

## Circular codes, symmetries and transformations

**Elena Fimmel · Simone Giannerini · Diego Luis Gonzalez · Lutz Strüngmann**

Received: 9 October 2013 / Revised: 19 February 2014  
© Springer-Verlag Berlin Heidelberg 2014

**Abstract** Circular codes, putative remnants of primeval comma-free codes, have gained considerable attention in the last years. In fact they represent a second kind of genetic code potentially involved in detecting and maintaining the normal reading frame in protein coding sequences. The discovering of an universal code across species suggested many theoretical and experimental questions. However, there is a key aspect that relates circular codes to symmetries and transformations that remains to a large extent unexplored. In this article we aim at addressing the issue by studying the symmetries and transformations that connect different circular codes. The main result is that the class of 216  $C^3$  maximal self-complementary codes can be partitioned into 27 equivalence classes defined by a particular set of transformations. We show that such transformations can be put in a group theoretic framework with an intuitive geometric interpretation. More general mathematical results about symmetry transformations which are valid for any kind of circular codes are also presented. Our results pave the way to the study of the biological consequences of the mathematical

---

E. Fimmel · L. Strüngmann  
Faculty of Computer Sciences, Institute of Applied Mathematics, Mannheim University  
of Applied Sciences, 68163 Mannheim, Germany  
e-mail: e.fimmel@hs-mannheim.de

L. Strüngmann  
e-mail: l.struengmann@hs-mannheim.de

S. Giannerini (✉) · D. L. Gonzalez  
Department of Statistical Sciences, University of Bologna, Bologna, Italy  
e-mail: simone.giannerini@unibo.it

D. L. Gonzalez  
CNR-IMM, Sezione di Bologna, Via Gobetti 101, Bologna, Italy  
e-mail: gonzalez@bo.imm.cnr.it

structure behind circular codes and contribute to shed light on the evolutionary steps that led to the observed symmetries of present codes.

**Keywords** Circular codes · Symmetry · Genetic code · Nucleotide transformations · Group theory

**Mathematics Subject Classification** 92D20 · 94B50 · 94B60 · 62P10

## 1 Introduction

Protein synthesis implies the decoding of the template information as sequences of nucleotides along nucleic acids into the amino acids that form a nascent protein. In this process accuracy is crucial: even a single error in the incorporation of the amino acid into the polypeptide chain may be the cause of diminished or even absent functionality in a biologically active protein. Accuracy of translation depends on accurate frame maintenance, on the correct pairing between codons and tRNA-anticodons inside the ribosome, and also on the correct charging of tRNAs with their cognate amino acids. This latter operation is ensured by a kind of enzymes called aaRSs (aminoacyl tRNA synthetases). Protein coding sequences are read by the synthesis machinery sequentially in groups of three nucleotides (codons). A codon is mapped to a specific amino acid through the genetic code, a translation table that connects the 64 possible codons with the 20 amino acids (plus punctuation marks). As mentioned, the maintenance of the correct reading frame is essential. A shift of one or two nucleotides of the position of the ribosome along the coding sequence leads to a frame-shift error that changes completely the identity of the coded amino acids. Excluding the case of programmed frame-shifts, where an alternative protein is coded in a frame different from the normal one, usually frame-shift errors are deleterious.

At the beginning of molecular genetics a possible simultaneous solution for the implementation of the genetic code and for frame synchronization was proposed. Before the discovery of the actual structure of the standard genetic code, Crick conjectured the existence of a code possessing the comma freeness property (Crick et al. 1957). Such kind of codes aroused interest from the point of view of coding theory because they are a particular type of error correcting codes (Golomb et al. 1958). In comma-free codes, a subset of the 64 possible codons are used for coding the 20 amino acids. The subset is chosen in such a way that a unique natural reading frame is allowed: the reading of a sequence out-of-frame produces invalid codons (codons that do not belong to the allowed subset). These codes are also called self synchronizing because they allow to discriminate the correct reading frame in any position along the sequence. They allow also to reject non-valid codons, that is, they allow the detection of errors in coding sequences. Unfortunately, the proposal of Crick turned out to be wrong (Hayes 1998). However, recent works have shown that a particular kind of related codes, i.e. circular codes, are indeed used in protein-coding sequences. Circular codes are a less restrictive version of comma-free codes and can be used for normal reading frame retrieval (Frey and Michel 2006; Michel et al. 2008). One such instance is the so-called  $X_0$  code empirically found both in eukaryotes and in

prokaryotes (Arquès and Michel 1996). In a recent study (Gonzalez et al. 2011), it has been shown that, on average, the code  $X_0$  has the best covering capability but there is a great variability as some codes are preferred over others, depending on the type of organism. This poses important biological questions about the existence of a unique universal code like  $X_0$  rather than thinking about the codes in terms of classes. The connections between protein-coding sequences and normal reading frame synchronization have been studied also by using a recently developed mathematical theory of the genetic code (Gonzalez 2004, 2008). In particular, it is possible to retrieve the reading frame of a protein coding sequence by using the information of dichotomic classes, quantities that are derived from the mentioned model (Giannerini et al. 2012). Dichotomic classes possess precise mathematical properties and a group structure that suggest that symmetries and transformations play a crucial role in the organization of genetic information. Moreover, in Michel and Pirillo (2011) and Michel et al. (2012) it is stressed the importance of the symmetric group for the study of circular codes. Thus, it is natural to ask whether symmetries and transformations also play an important role in characterizing circular codes and frame synchronization.

In the present work we study the symmetry properties of circular codes. In particular, we focus on the class of 216 maximal, self-complementary,  $C^3$  codes ( $X_0$  belongs to this class). In Sect. 2 we introduce the notation and define the set of transformations on the nucleotides and on the indices. Section 3 contains the results that connect circular codes and transformations. The main theorem proves that there is a subgroup of transformations of the nucleotides such that the set of 216 codes is invariant. These transformations are the bijections that commute with the complementary transformation and allow to classify the 216 codes into 27 equivalence classes. Moreover, we show that the 88 codes (Koch and Lehmann 1997; Lacan and Michel 2001) that can be obtained from the nucleotide frequencies in the positions of the codon are contained exactly in 11 of the 27 classes; in accordance with Lacan and Michel (2001) we call these 88 codes Nucleotide Frequency or “NF codes”. We present also an intuitive geometrical interpretation of the results and further theorems that establish the conditions under which circularity is preserved. The importance of the reverse symmetry follows naturally from our findings. Section 4 provides conclusions and perspectives.

## 2 Codons and transformations

The genetic code is written with words of three letters, codons, built over an alphabet

$$\mathcal{B} := \{U(T), C, A, G\}$$

of four letters, nucleotide bases Uracil (Thymine), Cytosine, Adenine, and Guanine, in short  $U(T)$ ,  $C$ ,  $A$ ,  $G$ . The *symmetric group* on the set  $\mathcal{B}$  is defined as

$$S_{\mathcal{B}} = \{\pi : \mathcal{B} \rightarrow \mathcal{B} \mid \pi \text{ is bijective}\}$$

with the usual group operation given by composition as functions. The group  $S_{\mathcal{B}}$  has 24 elements and is isomorphic to the symmetric group  $S_4$  on four elements (see

[Rotman 1995](#) for more details on group theory and symmetric groups). We will use standard notation as can be found in [Rotman \(1995\)](#), e.g. we will either write  $\pi_{AGCT}$  or  $\pi : (A, T, C, G) \rightarrow (G, A, T, C)$  if  $\pi$  satisfies  $\pi(A) = G, \pi(T) = A, \pi(C) = T$ , and  $\pi(G) = C$ . Bijective mappings  $\pi : \mathcal{B} \rightarrow \mathcal{B}$  can be applied componentwise to  $x \in \mathcal{B}^3$ , the set of codons, and thus induce a bijective map  $\mathcal{B}^3 \rightarrow \mathcal{B}^3$  which we will denote also by  $\pi$ .

Notice that, as shown in [Gonzalez et al. \(2008\)](#), there are 4 bijective transformations that are invariant with respect to the chemical characters of the nucleotides (we will use the notations from [Fimmel et al. 2013](#)). These are the Identity:

$$I \text{ (or } id) : (A, T, C, G) \rightarrow (A, T, C, G);$$

Strong/weak (SW) or complementary transformation:

$$SW \text{ (or } c) : (A, T, C, G) \rightarrow (T, A, G, C);$$

Pyrimidine/purine (YR) or parity transformation:

$$YR \text{ (or } p) : (A, T, C, G) \rightarrow (G, C, T, A);$$

and Keto/Amino (KM) or Rumer's transformation:

$$KM \text{ (or } r) : (A, T, C, G) \rightarrow (C, G, A, T).$$

In the following, we will use the convention that  $\{I, SW, YR, KM\}$  are used when we want to stress the biological context whereas  $\{id, c, p, r\}$  are used when we want to put the focus on the mathematical properties. Especially the complementary mapping SW (which we also denote with  $c$ ) will be used.

