Research Article

Dinucleotide Circular Codes

Christian J. Michel\(^1\) and Giuseppe Pirillo\(^2,3\)

\(^1\) Equipe de Bioinformatique Théorique, BFO, LSIT, UMR 7005, Université de Strasbourg, Pôle API, 300 Boulevard Sébastien Brant, 67400 Illkirch, France
\(^2\) Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Consiglio Nazionale delle Ricerche and Dipartimento di Matematica, "Ulisse Dini" Università di Firenze, Viale Morgagni 67/A, 50134 Firenze, Italy
\(^3\) Université de Marne-la-Vallée, 5 boulevard Descartes, 77454 Marne-la-Vallée Cedex 2, France

Correspondence should be addressed to Christian J. Michel; michel@dpt-info.u-strasbg.fr

Received 12 October 2012; Accepted 12 December 2012

Academic Editors: J. Chow, M. Jose, M. R. Roussel, and J. H. Wu

Copyright © 2013 C. J. Michel and G. Pirillo. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We begin here a combinatorial study of dinucleotide circular codes. A word written on a circle is called circular. A set of dinucleotides is a circular code if all circular words constructed with this set have a unique decomposition. Propositions based on a letter necklace allow to determine the 24 maximum dinucleotide circular codes (of 6 elements). A partition property is also identified with eight self-complementary maximum dinucleotide circular codes and two classes of eight maximum dinucleotide circular codes in bijective correspondence by the complementarity map.

1. Introduction

We continue our study of the combinatorial properties of circular codes in genes, that is, on the nucleotide alphabet \(A^2 = \{A, C, G, T\}\). A dinucleotide is a word of two letters (diletter) on \(A^2\). A trinucleotide is a word of three letters (triletter) on \(A^4\). The two sets of 16 dinucleotides and 64 trinucleotides are codes in the sense of language theory but not circular codes [1, 2]. In order to have an intuitive meaning of these notions, codes are written on a straight line, while circular codes are written on a circle, but, in both cases, unique decipherability is required.

Trinucleotide comma-free codes, a very particular case of trinucleotide circular codes, have been studied for a long time, see for example, [3–5]. After the discovery of a trinucleotide circular code in genes with strong mathematical properties [6], circular codes are mathematical objects studied in combinatorics, theoretical computer science, and theoretical biology. This theory underwent a rapid development, see for example, [7–27].

Trinucleotides are the fundamental words for genes, that is, the DNA sequences coding the amino acids constituting the protein sequences. However, dinucleotides are also words with important biological functions in genomes. Dinucleotides are involved in some genome sites, for example, the splice sites of introns in eukaryotic genomes are based on the dinucleotides \(GT\) and \(AT\) [28, 29]. Dinucleotides are also involved in some genome regions, for example, the dinucleotide \(CG\) in animal and plant genomes allows a positive or negative control over gene expression [30], and the dinucleotides \(CA\) [31, 32], \(CT\) [33], and \(TG\) [34] in eukaryotic genomes occur as concatenated words \((l_1 l_2)^+\), \(l_1, l_2 \in A^2\) (called tandem repeats in biology).

We begin here a new combinatorial study concerning the dinucleotide circular codes. Their number, their list, and a partition according to the complementarity map are determined with propositions based on a letter necklace.

2. Preliminaries

The following definitions and propositions are classical for any finite set of words on any finite alphabet [1]. We recall them for dinucleotides, that is, words of length 2 on a 4-letter alphabet. Let \(A^2 = \{A, C, G, T\}\) denote the genetic alphabet, lexicographically ordered by \(A < C < G < T\). The set
of nonempty words (resp., words) on $\mathcal{A}_4$ is denoted by $\mathcal{A}_4^*$ (resp., $\mathcal{A}_4^+$). The set of the 16 words of length 2 (dinucleotides or diletters) over $\mathcal{A}_4$ is denoted by $\mathcal{A}_4^2$. The set of the 64 words of length 3 (trinucleotides or trileters) over $\mathcal{A}_4$ is denoted by $\mathcal{A}_4^3$.

Definition 1. A set $X$ of words in $\mathcal{A}_4^2$ is a dinucleotide code if, for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $n, m \geq 1$, the condition $x_1 \cdots x_n = x'_1 \cdots x'_m$ implies $n = m$ and $x_i = x'_i$ for $i = 1, \ldots, n$.

Dinucleotide codes are read on a straight line.

Definition 2. A dinucleotide code $X$ in $\mathcal{A}_4^2$ is circular if, for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $n, m \geq 1$, $p \in \mathcal{A}_4^*$, $s \in \mathcal{A}_4^2$, the conditions $sx_1 \cdots x_n = psx'_1 \cdots x'_m$ and $x_1 = ps$ imply $n = m$, $p = \varepsilon$ (empty word), and $x_i = x'_i$ for $i = 1, \ldots, n$.

Dinucleotide circular codes are read on a circle.

Remark 3. The set $\mathcal{A}_4^2$ is a code but not a circular code.

Definition 4. Two dinucleotides $x$ and $y$ are conjugate if there exist two letters $l_1$ and $l_2, l_1, l_2 \in \mathcal{A}_4$, such that $x = l_1l_2$ and $y = l_2l_1$.

Proposition 5 (see [1]). A dinucleotide circular code cannot contain a word of the form $u^2$ with $u \neq \varepsilon$.

The periodic dinucleotides AA, CC, GG, and TT cannot be in a dinucleotide circular code.

