PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Cite this article: Fimmel E, Michel CJ, Strüngmann L. 2016 *n*-Nucleotide circular codes in graph theory. *Phil. Trans. R. Soc. A* **374**: 20150058.
http://dx.doi.org/10.1098/rsta.2015.0058

Accepted: 5 September 2015

One contribution of 21 to a theme issue 'DNA as information'.

Subject Areas:
graph theory, bioinformatics, biomathematics

Keywords:
circular code, comma-free code, tournament, reading frame

Author for correspondence:
Christian J. Michel
e-mail: c.michel@unistra.fr

n-Nucleotide circular codes in graph theory

Elena Fimmel[1], Christian J. Michel[2] and Lutz Strüngmann[1]

[1]Faculty for Computer Sciences, Institute of Mathematical Biology, Mannheim University of Applied Sciences, Mannheim 68163, Germany
[2]Theoretical bioinformatics, ICube, University of Strasbourg, CNRS, 300 Boulevard Sébastien Brant, Illkirch 67400, France

The circular code theory proposes that genes are constituted of two trinucleotide codes: the classical genetic code with 61 trinucleotides for coding the 20 amino acids (except the three stop codons $\{TAA, TAG, TGA\}$) and a circular code based on 20 trinucleotides for retrieving, maintaining and synchronizing the reading frame. It relies on two main results: the identification of a maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses (Michel 2015 *J. Theor. Biol.* **380**, 156–177. (doi:10.1016/j.jtbi.2015.04.009); Arquès & Michel 1996 *J. Theor. Biol.* **182**, 45–58. (doi:10.1006/jtbi.1996.0142)) and the finding of X circular code motifs in tRNAs and rRNAs, in particular in the ribosome decoding centre (Michel 2012 *Comput. Biol. Chem.* **37**, 24–37. (doi:10.1016/j.compbiolchem.2011.10.002); El Soufi & Michel 2014 *Comput. Biol. Chem.* **52**, 9–17. (doi:10.1016/j.compbiolchem.2014.08.001)). The universally conserved nucleotides A1492 and A1493 and the conserved nucleotide G530 are included in X circular code motifs. Recently, dinucleotide circular codes were also investigated (Michel & Pirillo 2013 *ISRN Biomath.* **2013**, 538631. (doi:10.1155/2013/538631); Fimmel *et al.* 2015 *J. Theor. Biol.* **386**, 159–165. (doi:10.1016/j.jtbi.2015.08.034)). As the genetic motifs of different lengths are ubiquitous in genes and genomes, we introduce a new approach based on graph theory to study in full generality n-nucleotide circular codes X, i.e. of length 2 (dinucleotide), 3 (trinucleotide), 4 (tetranucleotide), etc. Indeed, we prove that an n-nucleotide code X is circular if and only if the corresponding graph $\mathcal{G}(X)$ is acyclic. Moreover, the

THE ROYAL SOCIETY PUBLISHING

maximal length of a path in $\mathcal{G}(X)$ corresponds to the window of nucleotides in a sequence for detecting the correct reading frame. Finally, the graph theory of tournaments is applied to the study of dinucleotide circular codes. It has full equivalence between the combinatorics theory (Michel & Pirillo 2013 *ISRN Biomath.* **2013**, 538631. (doi:10.1155/2013/538631)) and the group theory (Fimmel *et al.* 2015 *J. Theor. Biol.* **386**, 159–165. (doi:10.1016/j.jtbi.2015.08.034)) of dinucleotide circular codes while its mathematical approach is simpler.

1. Introduction

Trinucleotide codes, such as the genetic code, provide a fascinating theory that combines the search for solutions to old and open problems with modern techniques from different fields of science. About 60 years ago, before the discovery of the genetic code, a class of trinucleotide codes, called comma-free codes, was proposed by Crick *et al.* [1] for explaining how the reading of a sequence of trinucleotides could code amino acids. In particular, how the correct reading frame can be retrieved and maintained. The four nucleotides $\{A, C, G, T\}$ as well as the 16 dinucleotides $\{AA, \ldots, TT\}$ are simple codes which are not appropriate for coding 20 amino acids. However, trinucleotides induce a redundancy in their coding. Thus, Crick *et al.* [1] conjectured that only 20 trinucleotides among the 64 possible trinucleotides $\{AAA, \ldots, TTT\}$ code the 20 amino acids. Such a bijective code implies that the coding trinucleotides are found only in one frame—the comma-freeness property. The determination of a set of 20 trinucleotides forming a comma-free code has several necessary conditions:

(i) A periodic trinucleotide from the set $\{AAA, CCC, GGG, TTT\}$ must be excluded from such a code. Indeed, the concatenation of AAA with itself, for instance, does not allow the (original) reading frame to be retrieved as there are three possible decompositions: $\ldots AAA, AAA, AAA \ldots$ (original frame), $\ldots A, AAA, AAA, AA \ldots$ and $\ldots AA, AAA, AAA, A \ldots$, the commas showing the adopted decomposition.

(ii) Two non-periodic permuted trinucleotides, i.e. two trinucleotides related by a circular permutation, e.g. ACG and CGA, must also be excluded from such a code. Indeed, the concatenation of ACG with itself, for instance, does not allow the reading frame to be retrieved as there are two possible decompositions: $\ldots ACG, ACG, ACG \ldots$ (original frame) and $\ldots A, CGA, CGA, CG \ldots$

Therefore, by excluding the four periodic trinucleotides and by gathering the 60 remaining trinucleotides in 20 classes of three trinucleotides such that, in each class, the three trinucleotides are deduced from each other by a circular permutation, e.g. ACG, CGA and GAC, we see that a comma-free code can contain only one trinucleotide from each class and thus has at most 20 trinucleotides. This trinucleotide number is identical to the amino acid number, thus leading to a code assigning one trinucleotide per amino acid without ambiguity. A few combinatorial results on trinucleotide comma-free codes were obtained by Golomb *et al.* [2,3]. However, no trinucleotide comma-free code was identified in genes statistically. Furthermore, at the beginning of the 1960s, the discovery that the trinucleotide TTT, an excluded trinucleotide in a comma-free code, codes phenylalanine [4] led to the abandonment of the concept of comma-freeness over the alphabet $\{A, C, G, T\}$. For several biological reasons, in particular the interaction between mRNA and tRNA, this concept was again taken up later over the purine/pyrimidine alphabet $\{R, Y\}$ ($R = \{A, G\}$, $Y = \{C, T\}$) with two trinucleotide comma-free codes: RRY [5] and $RNY = \{RRY, RYY\}$ (N being any letter on $\{R, Y\}$) [6]. Some statistical results studying and identifying these two comma-free codes were obtained at the sequence level by Shepherd [7] and at the population level by Michel [8]. In 1986, it was shown that introns, in contrast to exons, have no nucleotide periodicity modulo 3 ([9, fig. 2], with a statistical analysis of 90 introns). One year later, with the increase in sequence data, a nucleotide periodicity modulo 2 was identified in introns by two different statistical methods [10,11]. So far, no circular code has been found in introns.

In 1996, a statistical analysis of occurrence frequencies of the 64 trinucleotides $\{AAA, \ldots, TTT\}$ in the three frames of genes of both prokaryotes and eukaryotes showed that the trinucleotides are not uniformly distributed in these three frames [12]. By excluding the four periodic trinucleotides and by assigning each trinucleotide to a preferential frame (frame of highest occurrence frequency), three subsets $X = X_0$, X_1 and X_2 of 20 trinucleotides were found in the frames 0 (reading frame), 1 (frame 0 shifted by one nucleotide) and 2 (frame 0 shifted by two nucleotides) in genes of both prokaryotes and eukaryotes. This set X contains the following 20 trinucleotides:

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

The two sets X_1 and X_2, of 20 trinucleotides each, in the shifted frames 1 and 2, respectively, of genes can be deduced from X by a circular permutation. These three trinucleotide sets present several strong mathematical properties, particularly the fact that X is a maximal C^3 self-complementary trinucleotide circular code [12]. The subset $\{CAG, CTC, CTG, GAG\}$ of the circular code X is a trinucleotide comma-free code and, furthermore, C^3 and self-complementary [13]. Comma-free codes are important codes as they represent a limit class of circular codes with three particular properties: (i) no word of a comma-free code is found in a shifted frame; (ii) the length of reading frame retrieval is the shortest; and (iii) the probability of reading frame coding is maximal and equal to 1 [14]. In 2015, by quantifying the approach used in 1996 for identifying a preferential frame and by applying a massive statistical analysis of gene taxonomic groups, the circular code X was strengthened in genes of prokaryotes (7 851 762 genes, 2 481 566 882 trinucleotides) and eukaryotes (1 662 579 genes, 824 825 761 trinucleotides), and has now also been identified in genes of plasmids (237 486 genes, 68 244 356 trinucleotides) and viruses (184 344 genes, 45 688 798 trinucleotides) [15]. Several non-maximal C^3 self-complementary circular codes have been identified in genes of viruses (table 1 in [15]) which are all subsets of X (table 7d in [15]):

— $X \backslash \{CAG, CTG, GTA, TAC\}$ of 16 trinucleotides in genes of double-stranded DNA viruses (172 198 genes, 39 934 299 trinucleotides) and single-stranded RNA viruses (4492 genes, 3 510 773 trinucleotides);
— $X \backslash \{ACC, CAG, CTC, CTG, GAG, GCC, GGC, GGT\}$ of 12 trinucleotides in genes of double-stranded RNA viruses (973 genes, 654 931 trinucleotides);
— $X \backslash \{CAG, CTC, CTG, GAG, GCC, GGC, GTA, TAC\}$ of 12 trinucleotides in genes of single-stranded DNA viruses (3562 genes, 796 401 trinucleotides);
— $X \backslash \{AAC, ACC, CAG, CTG, GCC, GGC, GGT, GTA, GTT, TAC\}$ of 10 trinucleotides in genes of retro-transcribing viruses (559 genes, 269 070 trinucleotides); and
— $X \backslash \{CAG, CTG, CTC, GAG, GTA, TAC\}$ of 14 trinucleotides in genes of phages (2560 genes, 523 324 trinucleotides).