For a subset  $X \subseteq \mathcal{B}^3$  of the 64 codons in  $\mathcal{B}^3$  and a map  $\pi : \mathcal{B}^3 \rightarrow \mathcal{B}^3$  we define  $\pi(X)$  as

$$\pi(X) := \{\pi(x) | x \in X\}.$$

Now let us consider the *symmetric group*  $(S_3, \circ)$ , where

$$S_3 := \{\alpha : \{1, 2, 3\} \rightarrow \{1, 2, 3\} \mid \alpha \text{ is bijective}\}$$

and  $\circ$  denotes the composition of mappings. For instance:  $(132) \in S_3$  is the permutation such that  $1 \mapsto 3, 2 \mapsto 1, 3 \mapsto 2$ . Clearly, any such  $\alpha$  induces a mapping on the set of codons  $\mathcal{B}^3$  by permuting the order of the bases in the codons, e.g.  $(132)$  sends a codon  $(b_1, b_2, b_3)$  to  $(b_3, b_1, b_2)$ . Hence, for a given  $X \subseteq \mathcal{B}^3$ , we have that  $\pi(X)$  (for  $\pi : \mathcal{B}^3 \rightarrow \mathcal{B}^3$ ) is a transformation of the nucleotides of  $X$  whereas  $\alpha(X)$  (for  $\alpha \in S_3$ ) is a permutation of the positions of a codon. As we will show below, this is a difference that plays a crucial role from the biological point of view. In what follows we will focus on the subgroup of cyclical permutations of  $(S_3, \circ)$  denoted by

$$\mathcal{A}_3 := \{\alpha_0 = (1)(2)(3), \alpha_1 = (123), \alpha_2 = (132)\} \subset S_3$$

$(\mathcal{A}_3, \circ)$  is known as the *alternating subgroup* of  $(S_3, \circ)$  and its group table is given by

| $\circ$    | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ |
|------------|------------|------------|------------|
| $\alpha_0$ | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ |
| $\alpha_1$ | $\alpha_1$ | $\alpha_2$ | $\alpha_0$ |
| $\alpha_2$ | $\alpha_2$ | $\alpha_0$ | $\alpha_1$ |

In particular, we have

$$\alpha_1 \circ \alpha_1 = \alpha_2 \quad \text{and} \quad \alpha_2 \circ \alpha_2 = \alpha_1 \quad \text{and} \quad \alpha_1 \circ \alpha_2 = \alpha_2 \circ \alpha_1 = \alpha_0. \quad (1)$$

Moreover,  $\mathcal{A}_3$  forms a *normal* subgroup of  $S_3$  (in symbols  $\mathcal{A}_3 \triangleleft S_3$ ), i.e.  $\pi \mathcal{A}_3 \pi^{-1} = \mathcal{A}_3$  for all  $\pi \in S_3$  (see [Rotman 1995](#) for more details on normal subgroups). As mentioned above,  $S_3$  defines a group action on the set of codons and therefore we will call two codons  $x_1, x_2 \in \mathcal{B}^3$  *cyclically equivalent* if there exists a mapping  $\alpha \in \mathcal{A}_3$  such that

$$\alpha(x_1) = x_2.$$

In this case we write  $x_1 \sim x_2$ . For instance, given the codon *ATG* we have  $\alpha_1(\text{ATG}) = \text{TGA}$  and  $\alpha_2(\text{ATG}) = \text{GAT}$ , hence  $\text{ATG} \sim \text{TGA} \sim \text{GAT}$ . The relation  $\sim$  is an equivalence relation on the set of codons  $\mathcal{B}^3$  since  $\mathcal{A}_3$  forms a group. In this paper we are interested in the equivalence (conjugacy) classes induced by  $\sim$ . Clearly, the equivalence classes of the codons *AAA*, *CCC*, *GGG*, *TTT* contain only one element since the elements are permutation invariant. The remaining twenty equivalence classes have three elements each.

Another (biologically) important element of  $S_3$  is the reversing permutation of the indices (31)(2) which is not an element of  $\mathcal{A}_3$ . We will indicate this permutation by  $\leftarrow$  so that given a codon  $x = (b_1, b_2, b_3) \in \mathcal{B}^3$  we have  $\overleftarrow{x} := (b_3, b_2, b_1)$ . The reversing permutation is an important transformation that appears ubiquitously in genetic sequences, for example in inverse transpositions (see e.g. [Lewin 2004](#), pp. 469–470). Moreover, some proteins could be coded in the reverse sense, which is usual in mitochondria; also, it has been suggested that inversion is a primeval symmetry intimately related to the origin of protein coding ([Gonzalez et al. 2012](#)).

By the normality of  $\mathcal{A}_3$  in  $S_3$  it is easy to see that

$$\overleftarrow{\alpha_1(x)} = \alpha_2(\overleftarrow{x}) \quad \text{and} \quad \overleftarrow{\alpha_2(x)} = \alpha_1(\overleftarrow{x}) \quad (2)$$

for all  $x \in \mathcal{B}^3$ . In other words, the reversing permutation and the circular permutations do not commute but  $\alpha_1$  and  $\alpha_2$  are exchanged. On the other hand every transformation of the nucleotides  $\pi$  commutes with every permutation of the indices  $\alpha$ , i.e.

$$\pi \circ \alpha = \alpha \circ \pi, \quad (3)$$

where  $\pi \in S_{\mathcal{B}}$  and  $\alpha \in S_3$ . This property will be used later on.

### 3 Circular codes and transformations

Circular codes, as previously remarked, are a less restrictive version of comma-free codes and seem to play a key role in normal reading frame retrieval and maintenance; they are a sort of second genetic code (Arquès and Michel 1996; Michel 2008). Moreover, due to their properties, they could be the relics of some primeval comma-free codes, i.e., codes where punctuation signs are not needed for retrieving the correct reading frame. In the framework of protein synthesis we can define a circular code as a set of codons so that any arbitrary (circular) concatenation of codons of the set cannot be decomposed in a different frame by concatenating codons of the same code; in brief, there is only one valid frame for reading the sequence with only words of the code. In more mathematical terms a precise definition of a circular code and of some additional properties that they possess is as follows:

**Definition 1** Let  $X \subseteq \mathcal{B}^3$ . We will call a set of codons  $X$  a *trinucleotide circular code* if any word over the alphabet  $\mathcal{B}$  written on a circle has at most one decomposition into words from  $X$ . By word written on a circle it is intended that after the last letter the word starts again (from its first letter). We will call a trinucleotide circular code  $X$  *maximal* if it contains exactly 20 codons (i.e.  $|X| = 20$ ).

To illustrate the above definition let us consider the following example: Assume that  $ACG \in X$ . Then the word  $ACGACG$  can also be read on the circle as  $CGACGA$  or  $GACGAC$ . But these are exactly the words  $\alpha_1(ACG)\alpha_1(ACG)$  and  $\alpha_2(ACG)\alpha_2(ACG)$ . Thus we have the following remark.

*Remark 1* A trinucleotide circular code  $X \subseteq \mathcal{B}^3$  can contain at most one element from each complete equivalence class (with respect to  $\sim$ ) and cannot contain the codons  $AAA$ ,  $CCC$ ,  $GGG$ ,  $TTT$  since every codon from  $X$  is also a word over  $\mathcal{B}$ . Thus, a trinucleotide circular code can contain at most 20 codons and there are at most  $3^{20}$  potential different maximal trinucleotide circular codes.

Here are some examples of trinucleotide circular codes (verification is by easy calculations)

- $X = \{ATC, TCC, CAA\}$
- $X = \{GGT, GGC, ACT, ACC, AGC, AGT, GAC, GAT, GTC, GTT, AAT, ATT, AAC, ATC, GCT, GCC\}$

Among the trinucleotide circular codes there are some codes that turned out to be of special biological interest, namely those that are also self-complementary and  $C^3$ -codes.

**Definition 2** Let  $X \subseteq \mathcal{B}^3$ . We will say that  $X$  is a  $C^3$ -code if  $X$ , as well as  $X_1$  and  $X_2$  are circular, where  $X_1 := \alpha_1(X)$  and  $X_2 := \alpha_2(X)$ .

Note that by definition any  $C^3$ -code is also circular.

**Definition 3** Let  $X \subseteq \mathcal{B}^3$ . We will call  $X$  *self-complementary* if for each codon  $x \in X$  its anticodon  $\overleftarrow{c(x)}$  is also in  $X$ :

$$x \in X \Leftrightarrow \overleftarrow{c(x)} \in X.$$

We will also use the notation

$$X = \overleftarrow{c(X)}.$$

In Remark 2 we will see that there are maximal circular codes (even  $C^3$ -codes) that are not self-complementary and also self-complementary codes that are not  $C^3$ . As shown in Michel and Pirillo (2010) the class of self-complementary maximal  $C^3$ -codes, which we denote by  $\mathfrak{C}$ , contains 216 codes. The universal  $X_0$  code discovered in Arquès and Michel (1996) is one of these 216 codes.