Proposition 6 (see [1]). A dinucleotide circular code cannot contain conjugate dinucleotides.

Example 7. The dinucleotides AC and CA cannot be in the same circular code.

The set operations complementarity $\mathcal{C}$, permutation $\mathcal{P}$, and mirror image $\mathcal{M}$ defined later are involutions.

Definition 8. The nucleotide complementarity map $\mathcal{C} : \mathcal{A}_4 \rightarrow \mathcal{A}_4$ is defined by $\mathcal{C}(A) = T$, $\mathcal{C}(T) = A$, $\mathcal{C}(C) = G$, and $\mathcal{C}(G) = C$.

Definition 9. The dinucleotide complementarity map $\mathcal{C} : \mathcal{A}_4^2 \rightarrow \mathcal{A}_4$ is defined by $\mathcal{C}(l_1l_2) = \mathcal{C}(l_2)\mathcal{C}(l_1)$ for all $l_1, l_2 \in \mathcal{A}_4$.

Example 10. $\mathcal{C}(AC) = GT$.

Definition 11. The complementary dinucleotide set $\mathcal{C}(X)$ of a dinucleotide set $X$ is the set obtained by applying the dinucleotide complementarity map $\mathcal{C}$ to all the dinucleotides of $X$.

Remark 12. $\mathcal{C}^2(X) = \mathcal{C}(\mathcal{C}(X)) = X$.

Definition 13. A dinucleotide circular code $X$ is self-complementary if, for each $x \in X$, $\mathcal{C}(x) \in X$.

Definition 14. The (left) dinucleotide circular permutation map $\mathcal{P} : \mathcal{A}_4^2 \rightarrow \mathcal{A}_4^2$ permutes circularly each dinucleotide $l_1l_2 \in \mathcal{A}_4^2$.

Definition 15. The permuted dinucleotide set $\mathcal{P}(X)$ of a dinucleotide set $X$ is the set obtained by applying the circular permutation map $\mathcal{P}$ to all the dinucleotides of $X$.

Remark 16. $\mathcal{P}(X) = \mathcal{P}(\mathcal{P}(X)) = X$.

Definition 17. The mirror image of a dinucleotide $x = l_1l_2$ is $\overline{x} = l_2l_1$, $l_1, l_2 \in \mathcal{A}_4$.

Definition 18. The mirror image $\overline{X}$ of a dinucleotide set $X$ is the set of the mirror images of all the dinucleotides of $X$.

Remark 19. $\overline{X} = X$.

Remark 20. For a dinucleotide $x$ and for a dinucleotide set $X$, we have $\mathcal{P}(x) = \overline{x}$ and $\mathcal{P}(X) = \overline{X}$.

Proposition 21 (see [27]). A dinucleotide code $X$ is circular if and only if the dinucleotide code $\overline{X}$ is circular.

Proposition 22. A dinucleotide code $X$ is circular if and only if the permuted dinucleotide code $\mathcal{P}(X)$ is circular.

Proof. By Proposition 21 and Remark 20. \hfill \Box

Remark 23. Proposition 22 is not true with trinucleotides [6].

3. Results

In this paper, we identify the subsets of $\mathcal{A}_4^2$ which are circular codes. Based on a letter necklace, we prove a necessary and sufficient condition for a set of dinucleotides to be a circular code.

Definition 24. Let $l_1, l_2, l_3, \ldots, l_n, l_{n+1}$ be letters in $\mathcal{A}_4$. One says that the ordered sequence $l_1, l_2, l_3, \ldots, l_n, l_{n+1}$ is a $(n + 1)$-necklace for a subset $X \subseteq \mathcal{A}_4^2$ if each dinucleotide $l_1l_2, l_2l_3, \ldots, l_nl_{n+1}$ belongs to $X$.

Proposition 25. Let $X$ be a subset of $\mathcal{A}_4^2$. The following conditions are equivalent:

(1) $X$ is circular code,

(2) $X$ has no 5-necklace.

Proof. (1) $\rightarrow$ (2). Let $X$ be a circular code. We have to prove that $X$ has no 5-necklace. Suppose, by way of contradiction, that $l_1, l_2, l_3, l_4, l_5$ is a 5-necklace for $X$. As $\mathcal{A}_4$ contains four letters, for some $i, j \in \{1, 2, 3, 4, 5\}$, $i < j$, we have that $l_i = l_j$. Remark that the maximum value of $j - i$ is 4.
(i) If \( j - i = 1 \), then \( X \) has a periodic dinucleotide \( \alpha \alpha \). (Contradiction with Proposition 5.)

(ii) If \( j - i = 2 \), then \( X \) has two conjugated dinucleotides \( \alpha \beta \) and \( \beta \alpha \). (Contradiction with Proposition 6.)

(iii) If \( i = j = 3 \), then either \( l_1 = l_4 \) or \( l_2 = l_5 \).

(iiiia) If \( l_1 = l_4 \), the 5-necklace \( l_1, l_2, l_3, l_4, l_5 \) is \( l_1, l_2, l_3, l_4, l_5 \). So, \( l_1, l_2, l_3, l_4, l_5, l_6 \in X \). Consider the sequence \( x_1, x_2, x_3, x_4, x_5 \) of \( X \). Put \( x_1 = l_1, x_2 = l_1 x_1, x_3 = l_2, x_4 = l_2 x_3, x_5 = l_3 \). Note that \( x_1, x_2, x_3, x_4, x_5 \) belong to \( X \). Now, the following relations hold: \( l_2 x_1 l_3 l_4 l_5 = l_2 \cdot l_3 \cdot l_4 \cdot l_5 \cdot l_1 = l_2 x_2 x_3 x_4 x_5 \). (Contradiction with the assumption that \( X \) is a circular code.)

