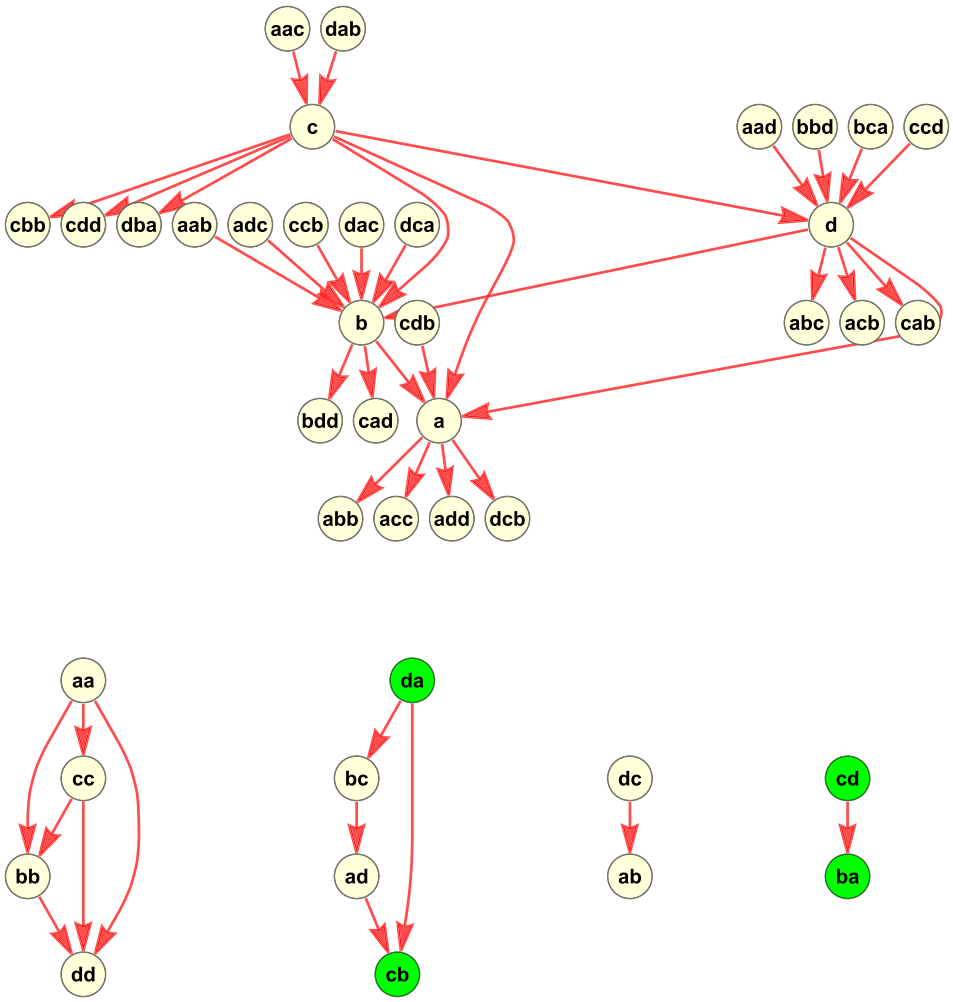


Fig. 14. (Example 14). The circular mixed set $\{ba, ca, cb, cd, da, db, aabb, aacc, aadd, adcb, bbdd, bcad, ccbb, ccdd, cdba, dabc, dacb, dcab\}$ of diletter and tetraletter words of size 18 on the alphabet $\Sigma = \{a, b, c, d\}$.

(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):

$$X = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}. \quad (1)$$

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 $CG|GCC|GGC$ 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 $\{AT, GC\}$: $D_1 = \{AT, GC, AC, GT, AG, CT\}$, $D_2 = \{AT, GC, AC, GT, GA, TC\}$, $D_3 = \{AT, GC, AG, CT, CA, TG\}$, $D_4 = \{AT, GC, CA, TG, GA, TC\}$.