In the following, we prove that self-complementary  $C^3$ -codes are intimately related both to bijective transformations and to the reversing permutation. This allows us to divide the 216 maximal self-complementary  $C^3$ -codes into 27 equivalence classes. In each of these equivalence classes we have 8 maximal self-complementary  $C^3$ -codes that are related by a set of transformations  $\pi \in S_B$ . Moreover, the equivalence classes have a geometrical interpretation implied by the symmetry group of the square. We have the following main results: the first one shows when circularity and the  $C^3$ -property are preserved under transformations and permutations.

**Theorem 1** *The following hold:*

- *The identical and the reversing permutations are the sole permutations of the positions of the bases of a codon which preserve the circularity of any circular code  $X \subseteq \mathcal{B}^3$ .*
- *Let  $X \subseteq \mathcal{B}^3$  be a trinucleotide circular code. For every  $\pi \in S_B$ ,  $\pi(X)$  is also a trinucleotide circular code. Furthermore, If  $X$  is a  $C^3$ -code, then  $\pi(X)$  is also a  $C^3$ -code.*

*Proof* See Appendix. □

The statement about the reversing permutation was already shown in Michel et al. (2012) (compare Proposition 5). We generalize this statement regarding all permutations of the positions of the bases and give an independent proof for the case of the reversing permutation.

The second result proves that only a few special transformations in  $S_B$  preserve self-complementarity.

**Theorem 2** *Let  $\pi \in S_B$ . Then*

$$\pi \circ c = c \circ \pi$$

*if and only if  $\pi(X)$  is a trinucleotide circular self-complementary code for every trinucleotide circular self-complementary code  $X \subseteq \mathcal{B}^3$ .*

*Proof* See Appendix. □

In simple words, the above theorems prove that there are only 8 base transformations in  $S_B$  that, when applied to a trinucleotide self-complementary  $C^3$ -code, generate a code of the same class. Theorems 1 and 2 have remarkable consequences. In fact, the

216 maximal self-complementary  $C^3$ -codes are naturally divided into 27 equivalence classes with 8 codes each. Given a code  $X \in \mathfrak{C}$  it is possible to obtain immediately the other 7 codes of  $\mathfrak{C}$  that are in the same equivalence class by simply applying the following bijective transformations<sup>1</sup> that build a subgroup of  $(S_{\mathcal{B}}, \circ)$ :

$$\begin{aligned} L := & \{id, c, p, r, \pi_{CG} : (A, C, G, T) \mapsto (A, G, C, T), \pi_{AT} : (A, C, G, T) \mapsto (T, C, G, A), \\ & \pi_{ACTG} : (A, C, G, T) \mapsto (C, T, A, G), \pi_{AGTC} : (A, C, G, T) \mapsto (G, A, T, C)\}. \end{aligned}$$

This group  $(L, \circ)$  is not a normal subgroup of  $(S_{\mathcal{B}}, \circ)$  but it is isomorphic to the *dihedral group*  $D_8$  (see [Rotman 1995](#) for more details).  $D_8$  is known in geometry as the *symmetry group of the square*, i.e. all symmetry (distance preserving) mappings of the square. The well-known fact that the centralizer  $\text{Cent}(D_8) = \{\pi \in S_4 : \pi \circ \sigma = \sigma \circ \pi \text{ for all } \sigma \in D_8\}$  of  $D_8$  inside  $S_4$  is the group  $\{id, c\}$  is reflected by the above Theorem 2 (for more details on group theory see [Rotman 1995](#)). In Sect. 3.1 we will explain geometrically why  $(L, \circ)$  is exactly the set of maps from  $S_{\mathcal{B}}$  that commute with  $c$ .

In Table 1 we show the list of the 216 codes, as taken from the lists in [Michel et al. \(2008\)](#), divided into the 27 equivalence classes. The universal  $X_0$  code discovered in [Arquès and Michel \(1996\)](#) is labelled with the number 23. As mentioned before, in [Koch and Lehmann \(1997\)](#) it has been proposed that circular codes are in some sense a byproduct of the frequencies of the bases in the different positions of a codon. However in [Lacan and Michel \(2001\)](#) it has been proved that the universal code  $X_0$  common to Prokaryotes and Eukaryotes cannot be generated in this way. They showed that only 88 of the complete set of 216 maximal self-complementary  $C^3$ -codes can be generated from the proportion of bases (“NF codes”). Thus, the set of 216 codes is bi-partitioned in one subset containing the 88 NF codes and one containing the 128 codes of the  $X_0$  type (non-NF). In Table 1 we have highlighted the 88 codes of NF type; surprisingly, they cover exactly 11 of the 27 equivalence classes. This suggests that the symmetries of the codes reflect indeed their capability of describing protein coding sequences. Moreover, in [Gonzalez et al. \(2011\)](#) it has been shown that, for every given Hamming distance from the  $X_0$  code, the codes that have a good coverage over a sample set of coding sequences are never of the NF type. It becomes clear that the two sub-classes of codes are invariant sets under the transformations in Theorem 2 and Theorem 1. In other words, any transformation of this kind applied to a NF code gives another NF code and the same holds for non-NF codes.

It is interesting to see some other consequences of Theorems 1 and 2 with respect to the properties of a maximal self-complementary  $C^3$ -code. For example, we have mentioned that codons composed of three equal trinucleotides cannot be part of such codes. But also four other codons cannot be part of them:

**Lemma 1** *Let  $X \subseteq \mathcal{B}^3$  be a trinucleotide circular self-complementary code. Then  $X$  cannot contain any codons of the form  $Nc(N)N$  for  $N \in \mathcal{B}$ .*

---

<sup>1</sup> Of course, excluding the identity.

**Table 1** Classification of the 216 circular codes of the class  $\mathfrak{C}$  into the 27 equivalence classes defined by Theorems 2 and 1

| I                           | SW ( $c$ ) | YR ( $p$ ) | KM ( $r$ ) | $\pi_{CG}$ | $\pi_{AT}$ | $\pi_{ACTG}$ | $\pi_{AGTC}$ |
|-----------------------------|------------|------------|------------|------------|------------|--------------|--------------|
| 1                           | 1          | 100        | 29         | 71         | 2          | 99           | 79           |
| 2                           | 3          | 101        | 15         | 90         | 5          | 103          | 43           |
| 3                           | 4          | 102        | 16         | 86         | 6          | 104          | 42           |
| 4                           | 7          | 97         | 54         | 51         | 9          | 95           | 58           |
| 5                           | 8          | 98         | 53         | 52         | 10         | 96           | 55           |
| <b>6</b>                    | <b>11</b>  | <b>91</b>  | <b>21</b>  | <b>78</b>  | <b>39</b>  | <b>68</b>    | <b>74</b>    |
| 7                           | 12         | 88         | 30         | 72         | 38         | 64           | 84           |
| 8                           | 13         | 87         | 23         | 81         | 37         | 65           | 77           |
| 9                           | 14         | 92         | 28         | 70         | 36         | 66           | 82           |
| 10                          | 17         | 89         | 20         | 80         | 40         | 69           | 75           |
| 11                          | 18         | 94         | 63         | 44         | 41         | 67           | 93           |
| 12                          | 24         | 83         | 49         | 57         | 73         | 32           | 48           |
| 13                          | 25         | 85         | 50         | 56         | 76         | 35           | 47           |
| <b>14</b>                   | <b>105</b> | <b>147</b> | <b>123</b> | <b>143</b> | <b>106</b> | <b>150</b>   | <b>141</b>   |
| 15                          | 107        | 148        | 120        | 146        | 112        | 156          | 127          |
| <b>16</b>                   | <b>108</b> | <b>152</b> | <b>125</b> | <b>140</b> | <b>110</b> | <b>149</b>   | <b>142</b>   |
| 17                          | 109        | 153        | 121        | 144        | 114        | 154          | 128          |
| 18                          | 111        | 151        | 119        | 145        | 116        | 159          | 126          |
| <b>19</b>                   | <b>113</b> | <b>158</b> | <b>134</b> | <b>132</b> | <b>115</b> | <b>155</b>   | <b>136</b>   |
| <b>20</b>                   | <b>117</b> | <b>157</b> | <b>130</b> | <b>135</b> | <b>118</b> | <b>160</b>   | <b>131</b>   |
| <b>21</b>                   | <b>161</b> | <b>211</b> | <b>168</b> | <b>207</b> | <b>162</b> | <b>214</b>   | <b>179</b>   |
| <b>22</b>                   | <b>163</b> | <b>215</b> | <b>190</b> | <b>188</b> | <b>165</b> | <b>212</b>   | <b>194</b>   |
| <b>23</b>                   | <b>164</b> | <b>216</b> | <b>189</b> | <b>187</b> | <b>166</b> | <b>213</b>   | <b>191</b>   |
| <b>24</b>                   | <b>167</b> | <b>208</b> | <b>171</b> | <b>204</b> | <b>178</b> | <b>200</b>   | <b>201</b>   |
| 25                          | 169        | 210        | 198        | 180        | 177        | 199          | 209          |
| <b>26</b>                   | <b>172</b> | <b>205</b> | <b>181</b> | <b>195</b> | <b>202</b> | <b>175</b>   | <b>184</b>   |
| <b>27</b>                   | <b>173</b> | <b>206</b> | <b>183</b> | <b>192</b> | <b>203</b> | <b>176</b>   | <b>182</b>   |
| The 88 NF codes are in bold |            |            |            |            |            |              |              |
|                             |            |            |            |            |            |              | <b>193</b>   |