(iiiib) The case of \( l_2 = l_5 \) is analogous.

(iv) If \( j - i = 4 \), then \( l_1 = l_5 \), and the 5-necklace \( l_1, l_2, l_3, l_4, l_5 \) is \( l_1, l_2, l_3, l_4, l_5 \). So, \( l_1, l_2, l_3, l_4, l_5, l_6 \in X \). Consider the sequence \( x_1, x_2, x_3, x_4, x_5 \) of \( X \). Put \( x_1 = l_1, x_2 = l_2, x_3 = l_3, x_4 = l_4, x_5 = l_5 \). Note that \( x_1, x_2, x_3, x_4, x_5 \) belong to \( X \). Now, the following relations hold: \( l_2 x_1 l_3 l_4 l_5 = l_2 \cdot l_3 \cdot l_4 \cdot l_5 \cdot l_1 = l_2 x_2 x_3 x_4 x_5 \). (Contradiction with the assumption that \( X \) is a circular code.)

(2) \( \rightarrow \) (1) Let \( X \) be without 5-necklace and suppose, by way of contradiction, that \( X \) is not a circular code. As \( X \) is a uniform code, there exist \( x_1, \ldots, x_n, x_1', \ldots, x_n' \in \mathcal{A'} \), such that \( s x_2 \cdots x_n p_1 = x_1' \cdots x_n' \). Moreover, as all the elements of \( X \) have a length of 2, there exist \( p_2, \ldots, p_n, s_2, \ldots, s_n \in \mathcal{A} \), such that \( p_2 s_2 = x_2 \cdots \), \( p_{n-1} s_{n-1} = x_{n-1} \), \( p_n s_n = x_n \), and \( s_1 p_2 = x_1' \). Now, if \( n \geq 3 \), then \( p_1, s_1, p_2, s_2, p_3 \) is a 5-necklace, if \( n = 2 \), then \( p_1, s_1, p_2, s_2, p_1 \) is a 5-necklace, if \( n = 1 \), then \( p_1, s_1, p_1 \) is a 5-necklace.

In any case, there is a 5-necklace. Contradiction.

If a dinucleotide set \( X \) is a circular code, then there exists no word of \( X \) with two different decompositions of their products written on a circle.

Example 26. Consider the set \( X \) containing only the dinucleotides \( AC \) and \( GT \). Let \( x_1, x_2, x_3, \ldots \) be any sequence with \( x_1 = AC \) or \( x_1 = GT \). As \( X \) does not contain \( CA, CG, TA \) and \( TG \), the sequence \( x_1, x_2, x_3, \ldots \) cannot have a double decomposition on a circle. The set \( X \) has no 5-necklace as, if \( l_1 \) is \( C \) (resp., \( T \)) then \( l_1 \) must be \( A \) (resp., \( G \)), but \( A \) (resp., \( G \)) is never a suffix in \( X \). There are sets with six dinucleotides which are circular codes. For example, any sequence with the set of dinucleotides \( x_1 = AC, x_2 = AG, x_3 = AT, x_1' = CG, x_2' = CT, x_3' = GT \) has no double decomposition on a circle.

More generally, write the sequence \( l_1, l_2, l_3, l_4, l_5, l_6 \) of \( \mathcal{A} \) on a circle. If \( x_1 = l_1 l_2, x_2 = l_3 l_4, x_3 = l_5 l_6, x_1' = l_2 l_3, x_2' = l_4 l_5, x_3' = l_6 l_1 \) belong to a set \( Y \), then \( Y \) cannot be a circular code because the sequence \( l_1 l_2 l_3 l_4 l_5 l_6 \) can be read in two ways: \( x_1 x_2 x_3' \) (with \( l_1 \) as the first letter) and \( x_1' x_2' x_3 \) (with \( l_2 \) as the first letter). There is a double reading of the sequence \( l_1, l_2, l_3, l_4, l_5, l_6 \) (corresponding to a double reading of the sequence \( l_2 l_3 l_4 l_5 l_6 l_1 \) in this case, \( l_1, l_2, l_3, l_4, l_5, l_6 \) is a 5-necklace for \( Y \).

Example 27. If \( x_1 = l_1 l_2 = AC, x_2 = l_3 l_4 = GT, x_3 = l_5 l_6 = AG \), \( x_1' = l_2 l_3 = CG, x_2' = l_4 l_5 = TA, x_3' = l_6 l_1 = GA \) are dinucleotides of \( Y \), we have the following relations: \( ACGT AG = AC \cdot GT \cdot AG = x_1 x_2 x_3, CGT AGA = CG \cdot TA \cdot GA = x_1' x_2' x_3' = CGT AGA \) and \( AC = x_1 \). So, \( Y \) is not a dinucleotide circular code (also a consequence of the fact that \( Y \) contains two conjugate dinucleotides \( AG \) and \( GA \)).

Proposition 28. A dinucleotide circular code has at most 6 elements.

