Hindawi Publishing Corporation
ISRN Biomathematics
Volume 2013, Article ID 538631, 8 pages
http://dx.doi.org/10.1155/2013/538631

Research Article

Dinucleotide Circular Codes

Christian J. Michel[1] and Giuseppe Pirillo[2,3]

[1] *Equipe de Bioinformatique Théorique, BFO, LSIIT, 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

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 $\mathscr{A}_4 = \{A, C, G, T\}$. A dinucleotide is a word of two letters (diletter) on $\mathscr{A}_4$. A trinucleotide is a word of three letters (triletter) on $\mathscr{A}_4$. 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 $(l_1 l_2)^+$, $l_1, l_2 \in \mathscr{A}_4$ (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 $\mathscr{A}_4 = \{A, C, G, T\}$ 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 triletters) 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 $x_1, \dots, x_n, x_1', \dots, x_m' \in X$, $n, m \geq 1$, the condition $x_1 \cdots x_n = x_1' \cdots x_m'$ implies $n = m$ and $x_i = x_i'$ for $i = 1, \dots, n$.

Dinucleotide codes are read on a straight line.

Definition 2. A dinucleotide code X in $\mathcal{A}_4^2$ is circular if, for each $x_1, \dots, x_n, x_1', \dots, x_m' \in X$, $n, m \geq 1$, $p \in \mathcal{A}_4^*, s \in \mathcal{A}_4^+$, the conditions $sx_2 \cdots x_n p = x_1' \cdots x_m'$ and $x_1 = ps$ imply $n = m$, $p = \varepsilon$ (empty word), and $x_i = x_i'$ for $i = 1, \dots, n$.

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 $x = l_1 l_2$ and $y = l_2 l_1$.

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 $\mathscr{C}$, permutation $\mathscr{P}$, and mirror image $\tilde{}$ defined later are involutions.

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

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

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

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

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

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

Definition 14. The (left) dinucleotide circular permutation map $\mathscr{P} : \mathcal{A}_4^2 \to \mathcal{A}_4^2$ permutes circularly each dinucleotide $\mathscr{P}(l_1 l_2) = l_2 l_1$, $l_1, l_2 \in \mathcal{A}_4$.

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

Remark 16. $\mathscr{P}^2(X) = \mathscr{P}(\mathscr{P}(X)) = X$.

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

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

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

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

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

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

Proof. By Proposition 21 and Remark 20. □

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 $l_1, l_2, l_3, \dots, l_n, l_{n+1}$ be letters in $\mathcal{A}_4$. One says that the ordered sequence $l_1, l_2, l_3, \dots, l_n, l_{n+1}$ is a $(n+1)$-necklace for a subset $X \subset \mathcal{A}_4^2$ if each dinucleotide $l_1 l_2, l_2 l_3, \dots, l_n l_{n+1}$ 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) $\to$ (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 $j - i = 3$, then either $l_1 = l_4$ or $l_2 = l_5$.

(iiia) 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_1, l_5. So, $l_1l_2, l_2l_3, l_3l_1, l_1l_5 \in X$. Consider the sequence $l_1, l_2, l_3, l_1, l_2, l_3$. Put $x_1 = l_1l_2$, $x_2 = l_3l_1$, $x_3 = l_2l_3$, $x_1' = l_2l_3$, $x_2' = l_1l_2$, $x_3' = l_3l_1$. Note that $x_1, x_2, x_3, x_1', x_2', x_3'$ belong to X. Now, the following relations hold: $l_2l_3l_1l_2l_3l_1 = l_2 \cdot l_3l_1 \cdot l_2l_3 \cdot l_1 = l_2x_2x_3l_1 = x_1'x_2'x_3'$ and $l_1l_2 = x_1$. (Contradiction with the assumption that X is a circular code.)