*Proof* Assume that  $X$  contains a codon of the form  $Nc(N)N$ ,  $N \in \mathcal{B}$ . Then its anti-codon  $c(N)Nc(N)$  is also in  $X$  and the word

$$w = Nc(N)Nc(N)Nc(N)$$

has two different decompositions into the words of  $X$  on a circle:

$$w = Nc(N)Nc(N)Nc(N) \quad \text{and} \quad w' = c(N)Nc(N)Nc(N)N.$$

It is a contradiction to the circularity of the code  $X$ .  $\square$

Another consequence of Theorems 1 and 2 is the known fact that the circular permutations of a maximal self-complementary  $C^3$ -code generate circular codes that

are not self-complementary (Arquès and Michel 1996; Bussoli et al. 2012). We give an independent proof for this fact:

**Lemma 2** *Let  $X \subseteq \mathcal{B}^3$  be a self-complementary  $C^3$ -code. Then  $X_1 := \alpha_1(X)$  and  $X_2 := \alpha_2(X)$  are not self-complementary.*

*Proof* Without loss of generality assume that  $X_1$  is self-complementary. Let

$$x := N_1 N_2 N_3 \in X$$

be an arbitrary codon from  $X$ . Since  $X$  is self-complementing  $X$  contains the anticodon of  $x$

$$\overleftarrow{c(x)} = c(N_3)c(N_2)c(N_1) \in X.$$

Then  $X_1$  contains

$$x_1 := N_3 N_1 N_2 \in X_1 \quad \text{and} \quad \alpha_1(\overleftarrow{c(x)}) = c(N_1)c(N_3)c(N_2) \in X_1.$$

We assumed that  $X_1$  is self-complementary. Then  $X_1$  must contain the anticodon of  $x_1$

$$\overleftarrow{c(x_1)} = c(N_2)c(N_1)c(N_3) \in X_1.$$

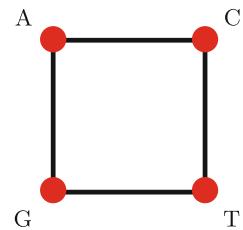
This is a contradiction to the circularity of  $X_1$  since  $c(N_2)c(N_1)c(N_3)$  and also  $c(N_1)c(N_3)c(N_2)$  are in the same conjugacy class.  $\square$

*Remark 2* Note that there are maximal self-complementary codes that are not  $C^3$  as well as maximal  $C^3$ -codes that are not self-complementary. Accordingly to Arquès and Michel (1996) there are 528 maximal self-complementary circular codes therefore only 216 have the  $C^3$ -property. Consequently, there are 312 maximal self-complementary circular codes which are not  $C^3$ -codes. On the other hand, there are 221,328 maximal  $C^3$ -codes (Michel 2013), hence, there are 221,112 maximal  $C^3$ -codes which are not self-complementary.

### 3.1 Geometry of self-complementary circular codes

As we have seen above, there is a subgroup  $L$  of  $S_{\mathcal{B}}$  that preserves the self-complementarity of circular codes ( $C^3$ -codes). In this section we are interested in explaining geometrically why the group  $L$  is exactly the subgroup of  $S_{\mathcal{B}}$  that has this property. Note that by Theorem 2 any map  $\pi \in S_{\mathcal{B}}$  maps a  $C^3$ -code to a  $C^3$ -code.

In the following, a *square* will mean an undirected simple triangle-free graph  $Q = (V(Q), E(Q))$  with sets of vertices  $V(Q)$ ,  $|V(Q)| = 4$  and of edges  $E(Q)$ ,  $|E(Q)| = 4$  between the vertices where edges are unordered pairs  $e = [v, w] \in E(Q)$ ,  $v, w \in V(Q)$ . Our main example will of course be a square  $Q_{\mathcal{B}}$

**Fig. 1** The square  $Q_B$ 

related to the set  $\mathcal{B}$  of bases, i.e.  $Q_B = (V_B, E_B)$  with  $V(Q_B) = \{A, C, G, T\} = \mathcal{B}$  and  $E(Q_B) = \{[A, C], [C, T], [T, G], [G, A]\}$  (see Fig. 1).

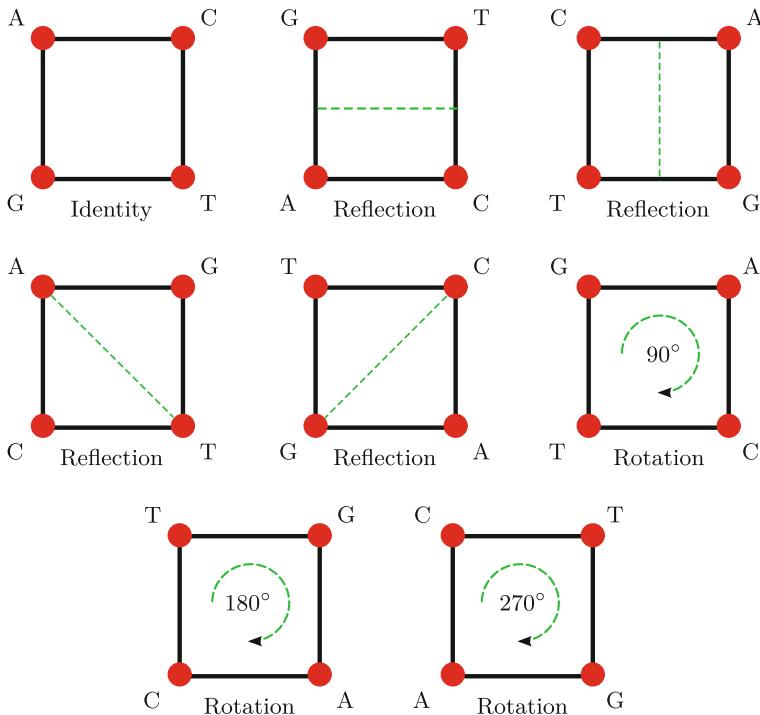
Let us recall that in graph theory, an *isomorphism* of graphs  $G$  and  $H$  is a bijection between the vertex sets of  $G$  and  $H$   $\sigma : V(G) \rightarrow V(H)$  such that any two vertices  $v$  and  $w$  of  $G$  are adjacent in  $G$  if and only if  $\sigma(v)$  and  $\sigma(w)$  are adjacent in  $H$ , i.e.  $\sigma$  is an 'edge-preserving bijection'. In the case, when  $G$  and  $H$  are one and the same graph, the bijection is called an *automorphism* or *isometry (symmetry map)* of  $G$ . It is easy to see that there are only eight automorphisms of a square, namely the identity, the (clockwise) *rotations* of  $90^\circ$ ,  $180^\circ$  and  $270^\circ$ , and the four *reflections*, i.e. two reflections about lines joining midpoints of opposite sides, and two reflections about diagonals. These eight automorphisms, shown in Fig. 2, together with the usual composition as operation, form a group  $\text{Sym}(Q)$ , the *symmetry group* of  $Q$  that is isomorphic to the dihedral group  $D_8$ . For our main example  $Q_B$  defined above, we obtain the group  $L$  as its symmetry group where

$$\begin{aligned} L := & \{id, c, p, r, \pi_{CG} : (A, C, G, T) \mapsto (A, G, C, T), \pi_{AT} : (A, C, G, T) \mapsto (T, C, G, A), \\ & \pi_{ACTG} : (A, C, G, T) \mapsto (C, T, A, G), \pi_{AGTC} : (A, C, G, T) \mapsto (G, A, T, C)\}. \end{aligned}$$