Proof. There are 16 dinucleotides. Four dinucleotides are periodic: \( AA, CC, GG, TT \). The remaining 12 dinucleotides are partitioned in six conjugation classes: \{AC, CA\}, \{AG, GA\}, \{AT, TA\}, \{CG, GC\}, \{CT, TC\}, and \{GT, TG\}. By Proposition 6, a dinucleotide circular code has at most one dinucleotide in each of these conjugation classes. So, a dinucleotide circular code has at most 6 elements.

Proposition 29. Let \( (i, j, h, k) \) be a permutation of \( (A, C, G, T) \). If

\[
X = \{ij, ih, ik, jh, jk, hk\},
\]

then \( X \) is a dinucleotide circular code.

Proof. Suppose, by way of contradiction, that \( X \) is not a dinucleotide circular code, and let \( l_1, l_2, l_3, l_4, l_5 \) be a 5-necklace of \( X \). Note that, with the exception of \( l_1 \), the other letters \( l_2, l_3, l_4, l_5 \) composing the necklace must be a suffix of a dinucleotide of \( X \).

Claim 1. For \( \alpha \in \{2, 3, 4, 5\}, l_\alpha \neq i \).

Proof of Claim 1. By inspection, \( i \) is never a suffix of a dinucleotide of \( X \).

Claim 2. For \( \alpha \in \{3, 4, 5\}, l_\alpha \neq j \).

Proof of Claim 2. By inspection, \( j \) is a suffix only of \( ij \). For \( \alpha \in \{3, 4, 5\}, \) if \( l_\alpha = j \), then \( l_{\alpha - 1} = i \) which is impossible by Claim 1.

Claim 3. For each \( \alpha \in \{4, 5\}, l_\alpha \neq h \).

Proof of Claim 3. By inspection, \( h \) is a suffix only of \( ih \) and \( jh \).

Suppose, by way of contradiction, that \( l_5 = h \). Then, \( l_4 = i \) or \( l_4 = j \). If \( l_4 = i \), we are in contradiction with Claim 1 and if \( l_4 = j \), we are in contradiction with Claim 2. Suppose, by way of contradiction, that \( l_5 = h \). Then, \( l_5 = i \) or \( l_5 = j \). If \( l_5 = j \), we are in contradiction with Claim 1 and if \( l_5 = i \), we are in contradiction with Claim 2.

Claim 4. \( l_5 \neq k \).
Proof of Claim 4. By inspection, \( k \) is a suffix only of \( i, j, \) and \( h \). Suppose, by way of contradiction, that \( \ell_5 = k \). Then, \( \ell_4 = i, \ell_4 = j, \) or \( \ell_4 = h \). In the first case, we are in contradiction with Claim 1; in the second case, we are in contradiction with Claim 2; and in the third case, we are in contradiction with Claim 3.

By Claims 1, 2, 3, and 4, we have \( \ell_4 \neq i, \ell_4 \neq j, \ell_4 \neq h, \ell_5 \neq k \), and so, \( X \) has no 5-necklace. Consequently, \( X \) is a dinucleotide circular code. \( \square \)

Definition 30. A maximum dinucleotide circular code is a dinucleotide circular code having 6 elements.

Remark 31. In Proposition 29, we have considered an arbitrary permutation \((i, j, h, k)\) of \((A, C, G, T)\), and we have proved that a maximum dinucleotide circular code corresponds to it. As the number of possible permutations is 24, the number of maximum dinucleotide circular codes is at least 24, and we will prove hereafter that it is exactly 24.

In the maximum dinucleotide circular code \( X = \{ij, ih, ik, jh, jk, hk\} \) (Proposition 29), the letter \( i \) has three occurrences in prefix of dinucleotides of \( X \) (shortly in prefix of \( X \)), the letter \( j \) has two occurrences in prefix of \( X \), and \( h \) has one occurrence in prefix of \( X \). The letter \( k \) never occurs in prefix of \( X \). This is a general fact, in the sense that in any maximum dinucleotide circular code \( X \) there is a letter, say \( i \), with three occurrences in prefix of \( X \), and a letter, say \( j \), with two occurrences in prefix of \( X \), and a letter, say \( h \), with one occurrence in prefix of \( X \), while the remaining letter, say \( k \), never occurs in prefix of \( X \).

We will prove formally this general fact. In the sequel, a set of 4 nonnegative numbers having a sum equal to 6 is called a 4-partition of 6. By “set” we rather mean a “multiset” as some numbers can be equal, for example, \([2, 2, 2, 0]\). Define \( \ell_k \) (resp., \( \ell_C, \ell_G, \ell_T \)) as the number of occurrences of \( A \) (resp., \( C, G, T \)) in prefix of a maximum dinucleotide circular code \( X \).

Lemma 32. If \( X \) is a maximum dinucleotide circular code, then \([\ell_A, \ell_C, \ell_G, \ell_T] \) is a 4-partition of 6.

Proof. By Proposition 28. \( \square \)

Lemma 33. In any dinucleotide circular code, one has \( \ell_A \leq 3, \ell_C \leq 3, \ell_G \leq 3, \ell_T \leq 3 \).

Proof. The alphabet \( A \) contains four letters, and a dinucleotide circular code cannot contain periodic dinucleotides. \( \square \)

Example 34. For \( X = \{ij, ih, ik, jh, jk, hk\} \), the 4-partition \([\ell_i, \ell_j, \ell_h, \ell_k] = [\ell_A, \ell_C, \ell_G, \ell_T]\) of 6 is \([3, 2, 1, 0]\).