A trinucleotide circular code has the fundamental property to always retrieve the reading frame in any position of any sequence generated with the circular code. In particular, initiation and stop trinucleotides as well as any frame signals are not necessary to define the reading frame. Indeed, a window of a few nucleotides, whose nucleotide length depends on the class of circular codes, positioned anywhere in a sequence generated with the circular code always retrieves the reading frame.

As an example let us take the word $w = \ldots AGGTAATTACCAG \ldots$ of the circular code X. Is the first nucleotide of w, i.e. A, the 1st, the 2nd or the 3rd nucleotide of a trinucleotide of X? By trying the three possible factorizations (frames) w_0, w_1 and w_2 (w_1 and w_2 being w_0 shifted by one and two nucleotides, respectively) into trinucleotides of X, only one factorization, i.e. w_1, is possible. Thus, the first nucleotide A of w is the 3rd nucleotide of a trinucleotide of X. Indeed, the factorization w_1 leads to the trinucleotides NNA, GGT, AAT, TAC and CAG (N being any appropriate letter of X) which belong to X. The factorizations w_0 and w_2 are impossible as no trinucleotide of X starts with the prefix AG. This case occurs immediately for w_0 and after

11 letters for w_2. Thus, the unique factorization of w is $w_1 = \ldots A, GGT, AAT, TAC, CAG, \ldots$. This word w can be located anywhere in a sequence of X, i.e. the sequence of X does not require an initiator codon, a stop codon or any frame signal to retrieve the reading frame. The word $w' = AGGTAATTACCA$ (w without the last G) with a length of 12 nucleotides is ambiguous as it has two factorizations w_1 and w_2 into trinucleotides of X. The word w' is called an ambiguous word of X. By definition of a circular code, all the ambiguous words are finite words. The word w', taken as an example here, is one of the four longest ambiguous words of X (see below). Thus, the window length l to retrieve the construction frame of a word of a circular code Y is the letter length of the longest ambiguous words w', plus one letter. With the circular code X, $l = 12 + 1 = 13$ nucleotides [16]. The window lengths l for the trinucleotide circular codes X_1 and X_2 are also equal to $l = 13$ nucleotides [16]. In conclusion, the retrieval of the reading frame with the circular code X, the frame 1 with the circular code X_1 and the frame 2 with the circular code X_2 needs the same window length l of 13 nucleotides ($l \geq 13$).

In 2012, in addition to the circular code X in genes (mRNA), a second major step of this circular code theory was revealed by the identification of X motifs, i.e. motifs generated with the circular code X, in the 5′ and/or 3′ regions of tRNAs of prokaryotes and eukaryotes [13,17] and 16S rRNAs, in particular in the ribosome decoding centre where the universally conserved nucleotides A1492 and A1493 and the conserved nucleotide G530 are included in the X motifs [13,18]. A three-dimensional visualization of X motifs in the ribosome shows several spatial configurations involving mRNA X motifs, tRNA X motifs and 16S rRNA X motifs [13,18]. These results led to the concept of a possible translation (framing) code based on the circular code which was proposed in Michel [13].

Trinucleotides are the fundamental words for genes. Dinucleotides are also words with important biological functions in genomes as they are involved in some genome sites, e.g. the splice sites of introns in eukaryotic genomes are based on the dinucleotides GT and AT [19,20]; and in some genome regions, e.g. the dinucleotide CG in animal and plant genomes allows a positive or negative control over gene expression [21]; the dinucleotides CA [22,23], CT [24] and TG [25] in eukaryotic genomes occur as concatenated words $(l_1 l_2)^+, l_1, l_2 \in \{A, C, G, T\}$ (called tandem repeats in biology), etc. Thus, dinucleotide circular codes have been studied according to two approaches, by the combinatorics theory [26] and the group theory [27].

As the genetic motifs of different lengths are ubiquitous in genes and genomes, we introduce here a new approach based on graph theory to study in full generality n-nucleotide circular codes X, i.e. of length 2 (dinucleotide), 3 (trinucleotide), 4 (tetranucleotide), etc. To each such code, a graph is associated and the main theorem states that an n-nucleotide code X is circular if and only if the corresponding graph $\mathcal{G}(X)$ is acyclic. Moreover, many properties of the codes can be seen in its representing graph.

2. The graph theory of n-nucleotide circular codes

Throughout this section, let $\mathcal{B} = \{A, C, G, T\}$ be the set of nucleotide bases, where A stands for *adenine*, C stands for *cytosine*, G stands for *guanine* and T stands for *thymine*. For $n \in \mathbb{N}$ with $n \geq 2$ an *n-nucleotide code* is a subset $X \subseteq \mathcal{B}^n$. The following definition relates a directed graph to any n-nucleotide code. Recall from graph theory [28] that a *graph* $\mathcal{G}$ consists of a finite set of *vertices (nodes)* V and a finite set of *edges* E. Here, an edge is a set $\{v, w\}$ of vertices from V. The graph is called *oriented* if the edges have an orientation, i.e. edges are considered to be ordered pairs $[v, w]$ in this case.

Definition 2.1. Let $X \subseteq \mathcal{B}^n$ be an n-nucleotide code ($n \in \mathbb{N}$). We define a directed graph $\mathcal{G}(X) = (V(X), E(X))$ with a set of vertices $V(X)$ and a set of edges $E(X)$ as follows:

— $V(X) = \{N_1 \ldots N_i, N_{i+1} \ldots N_n : N_1 N_2 N_3 \ldots N_n \in X, 1 \leq i \leq n-1\}$
— $E(X) = \{[N_1 \ldots N_i, N_{i+1} \ldots N_n] : N_1 N_2 N_3 \ldots N_n \in X, 1 \leq i \leq n-1\}$.

The graph $\mathcal{G}(X)$ is called the *representing graph* of X or the graph *associated* with X.

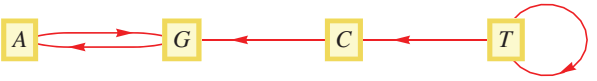

Figure 1. Graph representing the dinucleotide code {*AG*, *CG*, *GA*, *TC*, *TT*}. (Online version in colour.)

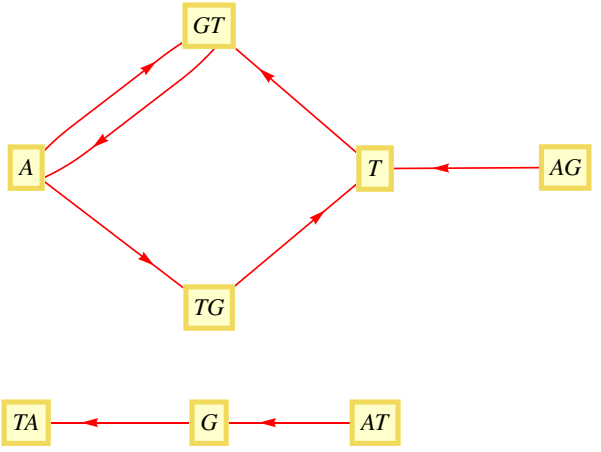

Figure 2. Graph representing the trinucleotide code {*AGT*, *ATG*, *GTA*, *TGT*}. (Online version in colour.)

Basically, the graph $\mathcal{G}(X)$ associated with a code X interprets n-nucleotide words from X in $(n-1)$ ways by pairs of i-nucleotides and $(n-i)$-nucleotides for $1 \le i \le n-1$. Figures 1–3 give examples of codes and their representing graphs in the case of $n=2$ (dinucleotide code), $n=3$ (trinucleotide code) and $n=4$ (tetranucleotide code).

As we can see, the graph of the tetranucleotide code has four disjoint parts. However, note that two parts are built by vertices labelled with dinucleotides and two parts are built by vertices labelled with nucleotides and trinucleotides. These parts are called *components* of $\mathcal{G}$. Recall that a subset V' of the set of vertices V is called *connected* if for any two nodes $v, w \in V'$ there is a path $[v, v_1][v_1, v_2] \dots [v_{n-1}, v_n][v_n, w]$ of vertices from V' connecting v and w. Any graph decomposes uniquely into connected components which are pairwise disjoint. Recall also that a graph is *bipartite* if its set of vertices V can be decomposed into two disjoint subsets V' and V'' such that the edges of $\mathcal{G}$ connect only nodes from V' with nodes from V'' and vice versa. Obviously, if X is an n-nucleotide code, then the components of $\mathcal{G}(X)$ are exactly the graphs

$$\mathcal{G}(X)_j = (V(X)_j, E(X)_j) \quad \textit{for } 1 \le j \le n-1$$

with

$$V(X)_j = \{N_1 \dots N_j, N_{j+1} \dots N_n, N_1 \dots N_{n-j}, N_{n-j+1} \dots N_n : N_1N_2N_3 \dots N_n \in X\}$$

and

$$E(X)_j = \{[N_1 \dots N_j, N_{j+1} \dots N_n], [N_1 \dots N_{n-j}, N_{n-j+1} \dots N_n] : N_1N_2N_3 \dots N_n \in X\}.$$

These components do not have to be connected, as we can see in figure 3. However, quite often they are. In fact, $\mathcal{G}(X)_j$ consists exactly of the nodes (and their corresponding edges) that interpret the elements of X in two ways: as a pair of a j-nucleotide and an $(n-j)$-nucleotide and as a pair of an $(n-j)$-nucleotide and a j-nucleotide. Note that by symmetry we have $\mathcal{G}(X)_j = \mathcal{G}(X)_{n-j}$ for all $j < n-1$. For instance, in figure 3 the two components of the graph associated with the tetranucleotide code are $\mathcal{G}(X)_1$ (=$\mathcal{G}(X)_3$) and $\mathcal{G}(X)_2$. The next observation is obvious.