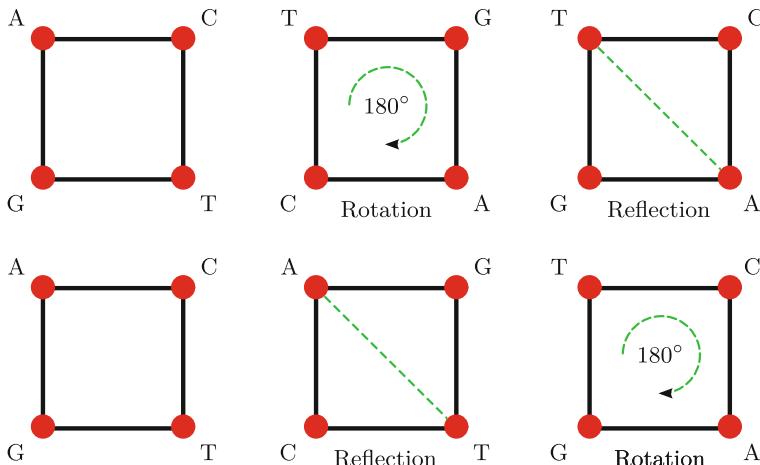
is the group defined in the previous section and that corresponds to the transformations in  $S_B$  preserving self-complementarity of  $C^3$ -codes. It is readily seen that the complementing map  $c$  corresponds to the rotation by  $180^\circ$ . We will denote this rotation by  $\text{rot}_{180}$ . This fact shows geometrically that  $c$  commutes with the maps in  $L$  as it was stated as one part of Theorem 2. In fact, the rotation by  $180^\circ$  is the only automorphism of the square that commutes with all other automorphisms, i.e.  $\{id, \text{rot}_{180}\}$  is the center of  $\text{Sym}(Q)$ . For instance, rotating by  $180^\circ$  and then reflecting about one of the diagonals (Fig. 3, first row) is the same as first reflecting about the diagonal and then rotating by  $180^\circ$  (Fig. 3, second row).

Now, we want to understand geometrically why there is no permutation of the vertices of a square (i.e. no other map in  $S_B$ ) that commutes with the rotation by  $180^\circ$  other than those coming from the automorphisms. In order to see this it is helpful to insert two more edges in the picture, namely the diagonals (see Fig. 4). For the reader's convenience we stay with our main example  $Q_B$ .

If we apply the rotation by  $180^\circ$   $\text{rot}_{180}$ , then the two diagonals are invariant, i.e.  $[A, T]$  goes to  $[A, T]$  and  $[C, G]$  goes to  $[C, G]$ . If  $\pi$  is a permutation of the vertices  $V_B$  that is not an automorphism, then it must map two vertices  $v, w$  to a set of vertices that is not connected by any edge of the square, i.e. the edge  $[v, w]$  must be mapped onto one of the diagonals, say  $d$ , under  $\pi$ . In this case,  $\pi$  would correspond to a



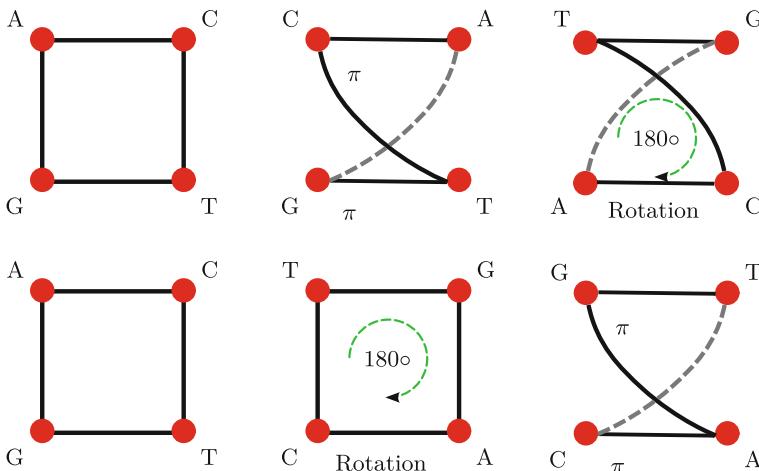
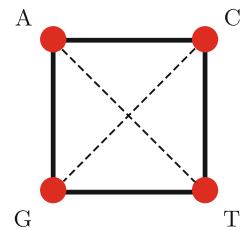
**Fig. 2** The symmetry group of the square  $Q_B$



**Fig. 3** The center of the symmetry group of  $Q_B$  contains  $\{id, \text{rot}_{180}\}$ . For instance it is shown that  $\text{rot}_{180}$  commutes with the reflection about a diagonal. The *first row* shows rotation plus reflection while the *second one* shows the effect of reflection plus rotation

transformation that does not belong to the symmetry group of  $Q_B$ . For instance, in Fig. 5  $\pi$  corresponds to twisting the upper part of the square. Still, it might be the case that such transformation commutes with rotation but Theorem 2 shows that this

**Fig. 4**  $Q_B$  with imaginary diagonals

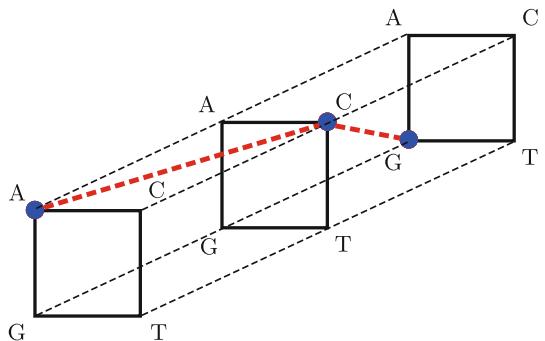


**Fig. 5** Any map  $\pi$  which is not in the symmetry group of the square  $Q_B$  does not commute with  $\text{rot}_{180}$ . This is shown in the figure where  $\pi \circ \text{rot}_{180}$  (first row) is different from  $\text{rot}_{180} \circ \pi$  (second row)

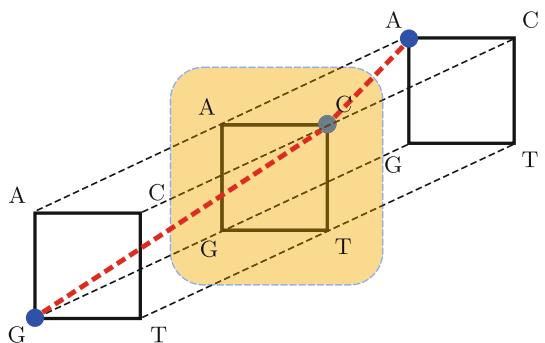
is not possible. Figure 5 shows clearly that if we first apply  $\pi$  and then rotation (first row, from left to right) we obtain a different result from applying rotation and then  $\pi$  (second row, from left to right).

Thus, we have a geometric verification of the fact that the maps in  $L$  are the only transformations in  $S_B$  that commute with the complementing map  $c$  (represented as a rotation by  $180^\circ$ ). Now we will show the geometrical meaning of self-complementarity: consider three squares/bases in a row and connect the corresponding vertices. The geometric figure we obtain is the cuboid shown in Fig. 6, where we have marked the codon  $ACG$ . Again, we are interested in the symmetry group of this object. However, for our purposes it is enough to see how, for a given codon, one can form its anticodon in a geometrical way. Assume that  $x \in \mathcal{B}^3$  is given, then its anticodon is defined as  $\overleftarrow{c}(x)$ . We have seen that forming the complement  $c(x)$  can be interpreted as applying  $\text{rot}_{180}$ , the rotation by  $180^\circ$ , in each of the squares in the cuboid. Moreover, it is obvious that reversion is given by reflection along the plane that is defined by the middle square (see Fig. 7). We will call this reflection  $\text{ref}$ . Thus, by applying in sequence the two automorphisms to a codon (rotation  $\text{rot}_{180}$  and reflection  $\text{ref}$ ) one can form the anticodon geometrically. Such operations are depicted in Figs. 7 and 8, respectively.

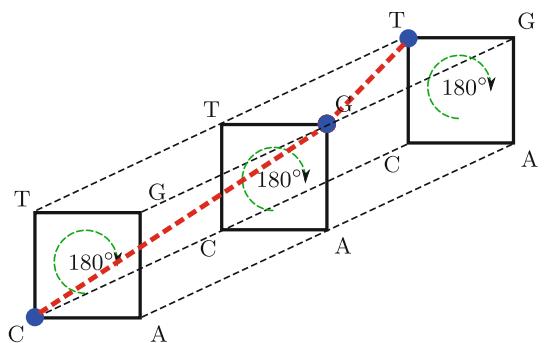
**Fig. 6** The cuboid with the codon  $ACG$  marked



**Fig. 7** The cuboid after reflection along the shaded plane defined by the *middle square*. The reversed codon  $GCA$  is marked



**Fig. 8** The cuboid after reflection ref and rotation  $\text{rot}_{180}$  with the anticodon  $CGT$  marked



The rotations inside the squares and the reflection of the cuboid commute with all the automorphisms of the square applied in each of the squares. Hence, given a self-complementary code  $X$  and an automorphism  $\pi \in L$  we see that geometrically forming the anticodon of  $x \in X$  commutes with applying the transformation  $\pi$  to  $x$ . Thus, the anticodon of  $\pi(x)$  will be the image of the anticodon of  $x$  under  $\pi$  and hence again in  $\pi(X)$  which must then also be self-complementary.