The following lemma will prove that the unique possible 4-partition \([\ell_A, \ell_C, \ell_G, \ell_T]\) for a maximum dinucleotide circular code is \([3, 2, 1, 0]\).

Lemma 35. For each maximum dinucleotide circular code \( X \), there exists a permutation \( \pi \) of \( A, C, G, T \) such that \( \pi(A) \) has three occurrences in prefix of \( X \), \( \pi(C) \) has two occurrences in prefix of \( X \), \( \pi(G) \) has one occurrence in prefix of \( X \), and \( \pi(T) \) has no occurrence in prefix of \( X \).

Proof. Putting the values \( \ell_A, \ell_C, \ell_G, \ell_T \) in nonincreasing order, by Lemma 33, we have to consider only the following cases: \([3, 3, 0, 0], [3, 2, 1, 0], [3, 1, 1, 1], [2, 2, 2, 0], [2, 2, 1, 1]\).

Case \([3, 3, 0, 0]\). Let \( i \) be the letter with three occurrences in prefix of \( X \). Let \( j, h, k \) be the three other letters of \( A, C, G, T \). We have \( ij, ih, ik \in X \). Without loss of generality, suppose that \( j \) has three occurrences in prefix of \( X \). Necessarily one of the two dinucleotides \( ji \) and \( jj \) must be in \( X \). But, in the first case, we are in contradiction with Proposition 6, and in the second case, we are in contradiction with Proposition 5.

So, the case \([3, 3, 0, 0]\) is impossible.

Case \([3, 1, 1, 1]\). Let \( i \) be the letter with three occurrences in prefix of \( X \). Let \( j, h, k \) be the three other letters of \( A, C, G, T \). Then, we have \( ij, ih, ik \in X \). \( ji \notin X \). Without loss of generality, we are in contradiction with Propositions 5 and 6.

Now, suppose that in \( X \), \( ji \notin X \). \( jh, kj \in X \). Without loss of generality, we have \( jh, jk, kj \in X \). But, \( jhjkj \) is a 5-necklace for \( X \). By Proposition 25, we are in contradiction.

So, the case \([3, 1, 1, 1]\) is impossible.

Case \([2, 2, 2, 0]\). Let \( i \) be one of the three letters with two occurrences in prefix of \( X \). Let \( j, h, k \) the three other letters of \( A, C, G, T \). Without loss of generality, we have \( ij, ih \in X \). With the two other letters having two occurrences in prefix of \( X \), we have three possibilities \([ij, h], [jk, h], [hk, j]\).

Case \([j, h]\). By Propositions 6 and 5, \( j, jh, hh \notin X \). But \( ij, ih, jh, jk, hj, hk \in X \). As \( jh, hj \) are conjugate, we are in contradiction with Proposition 6.

Case \([j, k]\). By Propositions 6 and 5, \( j, jh \notin X \). But \( ij, ih \notin X \). As \( kj \notin X \) (otherwise, we are in contradiction with Proposition 6), the two dinucleotides with \( k \) in prefix of \( X \) must be \( ki \) and \( kh \) and, consequently, \( ij, ih, jh, jk, ki, kh \in X \). But, \( kiji \) is a 5-necklace for \( X \). By Proposition 25, we are in contradiction.

So, the case \([2, 2, 2, 0]\) is impossible.

Case \([2, 2, 1, 1]\). Let \( i \) be one of the two letters with two occurrences in prefix of \( X \). Let \( j, h, k \) be the three other letters
of \{A, C, G, T\}. Without loss of generality, we have \(ij, ih \in X\). Consider the following cases:

(i) \(j\) has two occurrences in prefix of \(X\), and \(h\) and \(k\) have one occurrence in prefix of \(X\). By Propositions 6 and 5, \(ij, ih, hj, jk \in X\), and \(hk\) is the unique possible dinucleotide of \(X\) with \(h\) in prefix of \(X\). By Propositions 6 and 5, \(kj, kh, kk \notin X\). If \(kij\) \(\in X\) then \(ikjik\) is a 5-necklace of \(X\), and by Proposition 25, we are in contradiction. Thus, \(k\) cannot be a prefix of \(X\). 

(ii) \(h\) has two occurrences in prefix of \(X\), and \(j\) and \(k\) have one occurrence in prefix of \(X\). By Propositions 6 and 5, \(ij, ih, hj, jk \in X\), and \(hk\) is the unique possible dinucleotide of \(X\) with \(j\) in prefix of \(X\). By Propositions 6 and 5, \(kj, kh, kk \notin X\) and \(ki \notin X\) (otherwise, \(ijki\) is a 5-necklace for \(X\), and by Proposition 25, we are in contradiction). So, \(k\) cannot be a prefix of \(X\). 

(iii) \(k\) has two occurrences in prefix of \(X\), and \(j\) and \(h\) have one occurrence in prefix of \(X\). By Propositions 6 and 5, we have three possible cases \(ki, kj \in X\), \(ki, kh \in X\), and \(kj, kh \in X\).

(iiiia) \(ki, kj \in X\). So, \(ij, ih, ki, kj \in X\). By Propositions 6 and 5, \(ij, ih, kj \notin X\), but \(j \in X\). By Propositions 6 and 5, \(hj, hk \in X\). In the first case \(X = \{ij, ih, ki, kj, hj, h\}\), and by Proposition 6, we are in contradiction. In the second case, \(X = \{ij, ih, ki, kj, hj, hk\}\). But, \(ihkjh\) is a 5-necklace for \(X\). By Proposition 25, we are in contradiction.