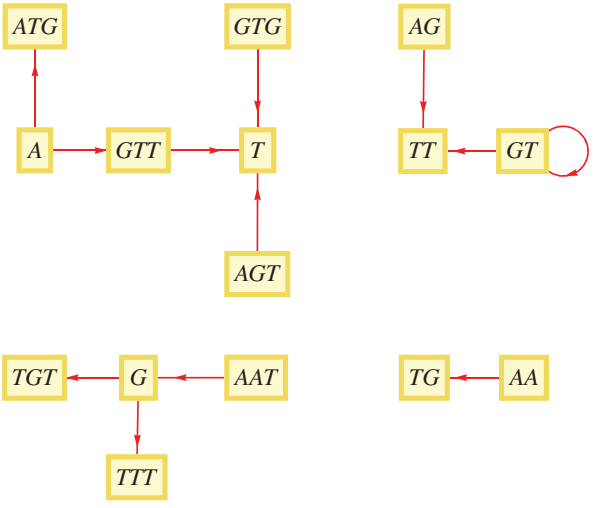

Figure 3. Graph representing the tetranucleotide code $\{AATG, AGTT, GTGT, GTTT\}$. (Online version in colour.)

Lemma 2.2. *Let X be an n-nucleotide code for some $n \in \mathbb{N}$. Then the following statements hold*:

(1) *If n is odd, then $\mathcal{G}(X)$ is a bipartite graph. In particular, all its components $\mathcal{G}(X)_j$ are bipartite.*
(2) *If n is even, then all components of $\mathcal{G}(X)$ are bipartite except for perhaps $\mathcal{G}(X)_{n/2}$.*

We now start to investigate our desired objects, namely n-nucleotide circular codes.

Definition 2.3. Let $X \subseteq \mathcal{B}^n$ be a code. We say that X is a *circular code* if for any concatenation $c_1 \dots c_m$ of n-nucleotide words from X there is only one partition into n-nucleotide words from X when read on a circle.

The first observation shows that for circular codes the associated graph is already simple. Recall from graph theory [28] that an oriented graph is *simple* if it does not contain loops, i.e. edges between a node and itself, and does not have multiple edges with the same orientation between two nodes. Note that the orientations for the multiple edges do play a role, i.e. for a simple oriented graph $\mathcal{G}$, we can still have $[x, y] \in E(\mathcal{G})$ and $[y, x] \in E(\mathcal{G})$, which means that there is a cycle (circle) of length 2. However, for circular codes this structure is also excluded. Recall that a *cycle* in $\mathcal{G}$ is an oriented closed path in $\mathcal{G}$. A *circle* is a cycle that visits no node twice except for the starting node (which is the end node at the same time). For instance, in figure 2, the sequence of vertices T, GT, A, TG, T is a circle while the sequence T, GT, A, GT, A, TG, T is a cycle that is not a circle.

Lemma 2.4. *Let $X \subseteq \mathcal{B}^n$ be a circular code. Then its representing graph is a simple oriented graph without circles of length* 2.

Proof. A proof can be found in appendix A. ■

Let us remark that lemma 2.4 simply says that for circular codes the representing graph has an *underlying* simple unoriented graph. Moreover, the circularity is only needed in a weaker form requiring definition 2.3 only for $m = 1$. These codes are called 1-*circular* (see, for instance, [29] for more details on these codes) and will appear again in §3.

Example 2.5. In figure 4, we show two examples of trinucleotide codes and their representing graphs. The code $\{ATG, CAC, CAT, GTG\}$ (a) is circular and has a simple graph, while the code $\{ATG, ATT, TGA, TGT\}$ (b) is non-circular and its graph is not simple.

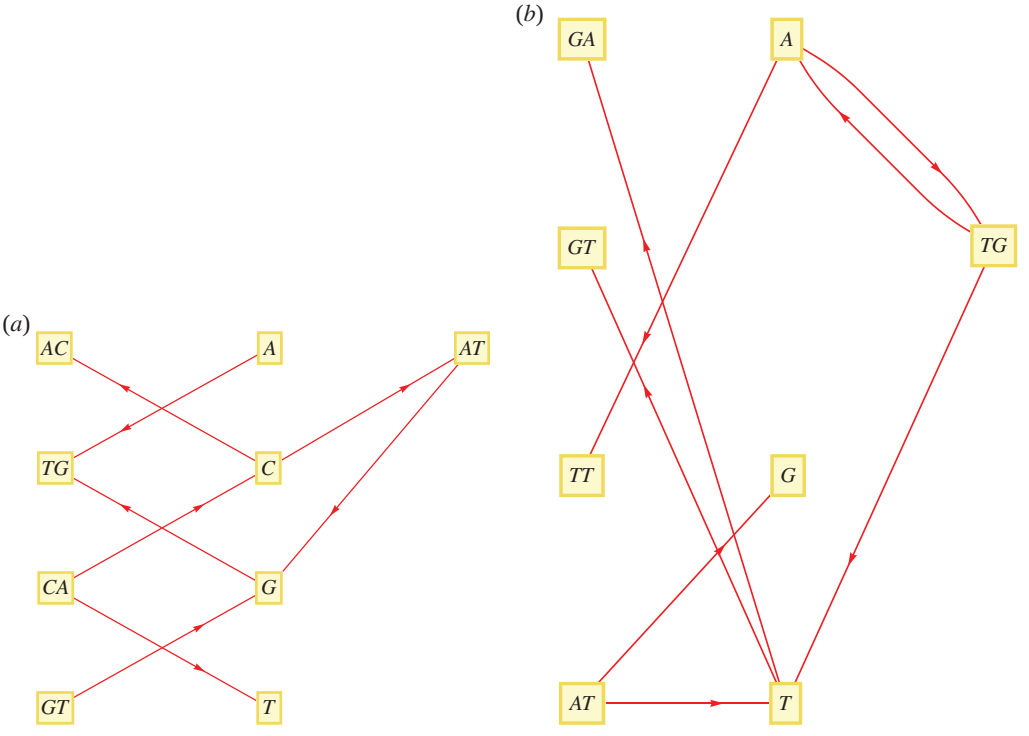

Figure 4. (*a*) The trinucleotide code {*ATG*, *CAC*, *CAT*, *GTG*} is circular and has a simple graph. (*b*) The trinucleotide code {*ATG*, *ATT*, *TGA*, *TGT*} is non-circular and its representing graph is not simple. (Online version in colour.)

We now state our first main theorem which proves the connection between the circularity of codes and the acyclicity of graphs. Recall from graph theory [28] that a graph is called *acyclic* if it does not contain cycles, i.e. oriented closed paths.

Theorem 2.6. *Given a code* $X \subseteq \mathcal{B}^n$ *the following statements are equivalent*:

(1) X *is circular.*
(2) $\mathcal{G}(X)$ *is acyclic.*

Proof. Let $X \subseteq \mathcal{B}^n$ be any code and assume that it is circular. If $G(X)$ is not acyclic, then one of its components $\mathcal{G}(X)_i$ is not acyclic. Hence there is a cycle in $\mathcal{G}(X)_i$ of the form

$$[N_1^1 \dots N_i^1, N_{i+1}^1 \dots N_n^1][N_{i+1}^1 \dots N_n^1, N_1^2 \dots N_i^2][N_1^2 \dots N_i^2, N_{i+1}^2 \dots N_n^2] \dots$$
$$[N_1^k \dots N_i^k, N_{i+1}^k \dots N_n^k]([N_{i+1}^k \dots N_n^k, N_1^1 \dots N_i^1])$$

for some natural number k. This means that the n-nucleotides

$$N_1^1 \dots N_n^1, \dots, N_1^k \dots N_n^k \in X$$

as well as the n-nucleotides

$$N_{i+1}^1 \dots N_n^1 N_1^2 \dots N_i^2, \dots, N_{i+1}^{k-1} \dots N_n^{k-1} N_1^k \dots N_i^k, N_{i+1}^k \dots N_n^k N_1^1 \dots N_i^1 \in X$$

are in the code. In the case $i \neq n - i$, the cycle has an even length. However, for even n, if $i = n - i$, i.e. $i = n/2$, it can happen that $N_{i+1}^k \dots N_n^k = N_1^1 \dots N_i^1$, and the last edge is missing. Hence, the cycle has odd length. In both cases taking every *second edge* from the cycle and starting with the edge

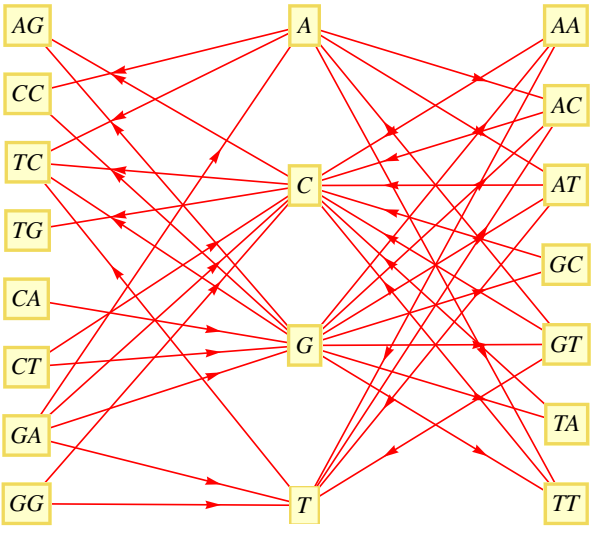

Figure 5. Graph $\mathcal{G}(X)$ of the maximal trinucleotide circular code X observed in genes of bacteria, eukaryotes, plasmids and viruses [12,15] (example 2.7). The four nucleotides $\{A, C, G, T\}$ of $\mathcal{G}(X)$ have ingoing and outgoing edges. The four dinucleotides $\{AG, CC, TC, TG\}$ of $\mathcal{G}(X)$ have no outgoing edge, the four dinucleotides $\{CA, CT, GA, GG\}$ of $\mathcal{G}(X)$ have no ingoing edge and the seven remaining dinucleotides $\{AA, AC, AT, GC, GT, TA, TT\}$ of $\mathcal{G}(X)$ have ingoing and outgoing edges. (Online version in colour.)