As we have seen it is geometrically clear that the maps in  $L$  preserve self-complementarity since they commute with the *geometrical construction of the anticodon*. It is therefore interesting to ask whether or not there are more automorphisms

(or even just bijective maps) of the cuboid (or even the set of codons) that preserve self-complementarity. The answer is negative if we restrict to maps of the cuboid as the following theorem shows:

**Theorem 3** *A permutation  $\pi$  of the set of vertices of the cuboid preserves self-complementarity of codes if and only if it is an automorphism of the cuboid and commutes with forming the anticodon, i.e.  $\pi(\overleftarrow{c(x)}) = \overleftarrow{c}(\pi(x))$ .*

*Proof* See Appendix. □

*Remark 3* If we consider any bijective maps  $\pi : \mathcal{B}^3 \rightarrow \mathcal{B}^3$  we will find much more possibilities, namely  $32! \cdot 2^{32}$  different maps preserving self-complementarity. Consider the following construction. We divide first the set of all codons  $\mathcal{B}^3$  into two equal-sized subsets  $H_1$  and  $H_2$  so that  $H_2 = \overleftarrow{c(H_1)}$ . There are  $2^{32}$  possibilities to do it: There are 32 codon-anticodon pairs, if we take from each such pair one element we get subsets  $H_1, H_2 := \mathcal{B}^3 \setminus H_1$ . Then we consider an arbitrary bijection  $\pi_1 : H_1 \rightarrow H_1$  ( $32!$  possibilities to choose it) and extend  $\pi_1$  onto  $H_2$  mapping an anticodon to  $b \in H_1$  on an anticodon to  $\pi_1(b)$ .

The bijective mapping  $\pi : \mathcal{B}^3 \rightarrow \mathcal{B}^3$

$$\pi(b) = \begin{cases} \overleftarrow{\pi_1(b)}, & b \in H_1 \\ c(\pi_1(\overleftarrow{c(b)})), & b \notin H_1 \end{cases}$$

will preserve self-complementarity. Such bijections, together with the operation of composition of mappings, form a group. However, not all of these mappings will preserve the  $C^3$ -property.

### 3.2 Structure of a circular code

Given a code  $X \in \mathfrak{C}$  we have that, in view of the self-complementary property, the 20 codons can be divided in two sets (in many different ways):  $X = x \cup c(\overleftarrow{x})$ . This bipartition is shown for the  $X_0$  code (Arquès and Michel 1996) in the first two columns of Table 2 where we also show the codes  $X_1 = \alpha_1(X_0)$  and  $X_2 = \alpha_2(X_0)$ . Notice that both  $X_1$  and  $X_2$  are maximal  $C^3$ -codes but not self-complementary. The structure of  $X_1$  and  $X_2$  can be derived directly from the property listed in Eq. 2. In fact we have

$$\begin{aligned} \alpha_1(c(\overleftarrow{x})) &= c(\overleftarrow{\alpha_2(x)}) \\ \alpha_2(c(\overleftarrow{x})) &= c(\overleftarrow{\alpha_1(x)}) \end{aligned}$$

In practice, since the reversing and the circular permutations do not commute, when we apply the circular permutation  $\alpha_1$  to the set  $c(\overleftarrow{x})$  we obtain the reverse complement of  $\alpha_2(x)$ . For instance, consider the pair of codons  $AAC, GTT \in X_0$ . Clearly,  $GTT$  is the reverse complement of  $AAC$ . Now, if we apply the two circular permutations  $\alpha_1, \alpha_2$  we obtain  $ACA, TTG \in X_1$  and  $CAA, TGT \in X_2$ . This time,  $TTG \in X_1$  ( $ACA \in X_1$ ) is the reverse complement of  $CAA \in X_2$  ( $TGT \in X_2$ ).

**Table 2** Structure of the AM code  $X_0$  (code number 23) and its circular permutations  $X_1 = \alpha_1(X_0)$  and  $X_2 = \alpha_2(X_0)$ 

|    | $X_0$ |                        | $\alpha_1(X_0) = X_1$ |                                  | $\alpha_2(X_0) = X_2$ |                                  |
|----|-------|------------------------|-----------------------|----------------------------------|-----------------------|----------------------------------|
|    | $x$   | $c(\overleftarrow{x})$ | $\alpha_1(x)$         | $\alpha_1(c(\overleftarrow{x}))$ | $\alpha_2(x)$         | $\alpha_2(c(\overleftarrow{x}))$ |
| 1  | AAC   | GTT                    | ACA                   | TTG                              | CAA                   | TGT                              |
| 2  | AAT   | ATT                    | ATA                   | TTA                              | TAA                   | TAT                              |
| 3  | ACC   | GGT                    | CCA                   | GTG                              | CAC                   | TGG                              |
| 4  | ATC   | GAT                    | TCA                   | ATG                              | CAT                   | TGA                              |
| 5  | CAG   | CTG                    | AGC                   | TGC                              | GCA                   | GCT                              |
| 6  | CTC   | GAG                    | TCC                   | AGG                              | CCT                   | GGA                              |
| 7  | GAA   | TTC                    | AAG                   | TCT                              | AGA                   | CTT                              |
| 8  | GAC   | GTC                    | ACG                   | TCG                              | CGA                   | CGT                              |
| 9  | GCC   | GGC                    | CCG                   | GCG                              | CGC                   | CGG                              |
| 10 | GTA   | TAC                    | TAG                   | ACT                              | AGT                   | CTA                              |

Theorem 2 states that as long as we apply one of the 8 admissible transformations (set  $L$ ) we keep all the property of a circular code. Indeed, the complementary transformation  $c$  plays a crucial role in the set  $L$ . Surprisingly, given a code  $X \in \mathfrak{C}$ , then  $c(X) = \overleftarrow{X}$ . In brief, a circular code in  $\mathfrak{C}$  is built in a way such that its complement coincides with its reverse. The result arises immediately by looking at Table 3 where we show a circular code  $X_0 \in \mathfrak{C}$ ,  $X_1 = \alpha_1(X_0)$  and  $X_2 = \alpha_2(X_0)$  together with its complement  $c(X_0)$  and the associated permuted codes  $c(X_1) = \alpha_1(c(X_0))$ ,  $c(X_2) = \alpha_2(c(X_0))$ . Since  $c$  commutes with  $\pi$ , then  $c(X_0)$  is the reverse of  $X_0$  and this is clear by looking at the elements of the two sets. Furthermore, we have the following relations:

$$\begin{aligned} X_1 &= \alpha_1(X_0) = \alpha_1(c(\overleftarrow{X_0})) = c(\alpha_1(\overleftarrow{X_0})) = c(\overleftarrow{\alpha_2(X_0)}) = \overleftarrow{c(X_2)} \\ X_2 &= \alpha_2(X_0) = \alpha_2(c(\overleftarrow{X_0})) = c(\alpha_2(\overleftarrow{X_0})) = c(\overleftarrow{\alpha_1(X_0)}) = \overleftarrow{c(X_1)}; \end{aligned}$$

In other words, the first circular permutation of a code  $X_0$  coincides with the reverse second circular permutation of the complement of  $X_0$  so that the pair  $X_0, c(X_0) = \overleftarrow{X_0}$  together with their circular permutations is in a precise relation. This nice property alert us on the important role of inversion symmetries along coding sequences. In fact, in Gonzalez et al. (2012) it has been shown that the symmetries of complementarity and inversion can be used to derive a complete version of a hypothetical primeval mitochondrial code composed of codons of four letters (tesserae). We conclude the section with a corollary and a remark:

**Corollary 1** *Let  $X \subseteq \mathcal{B}^3$  be a trinucleotide circular code. Then the set of the anti-codons of  $X$ :  $c(X)$  is also a trinucleotide circular code. Furthermore, If  $X$  is a  $C^3$ -code then  $c(X)$  is also a  $C^3$ -code.*

**Table 3** Structure of a code  $X_0 \in \mathfrak{C}$ , its circular permutations  $X_1$  and  $X_2$  and their complement  $c(X_0)$ ,  $c(X_1)$ ,  $c(X_2)$ 

| $X_0$                  | $X_1$                            | $X_2$                            | $c(X_0)$            | $c(X_1)$   | $c(X_2)$   |
|------------------------|----------------------------------|----------------------------------|---------------------|--|--|
| $x$                    | $\alpha_1(x)$                    | $\alpha_2(x)$                    | $c(x)$              | $\overleftarrow{\alpha_2(c(\overleftarrow{x}))}$ | $\overleftarrow{\alpha_1(c(\overleftarrow{x}))}$ |
| $c(\overleftarrow{x})$ | $\alpha_1(c(\overleftarrow{x}))$ | $\alpha_2(c(\overleftarrow{x}))$ | $\overleftarrow{x}$ | $\overleftarrow{\alpha_2(x)}$                    | $\overleftarrow{\alpha_1(x)}$                    |

*Proof* Since  $\overleftarrow{\cdot}$  is a composition of two mappings such that both of them preserve the circularity according to the lemma above the claim follows. If  $X$  is a  $C^3$ -code then  $X_1 := \alpha_1(X)$  and  $X_2 := \alpha_2(X)$  are also circular. The sets of the anticodons of  $X$ ,  $X_1$  and  $X_2$  are as just proved also circular codes. Besides, the properties

$$\overleftarrow{c(X_1)} = \alpha_2(\overleftarrow{c(X)}), \quad \overleftarrow{c(X_2)} = \alpha_1(\overleftarrow{c(X)})$$

take place. So  $\overleftarrow{c(X)}$  is also a  $C^3$ -code.  $\square$

*Remark 4* The union of a set of codons and their anticodons might not form a circular code even if the two sets separately are circular codes: for instance, *TAT* is the anticodon of *ATA*, the codes  $Y = \{\text{ATA}\}$  and  $Z = \{\text{TAT}\}$  are obviously circular but  $X = Y \cup Z$  is not a circular code.