(iiiib) \(ki, kh \in X\). So, \(ij, ih, ki, kh \in X\). By Propositions 6 and 5, \(j, ih, jk \notin X\), but \(h \in X\). By Propositions 6 and 5, \(j, kh \in X\). As \(j, hj \in X\) are conjugate, we are in contradiction with Proposition 6. So, the case \(2, 2, 1, 1\) is also impossible.

(iiiic) \(kj, kh \in X\). So, \(ij, ih, jk, kh \in X\). By Propositions 6 and 5, \(ij, jk \notin X\), but \(h \in X\). By Propositions 6 and 5, \(j, ih, h, jk \notin X\), but \(h \in X\). As \(j, hj \in X\) are conjugate, we are in contradiction with Proposition 6. So, the case \(2, 2, 1, 1\) is also impossible.

Only the 4-partition \(\{3, 2, 1, 0\}\) is realized by \(\{ij, ih, ik, jh, jk, hk\}\). It corresponds to the permutation \((i, j, h, k)\) of \((A, C, G, T)\). In other words, the permutation, whose existence is proved, is \(\pi(A) = i, \pi(C) = j, \pi(G) = k, \pi(T) = \) k. 

**Proposition 36.** There are 24 maximum dinucleotide circular codes.

**Proof.** By Proposition 29, each permutation \((i, j, h, k)\) of \((A, C, G, T)\) is associated with a maximum dinucleotide circular code \([ij, ih, ik, jh, jk, hk]\). As there are 24 permutations, the number of maximum dinucleotide circular codes is at least 24.

Now, let \(X\) be a maximum dinucleotide circular code. By Lemma 35, its 4-partition must be \(\{3, 2, 1, 0\}\). Let \(i\) (resp., \(j, h, k\)) be the letter of \(\sigma(A)\) having 3 (resp., 2, 1, 0) occurrences in prefix of \(X\). As \(i\) has three occurrences in prefix of \(X\), we must have \(ij, ih, ik \in X\). As \(j\) has two occurrences in prefix of \(X\), and as \(ih\) and \(jh\) are already in \(X\), we must also have \(j, hj \in X\). Finally, as \(h\) has only one occurrence in prefix of \(X\), and as \(ih\) and \(jh\) are already in \(X\), we must have \(hk \in X\). Consequently, \(X = \{ij, ih, ik, jh, jk, hk\}\), and \(X\) is one of the 24 maximum circular codes already considered. Thus, the number of maximum dinucleotide circular codes is exactly 24.

A computer calculus confirms that there are exactly 24 maximum dinucleotide circular codes (Table 1).

There are eight self-complementary maximum dinucleotide circular codes: \(X_1, X_4, X_{10}, X_{11}, X_{14}, X_{15}, X_{21}\), and \(X_{24}\) (Table 1). The 16 remaining ones are partitioned in two classes of eight maximum dinucleotide circular codes in bijective correspondence by the complementarity map (Table 1).

**Proposition 37.** If \(X\) is a maximum dinucleotide circular code, then \(\mathcal{C}(X)\) is also a maximum dinucleotide circular code.

**Proof.** By inspection (Table 1).

**Proposition 38.** If \(X\) is a maximum dinucleotide circular code, then 

\[
\mathcal{P}(\mathcal{C}(X)) = \mathcal{C}(\mathcal{P}(X)) \tag{2}
\]

**Proof.** By inspection (Table 1).

This proposition is not true with maximum trinucleotide circular codes, see for example, [6].

**4. Conclusion**

This new combinatorial study of circular codes in genes has proved that there are exactly 24 maximum dinucleotide circular codes on the 4-letter genetic alphabet \((A, C, G, T)\). They are listed in Table 1. Propositions 22, 37, and 38 lead to interesting properties with dinucleotide circular codes in DNA. Indeed, they ensure that several maximum dinucleotide circular codes can exist in the two strands of the DNA double helix simultaneously. Indeed, a maximum dinucleotide circular code \(X\) in a given strand \(s\) of DNA implies that its complementary set \(\mathcal{C}(s)\) in the complementary strand \(\mathcal{C}(s)\) of DNA is also a maximum dinucleotide circular code (Proposition 37) and according to two possibilities: \(\mathcal{C}(s) = X\) or \(\mathcal{C}(s) = Y\) with \(Y \neq X\) (Table 1). Furthermore, its permuted set \(\mathcal{P}(X)\) in \(s\), obtained by a frameshift of one letter of \(X\) in \(s\), is also a maximum dinucleotide circular code (Proposition 22). Finally, its complementary permuted set \(\mathcal{C}(\mathcal{P}(X))\) in \(\mathcal{C}(s)\) is also a maximum dinucleotide circular code (Proposition 38).