$[N_1^1 \dots N_i^1, N_{i+1}^1 \dots N_n^1]$ until we get the same edge repeated for the first time, we obtain the word

$$N_1^1 \dots N_i^1 N_{i+1}^1 \dots N_n^1 N_1^2 \dots N_n^2 N_1^3 \dots N_n^3 \dots N_1^k \dots N_i^k (N_{i+1}^k \dots N_n^k).$$

This word consists of k n-nucleotides $N_1^1 \dots N_n^1, \dots, N_1^k \dots, N_n^k$ if the cycle length is even and of $(2k-1)$ n-nucleotides $N_1^1 \dots N_n^1, \dots, N_1^k \dots N_n^k, N_{i+1}^1 \dots N_n^1 N_1^2 \dots N_i^2, \dots, N_{i+1}^{k-1} \dots N_n^{k-1} N_1^k \dots N_i^k$ if the cycle length is odd. Now, taking every *second edge* from the cycle but this time starting with the edge $[N_{i+1}^1 \dots N_n^1, N_1^2 \dots N_i^2]$ until we get the same edge repeated for the first time, we obtain a second decomposition of the word on the circle, namely

$$N_{i+1}^1 \dots N_n^1 N_1^2 \dots N_i^2, \dots, N_{i+1}^k \dots N_n^k (N_1^1 \dots N_i^1).$$

This is a contradiction to the circularity of X. The converse follows with similar arguments. ■

Clearly, the above result also gives a handy criterion for the C^3 *property* of trinucleotide codes, namely by the fact that a code is C^3 if and only if the graph of X as well as the graphs of the two *circularly permuted* sets of trinucleotides of X are acyclic. Recently, C^3 codes played an important role in the theory of error detection in genetic information. In particular, the maximal trinucleotide circular code X observed in genes of bacteria, eukaryotes, plasmids and viruses [12,15] initiated a renewed interest and had another interesting property, namely self-complementarity. Recall that an n-nucleotide code $X \subseteq \mathcal{B}^n$ is *self-complementary* if for each n-nucleotide from X the reversed complemented n-nucleotide is in X (see [30] for more details on C^3 codes). As an illustration we give the graph associated with the maximal self-complementary C^3 code found in [12,15].

Example 2.7. There are 12 964 440 maximal circular codes of 20 trinucleotides [12]. The maximal trinucleotide circular code X observed in genes of bacteria, eukaryotes, plasmids and viruses [12,15] has the following 20 trinucleotides:

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

The graph $\mathcal{G}(X)$ associated with X is shown in figure 5.

In order to see self-complementarity of a code X, we need to investigate first the reversing (mirroring) transformation which plays an important biological role. Recall that the *reversing transformation* inverts the order of bases in any n-nucleotide, i.e. for $x = N_1N_2 \dots N_{n-1}N_n \in \mathcal{B}^n$ we have $\overleftarrow{x} = N_nN_{n-1} \dots N_2N_1 \in \mathcal{B}^n$. If X is a code of n-nucleotides, then $\overleftarrow{X} = \{\overleftarrow{x} : x \in X\}$ is the *reversed code* of X. Similarly, the *complementing map* $c : \{A, C, G, T\} \to \{A, C, G, T\}$ that exchanges A and T as well as C and G induces the *complemented code* $c(X) = \{c(x) : x \in X\}$, where $c(N_1N_2 \dots N_{n-1}N_n) = c(N_1)c(N_2) \dots c(N_{n-1})c(N_n)$ for any n-nucleotide $x \in \mathcal{B}^n$. Note that for trinucleotides (codons) $x = N_1N_2N_3$ the anti-codon of x is exactly $\overleftarrow{c(x)}$.

The next lemma shows that the graphs $\mathcal{G}(X)$ and $\overleftarrow{\mathcal{G}(X)}$ of a code X and its reversed code $\overleftarrow{X}$ are *anti-isomorphic* while at the same time the graphs $\mathcal{G}(X)$ and $\mathcal{G}(c(X))$ of the complemented code $c(X)$ are isomorphic. Recall that an (*anti-*) *isomorphism* between two graphs $\mathcal{G}$ and $\mathcal{G}'$ is a bijective map $f : V \to V'$ that preserves edges in the sense that $[g_1, g_2] \in E$ if and only if $[f(g_1), f(g_2)] \in E'$ ($[g_1, g_2] \in E$ if and only if $[f(g_2), f(g_1)] \in E'$). For example, the graphs $G = (\{1, 2, 3\}, \{[1, 2], [2, 1], [1, 3]\})$ and $G' = (\{1, 2, 3\}, \{[1, 2], [2, 1], [3, 1]\})$ are anti-isomorphic. Their anti-isomorphism is easy to see considering the identical map $f = id$.

Lemma 2.8. *Let $X \subseteq \mathcal{B}^n$ be a code and $\mathcal{G}(X)$ its associated graph. Moreover, let c be the usual complementing map. Then*

(1) *The map $f_{\leftarrow} : V(X) \to \overleftarrow{V(X)}$ that sends a vertex $N_1 \dots N_j$ to $N_j \dots N_1$ is an anti-isomorphism between $\mathcal{G}(X)$ and $\overleftarrow{\mathcal{G}(X)}$.*
(2) *The map $f_c : V(X) \to c(V(X))$ that sends a vertex $N_1 \dots N_j$ to $c(N_1) \dots c(N_j)$ is an isomorphism between $\mathcal{G}(X)$ and $\mathcal{G}(c(X))$.*

Proof. The claims are obvious by the construction of the reversed code and the complemented code. ∎

We can now formulate self-complementarity in our graphs. The proof of the following result is obvious.

Theorem 2.9. *Let $X \subseteq \mathcal{B}^n$ be a code. Then X is self-complementary if and only if $\mathcal{G}(X)$ equals the reversed complemented graph $f_{\leftarrow}(f_c(\mathcal{G}(X)))$.*

The next theorem shows that we can even see the comma-freeness property of a code in its associated graph. Recall that an n-nucleotide code $X \subseteq \mathcal{B}^n$ is *comma-free* if for any two n-nucleotides $N_1 \dots N_n$ and $N'_1 \dots N'_n$ from X the n-nucleotides in frame 1 to $n-1$, i.e. $N_j \dots N_nN'_1 \dots N'_{j-1}$ for $2 \le j \le n$, do not belong to X. Comma-free codes are obviously circular but the converse is not true.

As an illustration, we give the graph associated with a maximal comma-free trinucleotide code.

Example 2.10. There are 408 maximal comma-free codes of 20 trinucleotides [2,3,31]. As an example, let Y be the following maximal comma-free code:

$$Y = \{AAC, AAG, AAT, CCA, GAC, TAC, GCA, GAG, TAG, TCA,$$
$$GAT, TAT, CCG, CCT, GCG, TCG, GCT, TCT, GGT, TTG\}.$$

The graph $\mathcal{G}(Y)$ associated with Y is shown in figure 6.

Our second main theorem shows that comma-freeness of a code X is connected to the maximal length of paths in its representing graph. In example 2.10, this length is 2, which is not a coincidence as we show now.

Theorem 2.11. *Given a code $X \subseteq \mathcal{B}^n$ the following statements are equivalent*:

(1) *The maximal length of a path in $\mathcal{G}(X)$ is 2.*
(2) *The code X is comma-free.*

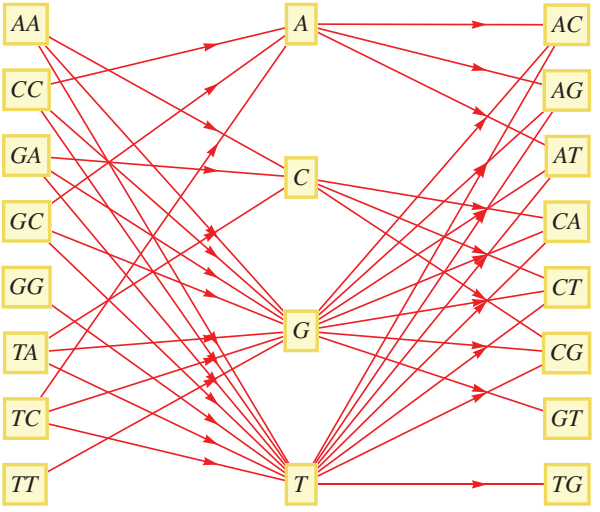

Figure 6. The graph $\mathcal{G}(Y)$ of the maximal trinucleotide comma-free code Y (example 2.10). The four nucleotides $\{A, C, G, T\}$ of $\mathcal{G}(Y)$ have ingoing and outgoing edges. The eight dinucleotides $\{AA, CC, GA, GC, GG, TA, TC, TT\}$ of $\mathcal{G}(Y)$ have no ingoing edges and the eight dinucleotides $\{AC, AG, AT, CA, CT, CG, GT, TG\}$ of $\mathcal{G}(Y)$ have no outgoing edges. (Online version in colour.)

Proof. A proof can be found in appendix B. ■

In general, the maximal length of a path in a representing graph $\mathcal{G}(X)$ of a code has a relation to the *error correcting window* of the code, i.e. the longest number of nucleotides that have to be read in an arbitrary sequence of words from the code X in order to retrieve the correct frame. In fact, any path in such a graph yields a sequence of nucleotides that can be read in two frames just by concatenating the labels of the vertices of the path. Conversely, any sequence (of words from the code) that can be read in two frames yields a path in the associated graph. The exact relation is not yet clear and has to be investigated in the future but we would like to present an example.