(iiib) 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_1. So, $l_1l_2, l_2l_3, l_3l_4, l_4l_1 \in X$. Consider the sequence l_1, l_2, l_3, l_4, l_1. Put $x_1 = l_1l_2$, $x_2 = l_3l_4$, $x_1' = l_2l_3$, $x_2' = l_4l_1$. Note that x_1, x_2, x_1', x_2' belong to X. Now, the following relations hold: $l_2l_3l_4l_1 = l_2 \cdot l_3l_4 \cdot l_1 = l_2x_2l_1 = x_1'x_2'$ and $l_1l_2 = x_1$. 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 $x_1, \ldots, x_n, x_1', \ldots, x_n' \in X$, $n \geq 1$, $p_1, s_1 \in \mathcal{A}_4^+$, such that $s_1x_2 \cdots x_np_1 = x_1' \cdots x_n'$ and $x_1 = p_1s_1$. 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 $p_2s_2 = x_2$,..., $p_{n-1}s_{n-1} = x_{n-1}$, $p_ns_n = x_n$ and $s_1p_2 = x_1'$, $s_2p_3 = x_2'$,..., $s_{n-1}p_n = x_{n-1}'$, $s_np_1 = x_n'$. 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, 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 $x_1x_2x_3, \ldots, x_n$ be any sequence with $x_i = AC$ or $x_i = GT$. As X does not contain CA, CG, TA and TG, the sequence $x_1x_2x_3, \ldots, x_n$ cannot have a double decomposition on a circle. The set X has no 5-necklace as, if l_5 is C (resp., T) then l_4 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 $\mathcal{A}_4$ on a circle. If $x_1 = l_1l_2$, $x_2 = l_3l_4$, $x_3 = l_5l_6$, $x_1' = l_2l_3$, $x_2' = l_4l_5$, $x_3' = l_6l_1$ belong to a set Y, then Y cannot be a circular code because the sequence $l_1l_2l_3l_4l_5l_6$ can be read in two ways: $x_1x_2x_3$ (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 is a 5-necklace for Y.

Example 27. If $x_1 = l_1l_2 = AC$, $x_2 = l_3l_4 = GT$, $x_3 = l_5l_6 = AG$, $x_1' = l_2l_3 = CG$, $x_2' = l_4l_5 = TA$, $x_3' = l_5l_6 = GA$ are dinucleotides of Y, we have the following relations: $ACGTAG = AC \cdot GT \cdot AG = x_1x_2x_3$, $CGTAGA = CG \cdot TA \cdot GA = x_1'x_2'x_3' = Cx_2x_3A$ 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, and 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*

$$X = \{ij, ih, ik, jh, jk, hk\}, \quad (1)$$

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 $l_4 = h$. Then, $l_3 = i$ or $l_3 = j$. If $l_3 = i$, we are in contradiction with Claim 1 and if $l_3 = j$, 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 ik, jk, and hk. Suppose, by way of contradiction, that $l_5 = k$. Then, $l_4 = i$, $l_4 = j$, or $l_4 = h$. 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 $l_5 \neq i, l_5 \neq j, l_5 \neq h, l_5 \neq k$, and so, X has no 5-necklace. Consequently, X is a dinucleotide circular code. □

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 each maximum dinucleotide circular code X there is a letter, say i, with three occurrence 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 l_A (resp., l_C, l_G, l_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 $\{l_A, l_C, l_G, l_T\}$ is a 4-partition of 6.*

Proof. By Proposition 28. □

Lemma 33. *In any dinucleotide circular code, one has $l_A \leq 3$, $l_C \leq 3$, $l_G \leq 3$, $l_T \leq 3$.*

Proof. The alphabet $\mathscr{A}_4$ contains four letters, and a dinucleotide circular code cannot contain periodic dinucleotides. □

Example 34. For $X = \{ij, ih, ik, jh, jk, hk\}$, the 4-partition $\{l_i, l_j, l_h, l_k\} = \{l_A, l_C, l_G, l_T\}$ of 6 is $\{3, 2, 1, 0\}$.