Chemical modification of nucleotides is ubiquitous in RNA and DNA. So far, a total of 107 modified nucleotides, for which chemical structures have been assigned, have been reported in RNA (see the RNA Modification Database at http://rna-mdb.cas.albany.edu/RNAmods/ [35]). The largest...
Table 1: The 24 maximum dinucleotide circular codes and their properties.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Dinucleotide circular code</th>
<th>ℵ</th>
<th>ℵ</th>
<th>ℵ</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>{AC, AG, AT, CG, CT, GT}</td>
<td>ℵ(X₁) = X₁</td>
<td>ℵ(X₁) = X₁</td>
<td>ℵ(ℵ(X₁)) = X₁</td>
</tr>
<tr>
<td>X₂</td>
<td>{AC, AG, AT, CG, TG}</td>
<td>ℵ(X₂) = X₁₃</td>
<td>ℵ(X₂) = X₁₃</td>
<td>ℵ(ℵ(X₂)) = X₁₃</td>
</tr>
<tr>
<td>X₃</td>
<td>{AC, AG, AT, CG, TC, TG}</td>
<td>ℵ(X₃) = X₁₇</td>
<td>ℵ(X₃) = X₁₇</td>
<td>ℵ(ℵ(X₃)) = X₁₇</td>
</tr>
<tr>
<td>X₄</td>
<td>{AC, AG, AT, CT, GC, GT}</td>
<td>ℵ(X₄) = X₄</td>
<td>ℵ(X₄) = X₄</td>
<td>ℵ(ℵ(X₄)) = X₄</td>
</tr>
<tr>
<td>X₅</td>
<td>{AC, AG, AT, GC, GT, TC}</td>
<td>ℵ(X₅) = X₉</td>
<td>ℵ(X₅) = X₉</td>
<td>ℵ(ℵ(X₅)) = X₉</td>
</tr>
<tr>
<td>X₆</td>
<td>{AC, AG, AT, GC, TG}</td>
<td>ℵ(X₆) = X₁₈</td>
<td>ℵ(X₆) = X₁₈</td>
<td>ℵ(ℵ(X₆)) = X₁₈</td>
</tr>
<tr>
<td>X₇</td>
<td>{AC, AG, CG, TA, TC, TG}</td>
<td>ℵ(X₇) = X₁₉</td>
<td>ℵ(X₇) = X₁₉</td>
<td>ℵ(ℵ(X₇)) = X₁₉</td>
</tr>
<tr>
<td>X₈</td>
<td>{AC, AG, GC, TA, TC, TG}</td>
<td>ℵ(X₈) = X₂₂</td>
<td>ℵ(X₈) = X₂₂</td>
<td>ℵ(ℵ(X₈)) = X₂₂</td>
</tr>
<tr>
<td>X₉</td>
<td>{AC, AT, CT, GA, GC, GT}</td>
<td>ℵ(X₉) = X₅</td>
<td>ℵ(X₉) = X₅</td>
<td>ℵ(ℵ(X₉)) = X₅</td>
</tr>
<tr>
<td>X₁₀</td>
<td>{AC, AT, GA, GC, GT, TC}</td>
<td>ℵ(X₁₀) = X₃₀</td>
<td>ℵ(X₁₀) = X₃₀</td>
<td>ℵ(ℵ(X₁₀)) = X₃₀</td>
</tr>
<tr>
<td>X₁₁</td>
<td>{AC, GA, GC, GT, TA, TC}</td>
<td>ℵ(X₁₁) = X₁₁</td>
<td>ℵ(X₁₁) = X₁₁</td>
<td>ℵ(ℵ(X₁₁)) = X₁₁</td>
</tr>
<tr>
<td>X₁₂</td>
<td>{AC, GA, GC, TA, TC, TG}</td>
<td>ℵ(X₁₂) = X₂₃</td>
<td>ℵ(X₁₂) = X₂₃</td>
<td>ℵ(ℵ(X₁₂)) = X₂₃</td>
</tr>
<tr>
<td>X₁₃</td>
<td>{AG, AT, CA, CG, CT, GT}</td>
<td>ℵ(X₁₃) = X₂</td>
<td>ℵ(X₁₃) = X₂</td>
<td>ℵ(ℵ(X₁₃)) = X₂</td>
</tr>
<tr>
<td>X₁₄</td>
<td>{AG, AT, CA, CG, TG}</td>
<td>ℵ(X₁₄) = X₁₄</td>
<td>ℵ(X₁₄) = X₁₄</td>
<td>ℵ(ℵ(X₁₄)) = X₁₄</td>
</tr>
<tr>
<td>X₁₅</td>
<td>{AG, CA, CG, TA, TC, TG}</td>
<td>ℵ(X₁₅) = X₁₅</td>
<td>ℵ(X₁₅) = X₁₅</td>
<td>ℵ(ℵ(X₁₅)) = X₁₅</td>
</tr>
<tr>
<td>X₁₆</td>
<td>{AG, CA, CG, TG}</td>
<td>ℵ(X₁₆) = X₃₀</td>
<td>ℵ(X₁₆) = X₃₀</td>
<td>ℵ(ℵ(X₁₆)) = X₃₀</td>
</tr>
<tr>
<td>X₁₇</td>
<td>{AT, CA, CG, TA, GC, GT}</td>
<td>ℵ(X₁₇) = X₃</td>
<td>ℵ(X₁₇) = X₃</td>
<td>ℵ(ℵ(X₁₇)) = X₃</td>
</tr>
<tr>
<td>X₁₈</td>
<td>{AT, CA, CT, GA, GC, GT}</td>
<td>ℵ(X₁₈) = X₆</td>
<td>ℵ(X₁₈) = X₆</td>
<td>ℵ(ℵ(X₁₈)) = X₆</td>
</tr>
<tr>
<td>X₁₉</td>
<td>{CA, CG, CT, GA, GT, TA}</td>
<td>ℵ(X₁₉) = X₇</td>
<td>ℵ(X₁₉) = X₇</td>
<td>ℵ(ℵ(X₁₉)) = X₇</td>
</tr>
<tr>
<td>X₂₀</td>
<td>{CA, CG, CT, GA, TA, TG}</td>
<td>ℵ(X₂₀) = X₁₆</td>
<td>ℵ(X₂₀) = X₁₆</td>
<td>ℵ(ℵ(X₂₀)) = X₁₆</td>
</tr>
<tr>
<td>X₂₁</td>
<td>{CA, CG, GA, TA, TC, TG}</td>
<td>ℵ(X₂₁) = X₂₁</td>
<td>ℵ(X₂₁) = X₂₁</td>
<td>ℵ(ℵ(X₂₁)) = X₂₁</td>
</tr>
<tr>
<td>X₂₂</td>
<td>{CA, CT, GA, GC, GT, TA}</td>
<td>ℵ(X₂₂) = X₈</td>
<td>ℵ(X₂₂) = X₈</td>
<td>ℵ(ℵ(X₂₂)) = X₈</td>
</tr>
<tr>
<td>X₂₃</td>
<td>{CA, GA, GC, GT, TA, TC}</td>
<td>ℵ(X₂₃) = X₁₂</td>
<td>ℵ(ℵ(X₂₃)) = X₁₂</td>
<td>ℵ(ℵ(X₂₃)) = X₁₂</td>
</tr>
<tr>
<td>X₂₄</td>
<td>{CA, GA, GC, TA, TC, TG}</td>
<td>ℵ(X₂₄) = X₂₄</td>
<td>ℵ(X₂₄) = X₂₄</td>
<td>ℵ(ℵ(X₂₄)) = X₂₄</td>
</tr>
</tbody>
</table>