Example 2.12. For the circular code X from example 2.7, the longest paths in $\mathcal{G}(X)$ have 12 nucleotides if we start with a nucleotide. They are as follows: $[G, GT, A, AT, T, AC, C, AG]$, $[G, GT, A, AT, T, AC, C, TC]$ and $[G, GT, A, AT, T, AC, C, TG]$. Thus, the two longest ambiguous words of 11 nucleotides which can be read in at least two frames, namely frame 0 and frame 1, are: $GGTAATTACCA$ and $GGTAATTACCT$ where $GGT \in X$ is in frame 0.

If we start with a dinucleotide, then the longest paths in $\mathcal{G}(X)$ have 14 nucleotides and are given by: $[CA, G, GT, A, AT, T, AC, C, AG]$, $[CA, G, GT, A, AT, T, AC, C, TC]$, $[CA, G, GT, A, AT, T, AC, C, TG]$, $[CT, G, GT, A, AT, T, AC, C, AG]$, $[CT, G, GT, A, AT, T, AC, C, TC]$, $[CT, G, GT, A, AT, T, AC, C, TG]$, $[GA, G, GT, A, AT, T, AC, C, AG]$, $[GA, G, GT, A, AT, T, AC, C, TC]$ and $[GA, G, GT, A, AT, T, AC, C, TG]$.

Thus, the four longest ambiguous words of 12 nucleotides which can be read in at least two frames, namely frame 0 and frame 1, are: $AGGTAATTACCA$, $AGGTAATTACCT$, $TGGTAATTACCA$ and $TGGTAATTACCT$ as AG and TG are suffixes of trinucleotides from X.

3. Application for dinucleotide circular codes

In this section, we investigate our graph theoretic approach for the case $n = 2$, i.e. for dinucleotides, nucleotide words of length two. Given a dinucleotide code $X \subseteq \mathcal{B}^2$ the associated graph $\mathcal{G}(X)$ has at most four vertices, labelled by the nucleotide bases, and at most 12 directed edges. Each of the vertices can have at most four ingoing and four outgoing edges (see [27] for a classification of these codes).

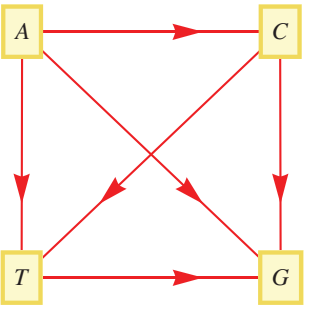

Figure 7. Graph representing the maximal dinucleotide circular code $X = \{AC, AG, AT, CG, CT, TG\}$. (Online version in colour.)

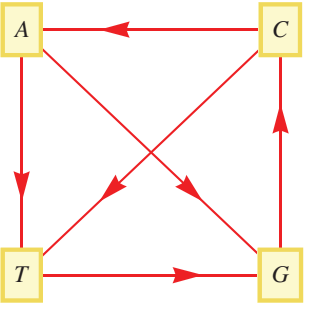

Figure 8. Graph representing the maximal dinucleotide 1-circular but not 2-circular code $X = \{AG, AT, CA, CT, GC, TG\}$. (Online version in colour.)

Example 3.1. Example of the maximal dinucleotide circular code $X = \{AC, AG, AT, CG, CT, TG\}$. The associated graph $\mathcal{G}(X)$ is shown in figure 7.

Recall that a dinucleotide code $X \subseteq \mathcal{B}^2$ is *k-circular* for $k \in \mathbb{N}$ if for any concatenation $c_1 \dots c_m$, $(m \leq k)$ of dinucleotides from X there is only one partition into dinucleotides from X when read on a circle [27]. Obviously, 1-circularity of a dinucleotide code X means that for each dinucleotide $N_1N_2 \in X$ the reversed dinucleotide N_2N_1 is not a member of X. This already implies that the associated graph $\mathcal{G}(X)$ of such a code can have at most six edges. Moreover, it is known [27] that for dinucleotide codes 3-circularity already implies circularity.

Example 3.2. Two examples of a 1-circular but not 2-circular dinucleotide code and a 2-circular but not 3-circular dinucleotide code.

(1) Let $X = \{AG, AT, CA, CT, GC, TG\}$, then X is 1-circular but not 2-circular and the associated graph $\mathcal{G}(X)$ is shown in figure 8.
(2) Let $X = \{AG, CA, CG, CT, GT, TA\}$, then X is 2-circular but not 3-circular and the associated graph $\mathcal{G}(X)$ is shown in figure 9.

Recall from graph theory [28] that a graph $\mathcal{G}$ is a *tournament* if it is obtained by assigning a direction to each edge of a complete (and hence simple) graph, i.e. it has $|V|(|V| - 1)/2$ edges. The following lemma shows that the tournaments on four vertices correspond exactly to the maximal 1-circular dinucleotide codes.

Lemma 3.3. *Given a code $X \subseteq \mathcal{B}^2$, the following statements are equivalent*:

(1) *X is a maximal 1-circular dinucleotide code.*

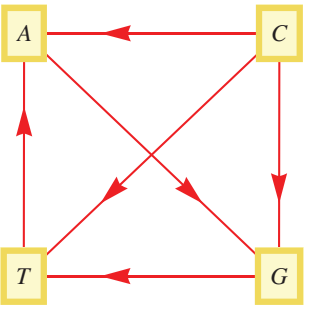

Figure 9. Graph representing the maximal dinucleotide 2-circular but not 3-circular code $X = \{AG, CA, CG, CT, GT, TA\}$. (Online version in colour.)

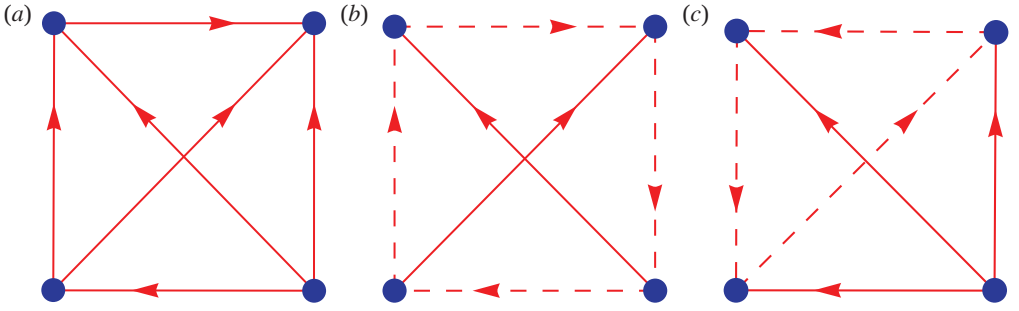

Figure 10. Graphs representing a maximal dinucleotide (*a*) circular, (*b*) 1- but not 2-circular and (*c*) 2-circular (but not circular) code. The Hamiltonian cycle from case (*b*) and the oriented cycle of length 3 from case (*c*) are highlighted as dashed lines. (Online version in colour.)

(2) *$\mathcal{G}(X)$ is a tournament on four vertices.*

Proof. A proof can be found in appendix C. ■

It follows immediately from lemma 3.3 that there are $2^6 = 64$ different maximal 1-circular dinucleotide codes since every tournament on four vertices has six edges and every edge can be oriented in two ways. We now characterize the graphs associated with 2- and 3-circular dinucleotide codes. Recall that a *Hamiltonian cycle* in some (oriented) graph is a (oriented) cycle that visits every node of the graph exactly once (except for the vertex that is both the start and end, which is visited twice). The proof of the following theorem can be found in appendix D.

Theorem 3.4. *Let $X \subseteq \mathcal{B}^2$ be a 1-circular dinucleotide code. Then*

(1) *X is circular if and only if $\mathcal{G}(X)$ is acyclic, i.e. $\mathcal{G}(X)$ does not contain any oriented cycle.*
(2) *X is a 1-circular but not 2-circular code if and only if $\mathcal{G}(X)$ contains a Hamiltonian cycle of length 4.*
(3) *X is a 2-circular but not 3-circular code if and only if $\mathcal{G}(X)$ contains an oriented cycle of length 3 and has no Hamiltonian cycle.*

Figure 10 visualizes the situations described in theorem 3.4.

As we have seen each maximal 1-circular dinucleotide code corresponds to a tournament on four vertices. The theory of tournaments is well studied in graph theory (see, for instance, [28]). Recall that the *score sequence* of a tournament is the set of out-degrees of its vertices $\{d^+(v) : v \in V\}$, where $d^+(v) = |\{[v, w] \in E : w \in V\}|$ for a vertex $v \in V$. Thus, we count how many edges start in

each vertex. For instance, in figure 10 the score sequences are (clockwise beginning from the upper left-hand vertex) (*a*) 1, 0, 3, 2; (*b*) 1, 1, 2, 2; (*c*) 1, 1, 3, 1.

Theorem 3.5. *The following statements are equivalent for a tournament $T = (V, E)$ on n vertices*:

(*T*-1) *T is acyclic.*
(*T*-2) *T does not contain a cycle of length 3.*
(*T*-3) *The score sequence of T is $\{0, 1, 2, \ldots, (n-1)\}$.*
(*T*-4) *T is transitive (i.e. from $[x, y] \in E$ and $[y, z] \in E$ it follows that $[x, z] \in E$).*
(*T*-5) *T has exactly one Hamiltonian path.*

We are now in a position to transfer theorem 3.5 one to one to circular dinucleotide codes showing a beautiful equivalence between the theory of tournaments on four vertices and the theory of maximal 1-circular dinucleotide codes. The equivalence of (*C*-1), (*C*-2) and (*C*-3) is known (see, for instance, [26,27]) but was proved using different techniques.