The following lemma will prove that the unique possible 4-partition $\{l_A, l_C, l_G, l_T\}$ for a maximum dinucleotide circular code is $\{3, 2, 1, 0\}$.

Lemma 35. *For each maximum dinucleotide circular code X, there exists a permutation π 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 l_A, l_C, l_G, l_T in nonincreasing order, by Lemma 33, we have to consider only the following cases: $\{3, 3, 0, 0\}$, $\{3, 2, 1, 0\}$, $\{3, 1, 1, 1\}$, $\{2, 2, 2, 0\}$, and $\{2, 2, 1, 1\}$.

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 $ij, ih, ik \in X$ and $ii, ji, hi, ki \notin X$; otherwise, we are in contradiction with Propositions 5 and 6.

Now, suppose that in $X - \{ij, ih, ik\}$, the same letter, say j without loss of generality, has two occurrences in suffix of X, that is, $hj, kj \in X$. The letter j cannot be a prefix of X. Indeed, $ji \notin X$, $jh, jk \in X$ are in contradiction with Proposition 6, and $jj \in X$ is in contradiction with Proposition 5.

So, in $X - \{ij, ih, ik\}$, the letters j, h, k must have only one occurrence in suffix of X. Without loss of generality, we have $jh, hk, kj \in X$. But, $jhkjh$ 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 $\{j, h\}$, $\{j, k\}$, $\{h, k\}$.

Case $\{j, h\}$. By Propositions 6 and 5, $ji, jj, hi, hh \notin X$, 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, $ji, jj \notin X$, but $jh, jk \in X$. 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, $kijki$ is a 5-necklace for X. By Proposition 25, we are in contradiction.

Case $\{h, k\}$. By Propositions 6 and 5, $hi, hh \notin X$, but $hj, hk \in X$. As $kh \notin X$ (otherwise, we are in contradiction with Proposition 6), the two dinucleotides with k in prefix of X must be ki and kj and, consequently, $ij, ih, hj, hk, ki, kj \in X$. But, $ihkih$ 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, $ij, ih, jh, jk \in X$, 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 $ki \in X$ then $ijkij$ is a 5-necklace of X, and by Proposition 25, we are in contradiction. So, k cannot be a prefix of X. Contradiction.

(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, $ij, ih, hj, hk \in X$, and jk 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, $ijkij$ is a 5-necklace for X, and by Proposition 25, we are in contradiction). So, k cannot be a prefix in X. Contradiction.

(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 $kj, kh \in X$.

(iiia) $ki, kj \in X$. So, $ij, ih, ki, kj \in X$. By Propositions 6 and 5, $ji, jj, jk \notin X$, but $jh \in X$. By Propositions 6 and 5, $hj, hk \in X$. In the first case $X = \{ij, ih, ki, kj, jh, hj\}$, and by Proposition 6, we are in contradiction. In the second case, $X = \{ij, ih, ki, kj, jh, hk\}$. But, $ihkih$ is a 5-necklace for X. By Proposition 25, we are in contradiction.

(iiib) $ki, kh \in X$. So, $ij, ih, ki, kh \in X$. By Propositions 6 and 5, $hi, hh, hk \notin X$, but $hj \in X$. By Propositions 6 and 5, jk is the unique possible dinucleotide of X with prefix j. So, $X = \{ij, ih, ki, kh, hj, jk\}$. But, $ihjki$ is a 5-necklace for X. By Proposition 25, we are in contradiction.