Table 2: Amino acids coded by the trinucleotides dl in the the standard genetic code where dl ∈ ℵ₄ and l being any letter of ℵ₄.

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Trinucleotides dl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala (A)</td>
<td>GCI</td>
</tr>
<tr>
<td>Arg (R)</td>
<td>CGI</td>
</tr>
<tr>
<td>Gly (G)</td>
<td>GGI</td>
</tr>
<tr>
<td>Leu (L)</td>
<td>CII</td>
</tr>
<tr>
<td>Pro (P)</td>
<td>CCI</td>
</tr>
<tr>
<td>Thr (T)</td>
<td>ACl</td>
</tr>
<tr>
<td>Ser (S)</td>
<td>TCI</td>
</tr>
<tr>
<td>Val (V)</td>
<td>GTI</td>
</tr>
</tbody>
</table>

number, that is, 81, with the greatest structural diversity, is found in tRNA, with 30 in rRNA, 12 in mRNA, and 13 in other RNA species, most notably snRNA. The four nucleotides can be chemically modified, for example, methylenosine, dimethylenosine, trimethylenosine, methylterminine, dimethylterminine, thiomethyluridine, dimethyluridine, thioruridine, pseudouridine, dihydroxymethylcytosine, but also inosine, lysine, wybutosine, wyosine, queoisine, and arachosine. In DNA, the cytosine in the CG dinucleotide, involved in gene regulation, can have two chemical forms (methylenosine, hydroxymethylcytosine). This chemical change allows to store additional information, thus expanding the alphabet {A, C, G, T} by two letters. Thus, the generalization of dinucleotide circular code propositions over larger alphabets is very interesting and should be investigated.

Dinucleotide circular codes may be involved in retrieval of the modulo 2 frame in genomes, for example, in the dinucleotide repeats.

Dinucleotide circular codes may also have a biological function in the coding process of amino acids. In the standard genetic code, eight amino acids Ala (A), Arg (R), Gly (G), Leu (L), Pro (P), Thr (T), Ser (S), and Val (V) are coded by sets of trinucleotides involving dinucleotides. Indeed, for each of these eight amino acids, there exists a dinucleotide dl ∈ ℵ₄ such that all the trinucleotides of the form dl (where l is any letter of ℵ₄) code the same amino acid (Table 2).

Now, Gly (G) and Pro (P) cannot be coded by a dinucleotide circular code as their dinucleotides are periodic {CC, GG}. Moreover, Ala (A) and Arg (R) cannot be coded simultaneously by a dinucleotide circular code as their dinucleotides are conjugate {CG, GC} and similarly for Leu (L) and Ser (S) with the conjugate dinucleotides {CT, TC}. On the other hand, as any subset of a maximum dinucleotide circular code is also a dinucleotide circular code, the following properties exist.

(i) The four amino acids Arg (R), Leu (L), Thr (T), and Val (V) can be coded by the dinucleotide circular code...
\{AC, CG, CT, GT\} which is a proper subset of the maximum dinucleotide circular code \(X_1\) (Table 1).

(ii) The four amino acids Ala (A), Leu (L), Thr (T), and Val (V) can be coded by the dinucleotide circular code \{AC, CT, GC, GT\} which is a proper subset of the maximum dinucleotide circular code \(X_1\) (Table 1).

(iii) The four amino acids Ala (A), Ser (S), Thr (T), and Val (V) can be coded by the dinucleotide circular code \{AC, GC, GT, TC\} which is a proper subset of the maximum dinucleotide circular code \(X_5\) (Table 1).

These results contribute to the research field analysing the mathematical properties of genetic codes.

Acknowledgments

The authors thank the reviewers and Jacques Justin for their advice. The second author thanks the Dipartimento di Matematica "U. Dini" for giving him a friendly hospitality.

References