Theorem 3.6. *Let $X \subseteq \mathcal{B}^2$ be a maximal 1-circular dinucleotide code. Then the following statements are equivalent*:

(*C*-1) *X is circular.*
(*C*-2) *X is 3-circular.*
(*C*-3) *X has the form $X = \{N_1N_2, N_1N_3, N_1N_4, N_2N_3, N_2N_4, N_3N_4\}$, $N_i \in \mathcal{B}, N_i \neq N_j$.*
(*C*-4) *X is transitive in the following sense: from $N_1N_2, N_2N_3 \in X$ it follows that $N_1N_3 \in X$.*
(*C*-5) *The relation $<$ defined on $\mathcal{B} = \{A, C, G, T\}$ by*

$$N_1 < N_2 \quad \Leftrightarrow \quad N_1N_2 \in X, \quad N_i \in \mathcal{B}$$

is a total order.

Proof. A proof can be found in appendix E. ■

As an immediate corollary, we obtain the classifications of all maximal k-circular dinucleotide codes that were obtained in [26,27]. The idea of the proof which can be found in appendix F is that for a maximal circular dinucleotide code X the associated tournament $\mathcal{G}(X)$ is acyclic and hence completely determined by its unique Hamiltonian path. Hence there are 24 such codes since there are 24 different Hamiltonian paths possible. For 2-circular but not 3-circular codes, $\mathcal{G}(X)$ has a Hamiltonian cycle which determines four such codes. Since there are only six possible Hamiltonian cycles we get 24 such codes in total.

Corollary 3.7. *Let $X \subseteq \mathcal{B}^2$ be a maximal 1-circular code. The following statements are true*:

(*1*) *There are 24 different maximal dinucleotide circular (=3-circular) codes.*
(*2*) *There are 24 different maximal dinucleotide 1- but not 2-circular codes.*
(*3*) *There are 16 different maximal dinucleotide 2-circular but not circular codes.*

Figure 10 illustrates the graphs associated with maximal dinucleotide circular, 2-circular but not circular and 1- but not 2-circular codes. By labelling of the vertices with nucleotide bases A, C, G, T, all 24 maximal dinucleotide circular codes can be obtained from figure 10*a*, all 24 maximal dinucleotide 1- but not 2-circular codes from figure 10*b* and eight different 2-circular but not circular codes from figure 10*c* (some of the labels will lead to the same code). The other eight of the 2-circular but not circular codes can be obtained from figure 10*c* by reversing all edges (compare lemma 2.8 and corollary 3.7).

Finally, we consider maximal dinucleotide comma-free codes which have been classified recently in [32]. Also in this case, our new graph theoretical approach recovers the same result in a more elegant way. We begin with embedding circular dinucleotide codes into maximal circular dinucleotide codes. According to theorem 3.4, every dinucleotide circular code can be represented

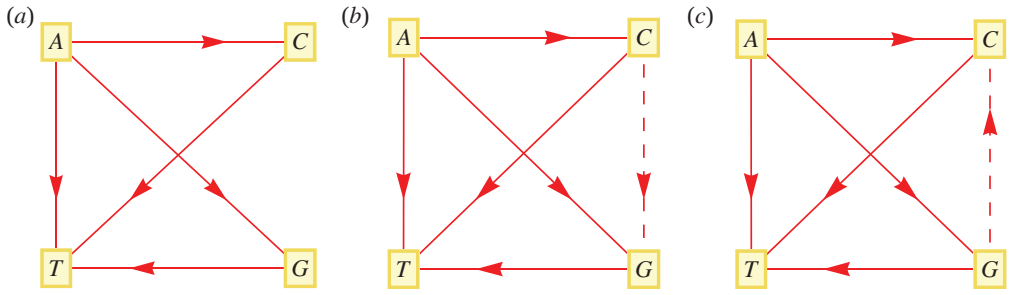

Figure 11. Graphs representing the maximal comma-free dinucleotide code (*a*) $\{AC, AG, AT, CT, GT\}$ and its two circular extensions (*b*) $\{AC, AG, AT, CG, CT, GT\}$ and (*c*) $\{AC, AG, AT, CT, GC, GT\}$. (Online version in colour.)

by an acyclic graph with at most four vertices. Straightforward calculations show that every such graph can be expanded to an acyclic tournament which represents a maximal dinucleotide circular code according to theorem 3.4. Hence we have the following.

Lemma 3.8. *Every dinucleotide circular code is contained in a maximal dinucleotide circular code. In particular, every comma-free dinucleotide code can be extended to a maximal circular dinucleotide code.*

Example 3.9. In figure 11, a maximal comma-free dinucleotide code and its two circular extensions are shown.

According to lemma 3.8, every comma-free code is contained in some maximal dinucleotide circular code. On the other hand, no maximal dinucleotide circular code is comma-free since its representing graph contains a Hamiltonian path of length 3 (compare theorem 2.11). If we remove one of the three edges which belongs to the (unique) Hamiltonian path $N_1 \to N_2 \to N_3 \to N_4$ from the corresponding acyclic tournament we obtain a graph which contains paths of at most length 2. According to theorem 2.11, we obtain in each case a graph that represents a dinucleotide comma-free code which will be maximal since it contains five edges (dinucleotides).

Example 3.10. In figure 12, a maximal dinucleotide circular code and its three maximal comma-free subcodes are shown.

Thus, we easily obtain the following.

Theorem 3.11 ([32]). *There are exactly* 36 *maximal dinucleotide comma-free codes. These are given as*

(1) *12 codes of the form* $\{N_1N_2, N_1N_3, N_1N_4, N_2N_4, N_3N_4\}$,
(2) *12 codes of the form* $\{N_1N_2, N_1N_3, N_1N_4, N_2N_3, N_2N_4\}$,
(3) *12 codes of the form* $\{N_2N_1, N_3N_1, N_4N_1, N_3N_2, N_4N_2\}$,

where $N_i \in \mathcal{B}, i = 1, 2, 3, 4$ *and* $N_i \neq N_j, i \neq j$.

Moreover, each of these codes can be obtained from a maximal dinucleotide circular code

$$X = \{N_1N_2, N_2N_3, N_3N_4, N_1N_4, N_1N_3, N_2N_4\}$$

with $N_i \in \mathcal{B}, N_i \neq N_j$ *by removing one of its dinucleotides* N_1N_2, N_2N_3 *or* N_3N_4.

4. Conclusion and perspectives

The circular code theory proposes that genes are constituted of two trinucleotide codes: the amino acid code and the circular code. The classical amino acid code contains 64 trinucleotides $\{AAA, \ldots, TTT\}$ with 61 trinucleotides coding the 20 amino acids and three stop codons which do not code for amino acid. The amino acid code in today's genes do not use all 64 available trinucleotides but a subset of 61 trinucleotides for coding the 20 amino acids. It

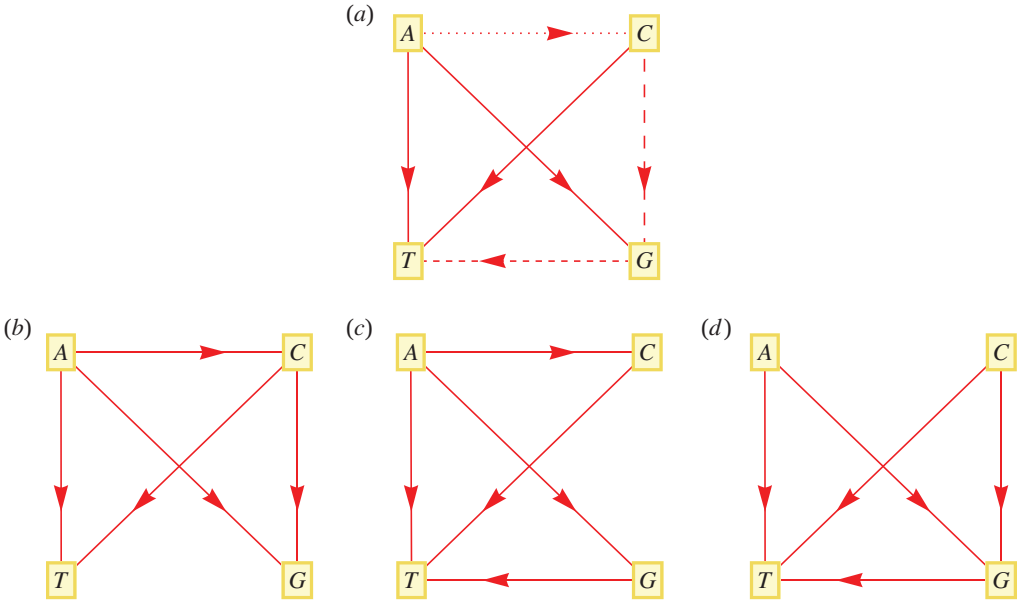

Figure 12. Graphs representing the maximal dinucleotide circular code (*a*) $\{AC, AG, AT, CG, CT, GT\}$ and its three maximal comma-free subcodes (*b*) $\{AC, AG, AT, CG, CT\}$, (*c*) $\{AC, AG, AT, CT, GT\}$ and (*d*) $\{AG, AT, CG, CT, GT\}$. (Online version in colour.)