(iiic) $kj, kh \in X$. So, $ij, ih, kj, kh \in X$. By Propositions 6 and 5, $ji, jj, jk \notin X$, but $jh \in X$. By Propositions 6 and 5, $hi, hh, hk \notin X$, but $hj \in X$. As $hj, jh \in X$ 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 $\pi(A) = i$, $\pi(C) = j$, $\pi(G) = k$, and $\pi(T) = k$. □

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 $\mathcal{A}_4$ 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 ij is already in X, we must also have $jh, jk \in X$. 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 $\mathscr{C}(X)$ is also a maximum dinucleotide circular code.*

Proof. By inspection (Table 1). □

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

$$\mathscr{P}(\mathscr{C}(X)) = \mathscr{C}(\mathscr{P}(X)). \quad (2)$$

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 $\mathscr{C}(X)$ in the complementary strand $\mathscr{C}(s)$ of DNA is also a maximum dinucleotide circular code (Proposition 37) and according to two possibilities: $\mathscr{C}(X) = X$ or $\mathscr{C}(X) = Y$ with $Y \neq X$ (Table 1). Furthermore, its permuted set $\mathscr{P}(X)$ 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 $\mathscr{C}(\mathscr{P}(X))$ in $\mathscr{C}(s)$ 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.

Symbol	Dinucleotide circular code	$\mathscr{C}$	$\mathscr{P}$	$\mathscr{PC}$
X_1	$\{AC, AG, AT, CG, CT, GT\}$	$\mathscr{C}(X_1) = X_1$	$\mathscr{P}(X_1) = X_{24}$	$\mathscr{P}(\mathscr{C}(X_1)) = X_{24}$
X_2	$\{AC, AG, AT, CG, CT, TG\}$	$\mathscr{C}(X_2) = X_{13}$	$\mathscr{P}(X_2) = X_{23}$	$\mathscr{P}(\mathscr{C}(X_2)) = X_{12}$
X_3	$\{AC, AG, AT, CG, TC, TG\}$	$\mathscr{C}(X_3) = X_{17}$	$\mathscr{P}(X_3) = X_{22}$	$\mathscr{P}(\mathscr{C}(X_3)) = X_8$
X_4	$\{AC, AG, AT, CT, GC, GT\}$	$\mathscr{C}(X_4) = X_4$	$\mathscr{P}(X_4) = X_{21}$	$\mathscr{P}(\mathscr{C}(X_4)) = X_{21}$
X_5	$\{AC, AG, AT, GC, GT, TC\}$	$\mathscr{C}(X_5) = X_9$	$\mathscr{P}(X_5) = X_{20}$	$\mathscr{P}(\mathscr{C}(X_5)) = X_{16}$
X_6	$\{AC, AG, AT, GC, TC, TG\}$	$\mathscr{C}(X_6) = X_{18}$	$\mathscr{P}(X_6) = X_{19}$	$\mathscr{P}(\mathscr{C}(X_6)) = X_7$
X_7	$\{AC, AG, CG, TA, TC, TG\}$	$\mathscr{C}(X_7) = X_{19}$	$\mathscr{P}(X_7) = X_{18}$	$\mathscr{P}(\mathscr{C}(X_7)) = X_6$
X_8	$\{AC, AG, GC, TA, TC, TG\}$	$\mathscr{C}(X_8) = X_{22}$	$\mathscr{P}(X_8) = X_{17}$	$\mathscr{P}(\mathscr{C}(X_8)) = X_3$