The codes D_3 and D_4 are non-circular, thus $X \cup D_3$ 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) $X \cup D_1$ 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.

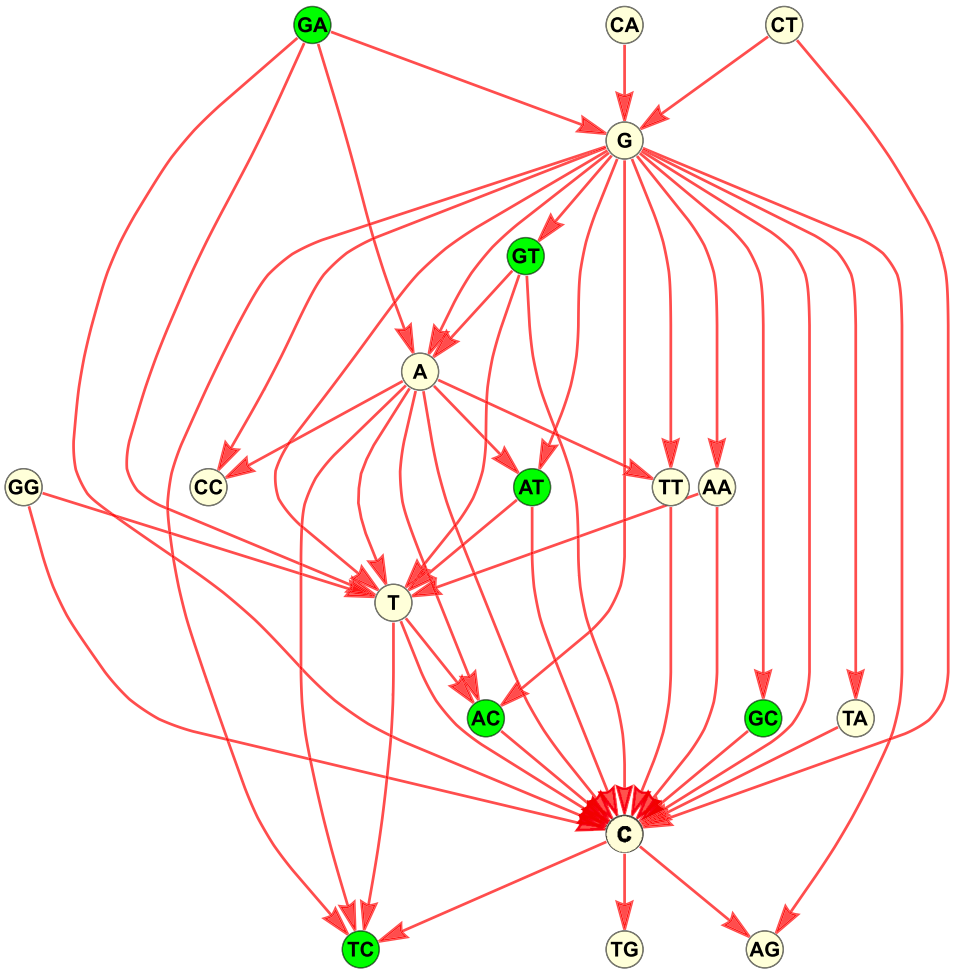


Fig. 15. The circular mixed set $X \cup D_2$ where X (1) is the maximal self-complementary circular trinucleotide code found in genes and $D_2 = \{AT, GC, AC, GT, GA, TC\}$ is a maximal self-complementary circular dinucleotide code.