is a surjective code. Furthermore, it contains two particular codes, a start code and a stop code, related to the reading frame of genes, precisely to initiate and close it. The main start trinucleotide code $\mathcal{T}_{\text{start}} = \{ATG\}$ is both a signal for the beginning of a gene and a code for the amino acid *Met*. The main stop trinucleotide code $\mathcal{T}_{\text{stop}} = \{TAA, TAG, TGA\}$ (also called stop codons) is only a signal for the end of a gene, i.e. without amino acid coding. The two codes $\mathcal{T}_{\text{start}}$ and $\mathcal{T}_{\text{stop}}$ have great variability among the variant genetic codes, showing an important evolution of the start and stop codes among species (see the genetic codes in http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=t). Indeed, in the standard code (code 1), the start trinucleotide code is extended to $\mathcal{T}_{\text{start}} = \{ATG, CTG, GTG, TTG\}$ coding the multi-set of amino acids $\{Met, Leu, Val, Leu\}$ (amino acid according to trinucleotide order) in eukaryotic genes. However, in the ciliate, dasycladacean and hexamita nuclear code (code 6), the euplotid nuclear code (code 10), the alternative flatworm mitochondrial code (code 14), the chlorophycean mitochondrial code (code 16) and the scenedesmus obliquus mitochondrial code (code 22), the start trinucleotide code is restricted to one trinucleotide $\mathcal{T}_{\text{start}} = \{ATG\}$ coding *Met*. By contrast, in the mould, protozoan and coelenterate mitochondrial code and the mycoplasma/spiroplasma code (code 4), the start trinucleotide code is extended up to eight trinucleotides $T_{\text{start}} = \{ATA, ATC, ATG, ATT, CTG, GTG, TTA, TTG\}$ coding the multi-set of amino acids $\{Ile, Ile, Met, Ile, Leu, Val, Leu, Leu\}$ (amino acid according to trinucleotide order). In the codes 6 and 14, the stop trinucleotide code is restricted to one trinucleotide $\mathcal{T}_{\text{stop}} = \{TGA\}$ and $\mathcal{T}_{\text{stop}} = \{TAG\}$, respectively. In the thraustochytrium mitochondrial code (code 23), the stop trinucleotide code is extended to four trinucleotides $\mathcal{T}_{\text{stop}} = \{TAA, TAG, TGA, TTA\}$.

The circular code X identified in genes of bacteria, eukaryotes, plasmids and viruses [12,15] is based on 20 trinucleotides with two mathematical properties involved in translation. It codes 12 amino acids $\boldsymbol{G}(X) = \{Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Phe, Thr, Tyr, Val\}$ according to the standard genetic code $\boldsymbol{G}$ [12, table 4(a)]. Thus, it is also a surjective code. Furthermore, it allows the reading frame to be retrieved, maintained and synchronized at any position in a gene generated by X. It is an extended mathematical property compared with the start and stop codes occurring only at the beginning and end positions of genes. The start code with a coding function and a frame function reduced to the first three nucleotides of a gene may be an evolutionary relic

of a circular code. Thus, the circular code X allows translation without the aid of proteins. As a consequence, we think that the circular code X has occurred before the classical amino acid code existing in today's genes. According to this hypothesis, primitive genes of short lengths would be directly decoded by the circular code X. This primitive coding of genes would be based on the property of reading frame retrieval of the circular code X and the coding of oligopeptides (peptides of short lengths) using the 12 amino acids $\boldsymbol{G}(X)$. It would not use the ribosome, which is a complex apparatus containing proteins, e.g. the 22 proteins S1–S22 in the small subunit and the 34 proteins L1–L36 in the large subunit of *Escherichia coli*, the two classes of aminoacyl tRNA synthetase, etc. Then, the complexity evolution of gene coding would have required an increase in the size of the protein alphabet (from 12 to 20), an increase in the length of proteins allowing diversity (the average size of today's genes is 1000 nucleotides for coding proteins of about 300 amino acids) and the development of the ribosomal apparatus as the property of reading frame retrieval of the circular code X would become inefficient with the size and topology of today's genes. The property of reading frame retrieval of the circular code has not completely disappeared during evolution as today's genes contain start and stop codes and as X motifs have been identified in tRNAs [13,17] and rRNAs, in particular in the ribosome decoding centre [13,18].

In this paper, we present a new approach to circular codes based on graph theory and extended to n-nucleotide codes. To each such code, a graph was assigned that interprets the n-nucleotides of the code as pairs of prefixes and suffixes. The general theorem established here identifies among the n-nucleotide codes those which are circular. Moreover, several properties of the circular codes can be seen in the representing graph, e.g. the error detecting window. Since dinucleotide circular codes may also have a biological function in the coding process of amino acids [26] we applied our approach to such codes and have established a beautiful correspondence between the theory of these codes and the theory of tournaments.

Tetranucleotide codes may also be involved in the amino acid coding [33–36] and thus our approach may help to investigate also the tetranucleotide circular codes in the future.

Authors' contributions. All authors contributed equally to this work.
Competing interests. We declare we have no competing interests.
Funding. We received no funding for this study.
Acknowledgements. E.F. and L.S. would like to thank the Karl Völker Foundation for its support.

Appendix A. Proof of lemma 2.4

Proof. A self-loop in $\mathcal{G}(X)$ can only arise for even n and would mean that $N_1 \dots N_{n/2} N_1 \dots N_{n/2} \in X$ for some $N_1, \dots, N_{n/2} \in \mathcal{B}$. The existence of multiple edges means that the same n-nucleotide is represented in the graph twice and the existence of inverted edges that $N_1 \dots N_i N_{i+1} \dots N_n \in X$ as well as $N_{i+1} \dots N_n N_1 \dots N_i \in X$. All this is forbidden by the construction of the graph and the circularity of the code X. ■

Appendix B. Proof of theorem 2.11

Proof. Let $X \subseteq \mathcal{B}^n$ be given and assume that X is comma-free. If $\mathcal{G}(X)$ contains a path of length at least 3, then it has one of length 3 which must belong to one of the components $\mathcal{G}(X)_j$ of $\mathcal{G}$. Thus, it is either of the form

$$[N_1 \dots N_j, N_{j+1} \dots N_n][N_{j+1} \dots N_n, N'_1 \dots N'_j][N'_1 \dots N'_j, N'_{j+1} \dots N'_n]$$

or

$$[N_1 \dots N_{n-j}, N_{n-j+1} \dots N_n][N_{n-j+1} \dots N_n, N'_1 \dots N'_{n-j}][N'_1 \dots N'_{n-j}, N''_1 \dots N''_{n-j+1}].$$

In the first case, the three n-nucleotides

$$N_1 \dots N_j N_{j+1} \dots N_n, \quad N_{j+1} \dots N_n N'_1 \dots N'_j \quad \text{and} \quad N'_1 \dots N'_j N'_{j+1} \dots N'_n$$

are in X. Thus, the concatenation $N_1 \dots N_j N_{j+1} \dots N_n N'_1 \dots N'_j N'_{j+1} \dots N'_n$ violates comma-freeness since $N_{j+1} \dots N_n N'_1$ is also in X. In the latter case, the three words

$$N_1 \dots N_{n-j} N_{n-j+1} \dots N_n, \quad N_{n-j+1} \dots N_n N'_1 \dots N'_{n-j} \quad \text{and} \quad N'_1 \dots N'_{n-j} N''_1 \dots N''_{n-j+1}$$

are in X. Thus, the concatenation $N_1 \dots N_{n-j} N_{n-j+1} \dots N_n N'_1 \dots N'_{n-j} N''_1 \dots N''_{n-j+1}$ violates comma-freeness since the n-nucleotide $N_{n-j+1} \dots N_n N'_1 \dots N'_{n-j}$ is also in X.

Conversely, assume that the maximal length of a path in $\mathcal{G}(X)$ is at most 2 and assume that X is not comma-free. Then there are two n-nucleotides $N_1 \dots N_n$ and $N'_1 \dots N'_n$ in X such that the concatenation $N_1 \dots N_n N'_1 \dots N'_n$ violates comma-freeness, i.e. there is $1 \leq j \leq n-1$ such that $N_{j+1} \dots N_n N'_1 \dots N'_j$ is in X as well. Thus, we obtain a path of length 3 in the jth components $\mathcal{G}(X)_j$ of $\mathcal{G}(X)$, namely the $[N_1 \dots N_j N_{j+1} \dots N_n][N_{j+1} \dots N_n N'_1 \dots N'_j][N'_1 \dots N'_j N'_{j+1} \dots N'_n]$ contradiction. ■

Appendix C. Proof of lemma 3.3

Proof. Let X be a maximal 1-circular code. Then X contains exactly six dinucleotides with all nucleotide bases appearing at least in one of its dinucleotides (see, for instance, [27]). Moreover, due to lemma 2.4 the associated graph $\mathcal{G}(X)$ is simple. Hence, $\mathcal{G}(X)$ is a complete graph on four vertices, which means that $\mathcal{G}(X)$ is a tournament on four vertices. The converse is obvious. ■

Appendix D. Proof of theorem 3.4

Proof. For claim (1) see theorem 2.6.

For claim (2) assume that X is 1- but not 2-circular. Then there are $N_1N_2, \tilde{N}_1\tilde{N}_2 \in X$ so that the word $N_1N_2\tilde{N}_1\tilde{N}_2$ has two decompositions on the circle. That means that also $N_2\tilde{N}_1, \tilde{N}_2N_1 \in X$. However, $N_1, N_2, \tilde{N}_1, \tilde{N}_2$ are different bases due to the 1-circularity of X and, thus, $N_1N_2\tilde{N}_1\tilde{N}_2N_1$ is a Hamiltonian cycle in $\mathcal{G}(X)$.

Let us assume now that $N_1 \to N_2 \to N_3 \to N_4 \to N_1, N_i \in \mathcal{B}$ is a Hamiltonian cycle in $\mathcal{G}(X)$. Then N_1N_2, N_2N_3, N_3N_4, $N_4N_1 \in X$ and the word $N_1N_2N_3N_4$ has two decompositions on the circle. Thus, X is not 2-circular.

Finally, for claim (3) let X be a 2- but not 3-circular code. Then according to (2) $\mathcal{G}(X)$ cannot have a Hamiltonian cycle. According to (1) $\mathcal{G}(X)$ cannot be acyclic since there are 2-circular codes which are not circular. The only case left is that $\mathcal{G}(X)$ has a cycle of length 3 but is not Hamiltonian.