X_9	$\{AC, AT, CT, GA, GC, GT\}$	$\mathscr{C}(X_9) = X_5$	$\mathscr{P}(X_9) = X_{16}$	$\mathscr{P}(\mathscr{C}(X_9)) = X_{20}$
X_{10}	$\{AC, AT, GA, GC, GT, TC\}$	$\mathscr{C}(X_{10}) = X_{10}$	$\mathscr{P}(X_{10}) = X_{15}$	$\mathscr{P}(\mathscr{C}(X_{10})) = X_{15}$
X_{11}	$\{AC, GA, GC, GT, TA, TC\}$	$\mathscr{C}(X_{11}) = X_{11}$	$\mathscr{P}(X_{11}) = X_{14}$	$\mathscr{P}(\mathscr{C}(X_{11})) = X_{14}$
X_{12}	$\{AC, GA, GC, TA, TC, TG\}$	$\mathscr{C}(X_{12}) = X_{23}$	$\mathscr{P}(X_{12}) = X_{13}$	$\mathscr{P}(\mathscr{C}(X_{12})) = X_2$
X_{13}	$\{AG, AT, CA, CG, CT, GT\}$	$\mathscr{C}(X_{13}) = X_2$	$\mathscr{P}(X_{13}) = X_{12}$	$\mathscr{P}(\mathscr{C}(X_{13})) = X_{23}$
X_{14}	$\{AG, AT, CA, CG, CT, TG\}$	$\mathscr{C}(X_{14}) = X_{14}$	$\mathscr{P}(X_{14}) = X_{11}$	$\mathscr{P}(\mathscr{C}(X_{14})) = X_{11}$
X_{15}	$\{AG, CA, CG, CT, TA, TG\}$	$\mathscr{C}(X_{15}) = X_{15}$	$\mathscr{P}(X_{15}) = X_{10}$	$\mathscr{P}(\mathscr{C}(X_{15})) = X_{10}$
X_{16}	$\{AG, CA, CG, TA, TC, TG\}$	$\mathscr{C}(X_{16}) = X_{20}$	$\mathscr{P}(X_{16}) = X_9$	$\mathscr{P}(\mathscr{C}(X_{16})) = X_5$
X_{17}	$\{AT, CA, CG, CT, GA, GT\}$	$\mathscr{C}(X_{17}) = X_3$	$\mathscr{P}(X_{17}) = X_8$	$\mathscr{P}(\mathscr{C}(X_{17})) = X_{22}$
X_{18}	$\{AT, CA, CT, GA, GC, GT\}$	$\mathscr{C}(X_{18}) = X_6$	$\mathscr{P}(X_{18}) = X_7$	$\mathscr{P}(\mathscr{C}(X_{18})) = X_{19}$
X_{19}	$\{CA, CG, CT, GA, GT, TA\}$	$\mathscr{C}(X_{19}) = X_7$	$\mathscr{P}(X_{19}) = X_6$	$\mathscr{P}(\mathscr{C}(X_{19})) = X_{18}$
X_{20}	$\{CA, CG, CT, GA, TA, TG\}$	$\mathscr{C}(X_{20}) = X_{16}$	$\mathscr{P}(X_{20}) = X_5$	$\mathscr{P}(\mathscr{C}(X_{20})) = X_9$
X_{21}	$\{CA, CG, GA, TA, TC, TG\}$	$\mathscr{C}(X_{21}) = X_{21}$	$\mathscr{P}(X_{21}) = X_4$	$\mathscr{P}(\mathscr{C}(X_{21})) = X_4$
X_{22}	$\{CA, CT, GA, GC, GT, TA\}$	$\mathscr{C}(X_{22}) = X_8$	$\mathscr{P}(X_{22}) = X_3$	$\mathscr{P}(\mathscr{C}(X_{22})) = X_{17}$
X_{23}	$\{CA, GA, GC, GT, TA, TC\}$	$\mathscr{C}(X_{23}) = X_{12}$	$\mathscr{P}(X_{23}) = X_2$	$\mathscr{P}(\mathscr{C}(X_{23})) = X_{13}$
X_{24}	$\{CA, GA, GC, TA, TC, TG\}$	$\mathscr{C}(X_{24}) = X_{24}$	$\mathscr{P}(X_{24}) = X_1$	$\mathscr{P}(\mathscr{C}(X_{24})) = X_1$

Table 2: Amino acids coded by the trinucleotides dl in the the standard genetic code where $d \in \mathscr{A}_4^2$ and l being any letter of $\mathscr{A}_4$.