(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 $X \cup D_1$ 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 $AT|GTA|CTC|GC$, 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

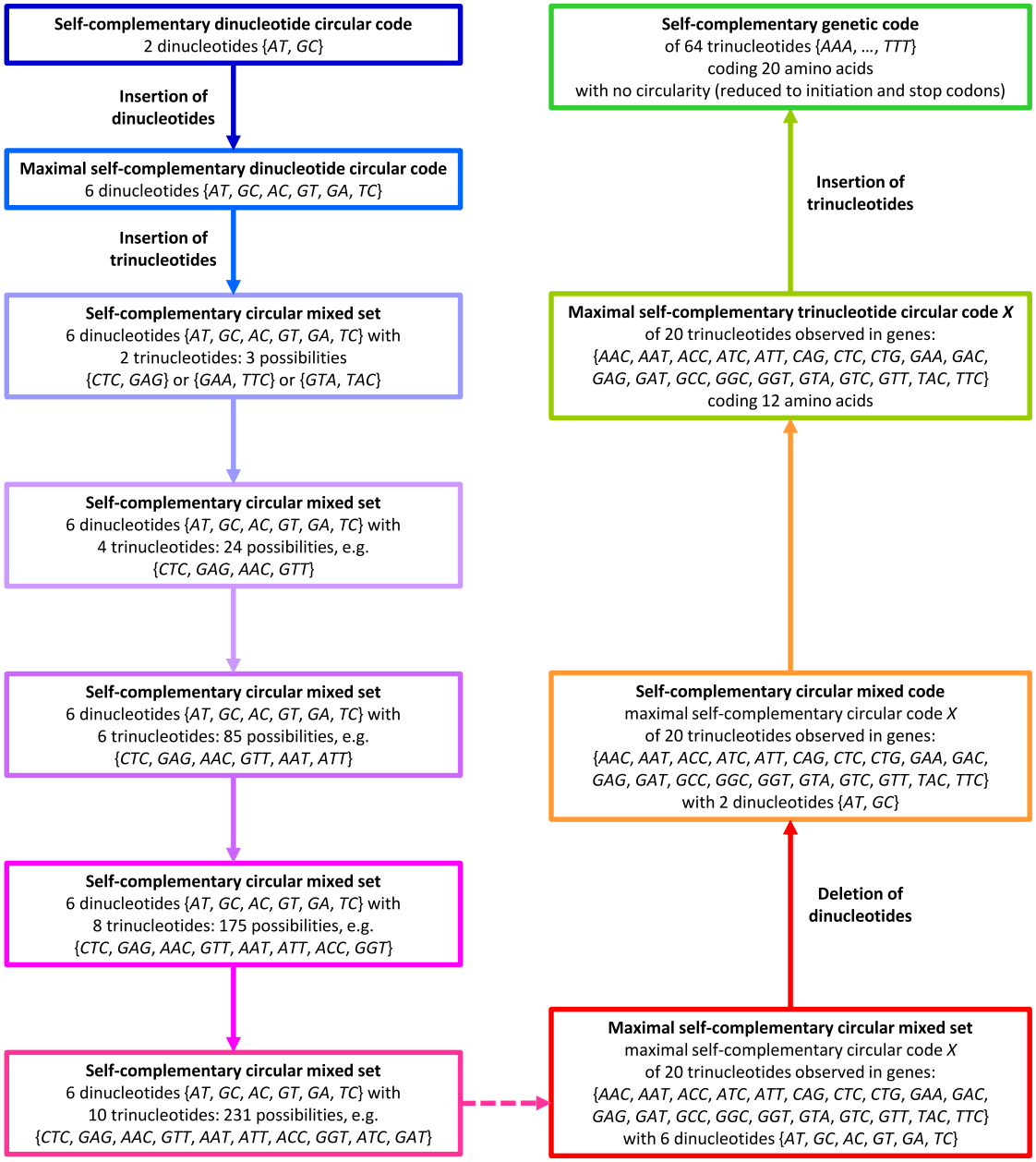


Fig. 16. A new evolution model of the genetic code based on circular mixed sets and on circular mixed codes of dinucleotides and trinucleotides.

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

Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. J. Theoret. Biol. 182, 45–58.