## 4 Conclusions

Circular codes represent a key aspect of the organization of genetic information related to the capability of maintaining the correct reading frame in protein synthesis. A particular kind of circular codes, namely, maximal,  $C^3$  and self-complementary codes show more or less universal properties across all domains of life, including prokaryotes and eukaryotes (Arquès and Michel 1996). In this work we have investigated the symmetry properties of circular codes and established clear connections with group theory and transformations. Previous studies (Arquès and Michel 1996) proved empirically that there is a circular code ( $X_0$ ) that, on average, has the best covering capability across organisms of different species. However, there is a high variability across organisms so that in some instances the code  $X_0$  has a very low coverage whereas other codes provide a much better description (Gonzalez et al. 2011). Hence, it is likely that the biological functions associated to circular codes be related to a set of codes rather than to a single one.

In this paper we have proved two general theorems that allow to predict the consequences of the action of the 24 possible nucleotide bijections on the structure of circular codes. The importance of the symmetric group  $S_B$  for the study of circular codes was also suggested in Michel and Pirillo (2011) and Michel et al. (2012). We found that any bijection preserves the properties of circularity and  $C^3$ . Moreover, the set of 216 maximal,  $C^3$  and self-complementary codes, are invariant under the action of a transformation subgroup of the symmetric group. This subgroup is isomorphic

to the dihedral group and its elements commute with the complementary transformation. The dihedral group allows the classification of circular codes in 27 equivalence classes. Such classification has a surprising biological implication. In fact, the set of 216  $C^3$  maximal and self-complementary codes can be partitioned in two subsets: the first one (NF) contains the codes that can be generated in the framework of the hypothesis proposed in Koch and Lehmann (1997), while the second one (non NF) contains the codes that cannot be generated in such a way (Lacan and Michel 2001). Now, we have shown that the set of NF codes covers exactly 11 of the 27 equivalence classes and this proves that the symmetry structure implied by the group theoretic framework characterizes codes that are related through a biological hypothesis. This results might be related to the findings of Michel (2013) where they find a similar partition of a particular set of 27 codes by means of a different approach based on the search of forbidden combinations of codons.

Since the dihedral group is isomorphic to the symmetry transformations of a square, we have provided an intuitive geometrical interpretation of these transformations. Among other properties, we have illustrated the combined action of the reverse and the complementary transformations in geometrical terms. Starting from a codon it is possible to derive the anti-codon through geometric arguments. Moreover, based on symmetry arguments, we have provided hints on the internal structure of circular codes and this confirms the importance of the complementary and reverse transformations that have been highlighted in many different contexts (Gonzalez et al. 2012).

The origin of circular codes is still controversial; their existence can be related to comma-free (self-synchronizable) codes in primeval organisms and they might play a fundamental role in maintaining the normal reading frame in protein synthesis. The study of their possible evolution, for example by transition and transversion mutations (Benard and Michel 2013), represent a challenging research area, being circularity properties necessarily associated to the hypothetical biological functions. By means of a clear theoretical framework, our work contributes also to shed light into the general conditions under which such mutations preserve or not these essential properties. Moreover, it highlights the essential role of symmetries, and in particular of the dihedral group, in classifying and interpreting genetic information.

**Acknowledgments** We would like to thank Alberto Danielli for useful discussions.

## Appendix A: Proofs

### Proof of Theorem 1

*Proof* We will write for a codon  $x_i \in X$   $x_i = B_1^i B_2^i B_3^i$ ,  $B_j^i \in \mathcal{B}$ ,  $j = 1, 2, 3$ .

1. Let us show first that  $\overleftarrow{X}$  is a trinucleotide circular code. The reverse codon to  $x_i$  has the form  $\overleftarrow{x_i} = B_3^i B_2^i B_1^i$ . Assume that  $\overleftarrow{X}$  is not circular and the word

$$w = \overleftarrow{x_1} \cdots \overleftarrow{x_k} = B_3^1 B_2^1 B_1^1 \cdots B_3^k B_2^k B_1^k, \quad x_i \in X$$

has at least two decompositions into the words from  $\overleftarrow{X}$  written on a circle. Without lost of generality let us assume that the second decomposition occurs with a shift by 1. That means that for all  $1 \leq i < k$

$$B_2^i B_1^i B_3^{i+1} \in \overleftarrow{X} \quad \text{and} \quad B_2^k B_1^k B_3^1 \in \overleftarrow{X}.$$

That means that for all  $1 \leq i < k$

$$B_3^{i+1} B_1^i B_2^i \in X \quad \text{and} \quad B_3^1 B_1^k B_2^k \in X.$$

So the word

$$w' = x_k x_{k-1} \cdots x_1 = B_1^k B_2^k B_3^k B_1^{k-1} B_2^{k-1} B_3^{k-1} \cdots B_1^1 B_2^1 B_3^1$$

has at least two decompositions into the words from  $X$  with a shift by 2.

Similar arguments work when the second decomposition was obtained by shift of 2 positions.

Let us show now with a counter-example that the remaining four permutations of the bases  $\alpha \in S_3 \setminus \{id, \leftrightarrow\}$  do not guarantee the circularity of  $\alpha(X)$ :

Let us denote the permutations

$$p_1 = (21)(3), \quad p_2 = (1)(32), \quad \alpha_1 = (213), \quad \alpha_2 = (312)$$

and consider for example  $X = \{TAA, ATT\}$ .  $X$  and  $Y := \alpha_1(X) = \{AAT, TTA\}$  are both circular. But  $\alpha_2(X) = \alpha_1(Y) = p_1(X) = p_2(Y) = \{ATA, TAT\}$  is not circular since the word  $w = ATATAT$  has two decompositions into the words of  $X$  on a circle:  $w = ATA, TAT$  and  $w' = TAT, ATA$ .

2. Assume that  $\pi(X)$  is not circular and the word

$$w = \pi(x_1) \cdots \pi(x_k) = \pi(B_1^1) \pi(B_2^1) \pi(B_3^1) \cdots \pi(B_1^k) \pi(B_2^k) \pi(B_3^k), \quad x_i \in X$$

has at least two decompositions into the words from  $\pi(X)$  written on a circle. Without lost of generality let us assume that the second decomposition occurs with a shift by 1. That means that for all  $1 \leq i < k$

$$\pi(B_2^i) \pi(B_3^i) \pi(B_1^{i+1}) \in \pi(X) \quad \text{and} \quad \pi(B_2^k) \pi(B_3^k) \pi(B_1^1) \in \pi(X).$$

It implies that for all  $1 \leq i < k$

$$B_2^i B_3^i B_1^{i+1} \in X \quad \text{and} \quad B_2^k B_3^k B_1^1 \in X.$$

In this case the word  $w' = \pi^{-1}(w)$  has at least two decompositions into the words from  $X$  on a circle. This is a contradiction to the circularity of  $X$ .

Similar arguments work when the second decomposition was obtained by shift of 2 positions.

For all  $\alpha \in S_3$  and  $\pi \in S_B$  the property

$$\alpha(\pi(X)) = \pi(\alpha(X))$$

is true. By the definition of a  $C^3$ -code  $X_1 := \alpha_1(X)$  and  $X_2 := \alpha_2(X)$  are trinucleotide circular codes. The arguments above show that  $\pi(X), \pi(X_1) = \alpha_1(\pi(X))$  and  $\pi(X_2) = \alpha_2(\pi(X))$  are circular codes. That means that  $\pi(X)$  is a  $C^3$ -code.

## Proof of Theorem 2

*Proof* According to the theorem above  $\pi(X)$  is circular. We prove that  $\pi(X)$  is self-complementary:

$$\overleftarrow{\overleftarrow{c(\pi(X))}} = \overleftarrow{\pi(c(X))} = \pi(\overleftarrow{c(X)}) = \pi(X)$$

because of the self-complementarity of  $X$ , the property  $\pi \circ c = c \circ \pi$  and the fact that for all  $\alpha \in S_3$  and  $\pi \in S_B$  the property

$$\alpha(\pi(X)) = \pi(\alpha(X))$$

is true.