Amino acid	Trinucleotides dl
Ala (*A*)	*GCl*
Arg (*R*)	*CGl*
Gly (*G*)	*GGl*
Leu (*L*)	*CTl*
Pro (*P*)	*CCl*
Thr (*T*)	*ACl*
Ser (*S*)	*TCl*
Val (*V*)	*GTl*

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, methyladenosine, dimethyladenosine, trimethyladenosine, methylcytidine, dimethylcytidine, thiocytidine, methylguanosine, dimethylguanosine, trimethylguanosine, methyluridine, dimethyluridine, thiouridine, pseudouridine, dihydrouridine, but also inosine, lysidine, wybutosine, wyosine, queuosine, and archaeosine. In DNA, the cytosine in the *CG* dinucleotide, involved in gene regulation, can have two chemical forms (methylcytosine, 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 $d \in \mathscr{A}_4^2$ such that all the trinucleotides of the form dl (where l is any letter of $\mathscr{A}_4$) 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_4 (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

[1] J. Berstel and D. Perrin, *Theory of Codes*, Academic Press, London, UK, 1985.

[2] J. L. Lassez, "Circular codes and synchronization," *International Journal of Computer and Information Sciences*, vol. 5, no. 2, pp. 201–208, 1976.

[3] F. H. C. Crick, J. S. Griffith, and L. E. Orgel, "Codes without commas," *Proceedings of the National Academy of Sciences*, vol. 43, pp. 416–421, 1957.

[4] S. W. Golomb, B. Gordon, and L. R. Welch, "Comma-free codes," *Canadian Journal of Mathematics*, vol. 10, pp. 202–209, 1958.

[5] S. W. Golomb, L. R. Welch, and M. Delbrück, "Construction and properties of comma-free codes," *Biologiske Meddelelser, Kongelige Danske Videnskabernes Selskab*, vol. 23, no. 9, 1958.

[6] D. G. Arquès and C. J. Michel, "A complementary circular code in the protein coding genes," *Journal of Theoretical Biology*, vol. 182, no. 1, pp. 45–58, 1996.

[7] A. J. Koch and J. Lehmann, "About a symmetry of the genetic code," *Journal of Theoretical Biology*, vol. 189, no. 2, pp. 171–174, 1997.

[8] M. P. Béal and J. Senellart, "On the bound of the synchronization delay of a local automaton," *Theoretical Computer Science*, vol. 205, no. 1-2, pp. 297–306, 1998.

[9] F. Bassino, "Generating functions of circular codes," *Advances in Applied Mathematics*, vol. 22, no. 1, pp. 1–24, 1999.

[10] R. Jolivet and F. Rothen, "Peculiar symmetry of DNA sequences and evidence suggesting its evolutionary origin in a primeval genetic code," in *Proceedings of the 1st European Workshop in Exo-/Astro-Biology*, P. Ehrenfreund, O. Angerer, and B. Battrick, Eds., ESA SP-496, pp. 173–176, Noordwijk, The Netherlands.

[11] G. Frey and C. J. Michel, "Circular codes in archaeal genomes," *Journal of Theoretical Biology*, vol. 223, no. 4, pp. 413–431, 2003.

[12] C. Nikolaou and Y. Almirantis, "Mutually symmetric and complementary triplets: differences in their use distinguish systematically between coding and non-coding genomic sequences," *Journal of Theoretical Biology*, vol. 223, no. 4, pp. 477–487, 2003.

[13] G. Pirillo, "A characterization for a set of trinucleotides to be a circular code," in *Determinism, Holism, and Complexity*, C. Pellegrini, P. Cerrai, P. Freguglia, V. Benci, and G. Israel, Eds., Kluwer Academic Publisher, New York, NY, USA, 2003.

[14] G. Pirillo and M. A. Pirillo, "Growth function of self-complementary circular codes," *Biology Forum*, vol. 98, no. 1, pp. 97–110, 2005.

[15] G. Frey and C. J. Michel, "Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes," *Computational Biology and Chemistry*, vol. 30, no. 2, pp. 87–101, 2006.

[16] J. L. Lassez, R. A. Rossi, and A. E. Bernal, "Crick's hypothesis revisited: the existence of a universal coding frame," in *Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW '07)*, pp. 745–751, Niagara Falls, Canada, May 2007.