Berstel, J., Perrin, D., Reutenauer, C., 2009. Codes. In: Codes and Automata (Encyclopedia of Mathematics and its Applications). Cambridge University Press, Cambridge, pp. 55–106.

Copley, S.D., Smith, E., Morowitz, H.J., 2005. A mechanism for the association of amino acids with their codons and the origin of the genetic code. Proc. Natl. Acad. Sci. USA 102, 4442–4447.

Diestel, R., 2005. Graph Theory, electronic ed. 2005 Springer-Verlag, Heidelberg, New York, 1997, 2000, 2005.

Dila, G., Ripp, R., Mayer, C., Poch, O., Michel, C.J., Thompson, J.D., 2019. Circular code motifs in the ribosome: a missing link in the evolution of translation? RNA 25, 1714–1730.

Eigen, M., Schuster, P., 1978. The hypercycle. a principle of natural self-organization. Part c: The realistic hypercycle. Naturwissenschaften 65, 341–369.

Fimmel, E., Giannerini, S., Gonzalez, D., Strüngmann, L., 2015. Dinucleotide circular codes and bijective transformations. J. Theoret. Biol. 386, 159–165.

Fimmel, E., Michel, C.J., Pirot, F., Sereni, J.-S., Starman, M., Strüngmann, L., 2020a. The relation between k-circularity and circularity of codes. Bull. Math. Biol. 82, 105, 1–34.

Fimmel, E., Michel, C.J., Pirot, F., Sereni, J.-S., Strüngmann, L., 2019. Mixed circular codes. Math. Biosci. 317, 1–14.

Fimmel, E., Michel, C.J., Strüngmann, L., 2016. *n*-Nucleotide circular codes in graph theory. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 374, 20150058.
Fimmel, E., Michel, C.J., Strüngmann, L., 2017a. Diletter circular codes over finite alphabets. Math. Biosci. 294, 120–129.
Fimmel, E., Michel, C.J., Strüngmann, L., 2017b. Strong comma-free codes in genetic information. Bull. Math. Biol. 79, 1796–1819.
Fimmel, E., Starman, M., Strüngmann, L., 2020b. Circular tessera codes in the evolution of the genetic code. Bull. Math. Biol. 82, 1–25.
Fimmel, E., Strüngmann, L., 2016. Codon distribution in error-detecting circular codes. Life 6, 1–17.
Fimmel, E., Strüngmann, L., 2018. Mathematical fundamentals for the noise immunity of the genetic code. BioSystems 164, 186–198.
Jungck, J.R., 1978. The genetic code as a periodic table. J. Mol. Evol. 11, 211–224.
Michel, C.J., 2008. A 2006 review of circular codes in genes. Comput. Math. Appl. 55, 984–988.
Michel, C.J., 2015. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. J. Theoret. Biol. 380, 156–177.
Michel, C.J., 2017. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, archaea, eukaryotes, plasmids and viruses. Life 7, 1–16.
Michel, C.J., Pirillo, G., 2013. Dinucleotide circular codes. ISRN Biomath. 538631.
Michel, C.J., Thompson, J.D., 2020. Identification of a circular code periodicity in the bacterial ribosome: origin of codon periodicity in genes? RNA Biol. 17, 571–583.
Parvathy, S.T., Udayasuriyan, V., Bhadana, V., 2022. Codon usage bias. Mol. Biol. Rep. 49, 539–565.
Rodin, A.S., Szathmáry, E., Rodin, S.N., 2011. On origin of genetic code and tRNA before translation. Biol. Direct 6, 14, 1–24.
Shepherd, J.C.W., 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci. USA 78, 1596–1600.
Thompson, J.D., Ripp, R., Mayer, C., Poch, O., Michel, C.J., 2021. Potential role of the X circular code in the regulation of gene expression. Biosystems 203, 104368, 1–15.
Yaman, T., Harvey, J.N., 2021. Computational analysis of a prebiotic amino acid synthesis with reference to extant codon–amino acid relationships. Life 11, 1343, 1–16.