Assume now that $\mathcal{G}(X)$ has a cycle of length 3 $N_1 \to N_2 \to N_3 \to N_1, N_i \in \mathcal{B}$. Then the word $N_1N_2N_3N_1N_2N_3$ has two decompositions on the circle since N_1N_2, N_2N_3, $N_3N_1 \in X$ and, thus, X is not 3-circular. In addition, X is 2-circular according to (2) since $\mathcal{G}(X)$ has no Hamiltonian cycle. ■

Appendix E. Proof of theorem 3.6

Proof. (C-1) $\Rightarrow$ (C-2): is obvious.

(C-2) $\Rightarrow$ (C-1): let X be a 3-circular code. According to theorem 3.4 (3) $\mathcal{G}(X)$ does not contain cycles of length 3 and, thus, according to theorem 3.5 $\mathcal{G}(X)$ is acyclic. It follows from theorem 2.6 (1) that X is circular.

(C-1) $\Leftrightarrow$ (C-3): clearly, if $X = \{N_1N_2, N_2N_3, N_3N_4, N_1N_4, N_1N_3, N_2N_4\}$, $N_i \in \mathcal{B}, N_i \neq N_j$ it follows that the score sequence of $\mathcal{G}(X)$ is $0, 1, 2, 3$. According to theorem 3.5, this is equivalent to the acyclicity of $\mathcal{G}(X)$ and, thus, to the circularity of X.

(C-1) $\Leftrightarrow$ (C-4): is obvious in view of the definition of $\mathcal{G}(X)$.

(C-1) $\Leftrightarrow$ (C-5): according to theorem 3.5, the acyclicity of $\mathcal{G}(X)$ and, thus, the circularity of X is equivalent to the existence of the unique Hamiltonian path in $\mathcal{G}(X)$. Given the unique Hamiltonian path $N_1 \to N_2 \to N_3 \to N_4$ in $\mathcal{G}(X)$ we define the order $N_1 < N_2 < N_3 < N_4$ which is a total order on $\mathcal{B}$ and vice versa. ■

Appendix F. Proof of corollary 3.7

Proof. For claim (1), the representing graph of a maximal dinucleotide circular code X is an acyclic tournament on four vertices. In such a tournament, there is exactly one Hamiltonian path [28] $N_1 \to N_2 \to N_3 \to N_4$. The remaining three edges can be oriented in the unique way to avoid cycles, namely $N_1 \to N_4$, $N_1 \to N_3$ and $N_2 \to N_4$. There are $24 = 4!$ possibilities to choose such a Hamiltonian path, which proves (1).

For claim (2), let X be a maximal dinucleotide 1- but not 2-circular code. Owing to theorem 3.4 the representing graph has a Hamiltonian cycle $N_1 \to N_2 \to N_3 \to N_4 \to N_1$, $N_i \in \mathcal{B}$. There are $6 = 3 \times 2$ possibilities to choose such a cycle and in each case $4 = 2 \times 2$ additional possibilities to orient the remaining two edges. It is easy to see that any orientation of the remaining two edges does not lead to new Hamiltonian cycles. Therefore, we have $24 = 6 \times 4$ different maximal dinucleotide 1- but not 2-circular codes.

For claim (3), the remaining codes which are not covered by (1) or (2) must be 2-circular but not circular. There are $16 = 64 - 24 - 24$ such codes. ■

References

1. Crick F, Griffith JS, Orgel LE. 1957 Codes without commas. *Proc. Natl Acad. Sci. USA* **43**, 416–421. (doi:10.1073/pnas.43.5.416)
2. Golomb SW, Gordon B, Welch LR. 1958 Comma-free codes. *Can. J. Math.* **10**, 202–209. (doi:10.4153/CJM-1958-023-9)
3. Golomb SW, Delbruck M, Welch LR. 1958 Construction and properties of comma-free codes. *Biologiske Meddelelser, Kongelige Danske Videnskabernes Selskab* **23**, 1–34.
4. Nirenberg MW, Matthaei JH. 1961 The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl Acad. Sci. USA* **47**, 1588–1602. (doi:10.1073/pnas.47.10.1588)
5. Crick FH, Brenner S, Klug A, Pieczenik G. 1976 A speculation on the origin of protein synthesis. *Orig. Life* **7**, 389–397. (doi:10.1007/BF00927934)
6. Eigen M, Schuster P. 1978 The hypercycle. A principle of natural self-organization. Part C: the realistic hypercycle. *Naturwissenschaften* **65**, 341–369. (doi:10.1007/BF00439699)
7. Shepherd JCW. 1981 Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl Acad. Sci. USA* **78**, 1596–1600. (doi:10.1073/pnas.78.3.1596)
8. Michel CJ. 1989 A study of the purine/pyrimidine codon occurrence with a reduced centered variable and an evaluation compared to the frequency statistic. *Math. Biosci.* **97**, 161–177. (doi:10.1016/0025-5564(89)90003-5)
9. Michel CJ. 1986 New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its evaluation. *J. Theor. Biol.* **120**, 223–236. (doi:10.1016/S0022-5193(86)80176-X)
10. Arquès DG, Michel CJ. 1987 Periodicities in introns. *Nucleic Acids Res.* **15**, 7581–7592. (doi:10.1093/nar/15.18.7581)
11. Konopka AK, Smythers GW. 1987 DISTAN—a program which detects significant distances between short oligonucleotides. *Bioinformatics* **3**, 193–201. (doi:10.1093/bioinformatics/3.3.193)
12. Arquès DG, Michel CJ. 1996 A complementary circular code in the protein coding genes. *J. Theor. Biol.* **182**, 45–58. (doi:10.1006/jtbi.1996.0142)
13. Michel CJ. 2012 Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. *Comput. Biol. Chem.* **37**, 24–37. (doi:10.1016/j.compbiolchem.2011.10.002)
14. Michel CJ. 2014 A genetic scale of reading frame coding. *J. Theor. Biol.* **355**, 83–94. (doi:10.1016/j.jtbi.2014.03.029)
15. Michel CJ. 2015 The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. *J. Theor. Biol.* **380**, 156–177. (doi:10.1016/j.jtbi.2015.04.009)
16. Michel CJ. 2008 A 2006 review of circular codes in genes. *Comp. Math. Appl.* **55**, 989–996. (doi:10.1016/j.camwa.2006.12.091)

17. Michel CJ. 2013 Circular code motifs in transfer RNAs. *Comput. Biol. Chem.* **45**, 17–29. (doi:10.1016/j.compbiolchem.2013.02.004)
18. El Soufi K, Michel CJ. 2014 Circular code motifs in the ribosome decoding center. *Comput. Biol. Chem.* **52**, 9–17. (doi:10.1016/j.compbiolchem.2014.08.001)
19. Burset M, Seledtsov IA, Solovyev VV. 2000 Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28**, 4364–4375. (doi:10.1093/nar/28.21.4364)
20. Mount SM. 1982 A catalogue of splice junction sequences. *Nucleic Acids Res.* **10**, 459–472. (doi:10.1093/nar/10.2.459)
21. Bird A. 2011 The dinucleotide CG as a genomic signalling module. *J. Mol. Biol.* **409**, 47–53. (doi:10.1016/j.jmb.2011.01.056)
22. Buerger H *et al.* 2004 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. *J. Pathol.* **203**, 545–550. (doi:10.1002/path.1542)
23. Gebhardt K, Zanker KS, Brandt B. 1999 Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J. Biol. Chem.* **274**, 13 176–13 180. (doi:10.1074/jbc.274.19.13176)
24. Schmidt AL, Mitter V. 2004 Microsatellite mutation directed by an external stimulus. *Mut. Res.* **568**, 233–243. (doi:10.1016/j.mrfmmm.2004.09.003)
25. Cuppens H *et al.* 1998 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. *J. Clin. Invest.* **101**, 487–496. (doi:10.1172/JCI639)
26. Michel CJ, Pirillo G. 2013 Dinucleotide circular codes. *ISRN Biomath.* **2013**, 538631. (doi:10.1155/2013/538631)
27. Fimmel E, Giannerini S, Gonzalez D, Strüngmann L. 2015 Dinucleotide circular codes and bijective transformations. *J. Theor. Biol.* **386**, 159–165. (doi:10.1016/j.jtbi.2015.08.034)
28. Clark J, Holton DA. 1991 *A first look at graph theory*. Singapore: World Scientific.
29. Fimmel E, Strüngmann L. 2015 On the hierarchy of trinucleotide *n*-circular codes and their corresponding amino acids. *J. Theor. Biol.* **364**, 113–120. (doi:10.1016/j.jtbi.2014.09.011)
30. Fimmel E, Giannerini S, Gonzalez D, Strüngmann L. 2014 Circular codes, symmetries and transformations. *J. Math. Biol.* **70**, 1623–1644. (doi:10.1007/s00285-014-0806-7)
31. Michel CJ, Pirillo G, Pirillo MA. 2008. Varieties of comma free codes. *Comput. Math. Appl.* **55**, 989–996. (doi:10.1016/j.camwa.2006.12.091)
32. Fimmel E, Strüngmann L. 2016 Maximal dinucleotide comma-free codes. *J. Theor. Biol.* **389**, 206–213. (doi:10.1016/j.jtbi.2015.10.022)
33. 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. Precedings*. (doi:10.1038/npre.2012.7136.1)
34. Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. 2003 Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* **13**, 145–158. (doi:10.1101/gr.335003)
35. Seligmann H. 2012 Putative mitochondrial polypeptides coded by expanded quadruplet codons, decoded by antisense tRNAs with unusual anticodons. *Biosystems* **110**, 84–106. (doi:10.1016/j.biosystems.2012.09.002)
36. Seligmann H, Labra A. 2013 Tetracoding increases with body temperature in Lepidosauria. *BioSystems* **114**, 155–163. (doi:10.1016/j.biosystems.2013.09.002)