[17] C. J. Michel, G. Pirillo, and M. A. Pirillo, "Varieties of comma-free codes," *Computers and Mathematics with Applications*, vol. 55, no. 5, pp. 989–996, 2008.

[18] C. J. Michel, G. Pirillo, and M. A. Pirillo, "A relation between trinucleotide comma-free codes and trinucleotide circular codes," *Theoretical Computer Science*, vol. 401, no. 1–3, pp. 17–26, 2008.

[19] G. Pirillo, "A hierarchy for circular codes," *RAIRO-Theoretical Informatics and Applications*, vol. 42, no. 4, pp. 717–728, 2008.

[20] G. Pirillo, "Some remarks on prefix and suffix codes," *Pure Mathematics and Applications*, vol. 19, pp. 53–60, 2008.

[21] C. J. Michel and G. Pirillo, "Identification of all trinucleotide circular codes," *Computational Biology and Chemistry*, vol. 34, no. 2, pp. 122–125, 2010.

[22] G. Pirillo, "Non sharing border codes," *The Advances in Applied Mathematics and Mechanics*, vol. 3, pp. 215–223, 2010.

[23] C. J. Michel and G. Pirillo, "Strong trinucleotide circular codes," *International Journal of Combinatorics*, vol. 2011, Article ID 659567, 14 pages, 2011.

[24] L. Bussoli, C. J. Michel, and G. Pirillo, "On some forbidden configurations for self-complementary trinucleotide circular codes," *Journal for Algebra Number Theory Academia*, vol. 2, pp. 223–232, 2011.

[25] D. L. Gonzalez, S. Giannerini, and R. Rosa, "Circular codes revisited: a statistical approach," *Journal of Theoretical Biology*, vol. 275, no. 1, pp. 21–28, 2011.

[26] L. Bussoli, C. J. Michel, and G. Pirillo, "On conjugation partitions of sets of trinucleotides," *Applied mathematics*, vol. 3, pp. 107–112, 2012.

[27] C. J. Michel, G. Pirillo, and M. A. Pirillo, "A classification of 20-trinucleotide circular codes," *Information and Computation*, vol. 212, pp. 55–63, 2012.

[28] M. Burset, I. A. Seledtsov, and V. V. Solovyev, "Analysis of canonical and non-canonical splice sites in mammalian genomes," *Nucleic Acids Research*, vol. 28, no. 21, pp. 4364–4375, 2000.

[29] S. M. Mount, "A catalogue of splice junction sequences," *Nucleic Acids Research*, vol. 10, no. 2, pp. 459–472, 1982.

[30] A. Bird, "The dinucleotide CG as a genomic signalling module," *Journal of Molecular Biology*, vol. 409, no. 1, pp. 47–53, 2011.

[31] F. Gebhardt, K. S. Zänker, and B. Brandt, "Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1," *Journal of Biological Chemistry*, vol. 274, no. 19, pp. 13176–13180, 1999.

[32] H. Buerger, J. Packeisen, A. Boecker et al., "Allelic length of a CA dinucleotide repeat in the egfr gene correlates with the frequency of amplifications of this sequence—first results of an inter-ethnic breast cancer study," *Journal of Pathology*, vol. 203, no. 1, pp. 545–550, 2004.

[33] A. L. Schmidt and V. Mitter, "Microsatellite mutation directed by an external stimulus," *Mutation Research*, vol. 568, no. 2, pp. 233–243, 2004.

[34] H. Cuppens, W. Lin, M. Jaspers et al., "Polyvariant mutant cystic fibrosis transmembrane conductance regulator genes: the polymorphic (TG)m locus explains the partial penetrance of the T5 polymorphism as a disease mutation," *Journal of Clinical Investigation*, vol. 101, no. 2, pp. 487–496, 1998.

[35] J. Rozenski, P. F. Crain, and J. A. McCloskey, "The RNA modification database: 1999 update," *Nucleic Acids Research*, vol. 27, no. 1, pp. 196–197, 1999.