Let us list all  $\pi \in S_B$  satisfying  $\pi \circ c = c \circ \pi$ : It is easy to prove that such maps build a subgroup of  $(S_B, \circ)$ . Consequently, the number of such maps must be a factor of 24. The following 8 bijective transformations have this property and build a subgroup of  $(S_B, \circ)$  (easy to check):

$$\begin{aligned} L := \{id, c, p, r, \pi_{CG} : (A, C, G, T) \mapsto (A, G, C, T), \pi_{AT} : (A, C, G, T) \mapsto (T, C, G, A), \\ \pi_{ACTG} : (A, C, G, T) \mapsto (C, T, A, G), \pi_{AGTC} : (A, C, G, T) \mapsto (G, A, T, C)\}. \end{aligned}$$

To show that we found all  $\pi \in S_B$  satisfying  $\pi \circ c = c \circ \pi$  and to exclude the cases of 24 or 12 elements let us add that for example for

$$\pi : A, C, G, T \mapsto C, A, G, T \text{ we have } c \circ \pi(A) = T \neq G = \pi \circ c(A)$$

and it cannot be that we have twelve such maps since 8 is not a factor of 12.

Each  $\pi \in S_B$  preserves according the theorem above the circularity of  $X$ . Let us show now with a counterexample that it is not the case with the self-complementarity if  $\pi \in S_B \setminus L$  does not commute with  $c$ : Consider for example the circular self-complementary code  $X := \{CTG, CAG\}$ . For

- 
- $\pi_{AC} : A, C, G, T \mapsto C, A, G, T$  we get  $\pi_{AC}(X) = \{ATG, ACG\}$ ,  
 $\pi_{AG} : A, C, G, T \mapsto G, C, A, T$  we get  $\pi_{AG}(X) = \{ATA, CGA\}$ ,  
 $\pi_{TG} : A, C, G, T \mapsto A, C, T, G$  we get  $\pi_{TG}(X) = \{CGT, CAT\}$ ,  
 $\pi_{TC} : A, C, G, T \mapsto A, T, G, C$  we get  $\pi_{TC}(X) = \{TCG, TAG\}$ ,  
 $\pi_{ATCG} : A, C, G, T \mapsto T, G, A, C$  we get  $\pi_{ATCG}(X) = \{GCA, GTA\}$ ,  
 $\pi_{ATGC} : A, C, G, T \mapsto T, A, C, G$  we get  $\pi_{ATGC}(X) = \{AGC, ATC\}$ ,  
 $\pi_{TACG} : A, C, G, T \mapsto C, G, T, A$  we get  $\pi_{TACG}(X) = \{GAT, GCT\}$ ,  
 $\pi_{TAGC} : A, C, G, T \mapsto G, T, C, A$  we get  $\pi_{TAGC}(X) = \{ACT, AGC\}$ ,  
 $\pi_{ATC} : A, C, G, T \mapsto T, A, G, C$  we get  $\pi_{ATC}(X) = \{ACG, ATG\}$ ,  
 $\pi_{TAC} : A, C, G, T \mapsto C, T, G, A$  we get  $\pi_{TAC}(X) = \{TAG, TCG\}$ ,  
 $\pi_{ATG} : A, C, G, T \mapsto T, C, A, G$  we get  $\pi_{ATG}(X) = \{CGA, CTA\}$ ,  
 $\pi_{TAG} : A, C, G, T \mapsto G, C, T, A$  we get  $\pi_{TAG}(X) = \{CAT, CGT\}$ ,  
 $\pi_{GTC} : A, C, G, T \mapsto A, G, T, C$  we get  $\pi_{GTC}(X) = \{GCT, GAT\}$ ,  
 $\pi_{TGC} : A, C, G, T \mapsto A, T, C, G$  we get  $\pi_{TGC}(X) = \{TGC, TAC\}$ ,  
 $\pi_{AGC} : A, C, G, T \mapsto G, A, C, T$  we get  $\pi_{AGC}(X) = \{ATC, AGC\}$ ,  
 $\pi_{GAC} : A, C, G, T \mapsto C, G, A, T$  we get  $\pi_{GAC}(X) = \{GTA, GCA\}$ .

In each case we get a non-self-complementary code.

### Proof of Theorem 3

*Proof* Let  $\pi$  be any permutation of the set of vertices of the cuboid and take a self-complementary code  $X$ . Let  $x \in X$ . Then the anticodon  $\overleftarrow{c(\pi(x))}$  of the image of  $x$  under  $\pi$  must be contained in  $\pi(X)$ , hence is of the form  $\pi(\overleftarrow{x'})$  for some  $x' \in X$ . Now choose a self-complementary code  $Y$  with  $X \cap Y = \{x, c(x)\}$ . Then again  $\overleftarrow{c(\pi(x))}$  must be in  $\pi(Y)$  but by assumption this can only be the case if  $c(\pi(x)) = \pi(c(x))$ , hence  $\pi$  commutes with forming the anticodon.

Last but not least assume that a permutation  $\pi$  of the set of vertices of the cuboid commutes with forming the anticodon, hence preserves self-complementarity, but is not an automorphism. It is easy to see that  $\pi$  must preserve degrees of vertices since it commutes with ref. Thus  $\pi$  implies a permutation on the middle square which therefore has to be an automorphism of the middle square because it is assumed to commute with  $\text{rot}_{180}$ . Again commuting with ref shows that also the outer squares must either be invariant or be reflected onto each other followed by an automorphism of the square.

### References

- Arquès DG, Michel CJ (1996) A complementary circular code in the protein coding genes. J Theor Biol 182:45–58

- Benard E, Michel CJ (2013) Transition and transversion on the common trinucleotide circular code. *Comput Biol J* ID 795418:10
- Bussoli L, Michel CJ, Pirillo G (2012) On conjugation partitions of sets of trinucleotides. *Appl Math* 3:107–112
- Crick FHC, Griffith JS, Orgel LE (1957) Codes without commas. *Proc Natl Acad Sci USA* 43(5):416–421
- Fimmel E, Danielli A, Strüngmann L (2013) On dichotomic classes and bijections of the genetic code. *J Theor Biol* 336:221–230
- Frey G, Michel CJ (2006) Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. *Comput Biol Chem* 30:87–101
- Giannerini S, Gonzalez DL, Rosa R (2012) DNA, dichotomic classes and frame synchronization: a quasi-crystal framework. *Philos Trans R Soc A Math Phys Eng Sci* 370(1969):2987–3006
- Golomb SW, Gordon B, Welch LR (1958) Comma-free codes. *Can J Math* 10:202–209
- Gonzalez DL (2004) Can the genetic code be mathematically described? *Med Sci Monit* 10(4):11–17
- Gonzalez DL (2008) The mathematical structure of the genetic code. In: Barbieri M, Hoffmeyer J (eds) *The codes of life: the rules of macroevolution*. Biosemiotics, vol 1, chap 8. Springer, Netherlands, pp 111–152
- Gonzalez DL, Giannerini S, Rosa R (2008) Strong short-range correlations and dichotomic codon classes in coding DNA sequences. *Phys Rev E* 78(5, ID 051918)
- Gonzalez DL, Giannerini S, Rosa R (2011) Circular codes revisited: a statistical approach. *J Theor Biol* 275(1):21–28
- Gonzalez DL, Giannerini S, Rosa R (2012) On the origin of the mitochondrial genetic code: towards a unified mathematical framework for the management of genetic information. *Nat Preced.* doi:[10.1038/npre.2012.7136.1](https://doi.org/10.1038/npre.2012.7136.1)
- Hayes B (1998) The invention of the genetic code. *Am Sci* 86(1):8–14
- Koch AJ, Lehmann J (1997) About a symmetry of the genetic code. *J Theor Biol* 189:171–174
- Lacan J, Michel CJ (2001) Analysis of a circular code model. *J Theor Biol* 213:159–170
- Lewin B (2004) *Genes 8*. Pearson Prentice Hall, Upper Saddle River
- Michel CJ (2008) A 2006 review of circular codes in genes. *Comput Math Appl* 55:984–988
- Michel CJ (2013) Private communication
- Michel CJ, Pirillo G (2010) Identification of all trinucleotide circular codes. *Comput Biol Chem* 34(2):122–125
- Michel CJ, Pirillo G (2011) Strong trinucleotide circular codes. *Int J Comb* 2011(ID 659567)
- Michel CJ, Pirillo G, Pirillo MA (2008) A relation between trinucleotide comma-free codes and trinucleotide circular codes. *Theor Comput Sci* 401(1–3):17–26
- Michel CJ, Pirillo G, Pirillo MA (2012) A classification of 20-trinucleotide circular codes. *Inf Comput* 212:55–63
- Rotman JJ (1995) *An introduction to the theory of groups*. Springer, Berlin