BioSystems 239 (2024) 105215

Contents lists available at ScienceDirect

BioSystems

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

Circular code in introns

Christian J. Michel

Theoretical Bioinformatics, ICube, C.N.R.S., University of Strasbourg, 300 Boulevard Sébastien Brant, 67400 Illkirch, France

ARTICLE INFO

Keywords:
Circular code periodicity
Periodicities modulo 2, 3, 6
Trinucleotides $N_1N_2N_1$
Introns

ABSTRACT

A massive statistical analysis based on the autocorrelation function of the circular code X observed in genes is performed on the (eukaryotic) introns. Surprisingly, a circular code periodicity 0 modulo 3 is identified in 5 groups of introns: birds, ascomycetes, basidiomycetes, green algae and land plants. This circular code periodicity, which is a property of retrieving the reading frame in (protein coding) genes, may suggest that these introns have a coding property. In a well-known way, a periodicity 1 modulo 2 is observed in 6 groups of introns: amphibians, fishes, mammals, other animals, reptiles and apicomplexans. A mixed periodicity modulo 2 and 3 is found in the introns of insects. Astonishing, a subperiodicity 3 modulo 6 is a common statistical property in these 3 classes of introns. When the particular trinucleotides $N_1N_2N_1$ of the circular code X are not considered, the circular code periodicity 0 modulo 3, hidden by the periodicity 1 modulo 2, is now retrieved in 5 groups of introns: amphibians, fishes, other animals, reptiles and insects. Thus, 10 groups of introns, taxonomically different, out of 12 have a coding property related to the reading frame retrieval. The trinucleotides $N_1N_2N_1$ are analysed in the 216 maximal C^3 self-complementary trinucleotide circular codes. A hexanucleotide code (words of 6 letters) is proposed to explain the periodicity 3 modulo 6. It could be a trace of more general circular codes at the origin of the circular code X.

1. Introduction

Introns were discovered in 1977 (Berget et al., 1977; Chow et al., 1977), which showed that the eukaryotic genes, in contrast to the prokaryotic ones, contain insertion sequences that are removed from pre-mRNAs shortly after transcription. Such insertion sequences were named introns ("INTRagenic regiONs"), while protein coding gene fragments separated by them were named exons ("EXpressed regiONs") (Gilbert, 1978). From an evolution point of view, the introns-early and introns-late hypotheses have been debated for decades. The introns-early hypothesis states that the protein coding sequences had already contained introns at the primitive stages of evolution in order to generate new proteins, via recombination for example. The introns-late hypothesis states that the introns have emerged in eukaryotic genomes only.

The periodicity modulo 3 (also called three-base periodicity TBP) is defined as the preferential spacing of nucleotides and higher order k-tuples, such as trinucleotides, by distances of 3, 6, 9, etc. nucleotides. It is a well-known intrinsic property of (protein coding) genes observed in the pioneer works (Shepherd, 1981a,b; Fickett, 1982; Michel, 1986; Arquès and Michel, 1987a,b, 1990a,b). It is related to a biased distribution of codons, a consequence from the degeneracy of the genetic code (most amino acids are coded by more than one codon) and specific codon usage bias in different organisms. One proposal is that the ancestral forms of present-day genes might have been coded by the comma-free codes RRY (Crick et al., 1976) and RNY (Eigen and Schuster, 1978) ($R = \{A, G\}$, $Y = \{C, T\}$, N being any nucleotide). To illustrate the notion of periodicity modulo 3, for example with a RRY code, a sequence $RRY|RRY|RRY|\ldots$ implies that any nucleotide Y is distant from another nucleotide Y by a multiple of 3 nucleotides (3, 6, etc.), and any trinucleotide RRY is also distant from another trinucleotide RRY by a multiple of 3 nucleotides (0, 3, 6, etc.), etc. In 1986, it was shown that (eukaryotic) introns have no periodicity modulo 3 (Fig. 2 in Michel (1986)), and one year later a nucleotide periodicity modulo 2 is identified in the pioneer works (Arquès and Michel, 1987c; Konopka and Smythers, 1987).

In real genetic sequences, the periodic signals modulo 2 and 3 are very noisy. Thus, sensitive statistical-signal analysis functions are necessary to study them. In this paper, we apply the circular code autocorrelation function to the introns. It is based on the maximal C^3 self-complementary trinucleotide circular code X that is in average overrepresented in the reading frame of genes of bacteria, archaea, eukaryotes, plasmids and viruses (Arquès and Michel, 1996; Michel, 2015, 2017) where

$$X = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, \\ GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}. \tag{1.1}$$

E-mail address: c.michel@unistra.fr.
URL: https://dpt-info.di.unistra.fr/~c.michel/.

https://doi.org/10.1016/j.biosystems.2024.105215
Received 12 January 2024; Received in revised form 14 April 2024; Accepted 15 April 2024
Available online 17 April 2024

A code is circular if any word written on a circle (the last nucleotide becoming the first in the circle) has a unique decomposition into codewords. A trinucleotide circular code allows to retrieve and maintain the reading frame (property of synchronisation) by using an appropriate window of nucleotides. In any sequence generated by a trinucleotide circular code, the reading frame can be retrieved in a window length of at most 13 consecutive nucleotides (4 trinucleotides plus one nucleotide). The theoretical concepts of circular code are not necessary to understand the methods and results obtained in this work. However, the reader wishing to go into more detail can refer to the two reviews (Michel, 2008; Fimmel and Strüngmann, 2018) and to the recent theoretical works (Fimmel et al., 2019, 2020; Michel et al., 2022; Michel and Sereni, 2022, 2023; Fimmel et al., 2023a,b). The historical context of the circular code result is described in Sections 4 and 5 in Michel (2020).

This article is organised as follows. The method Section 2 is divided into four parts. Section 2.1 gives the definition of the autocorrelation function of the circular code X(1.1) observed in genes that will be applied to the study of introns. Section 2.2 recalls the circular code periodicity 0 modulo 3 in genes. Section 2.3 presents a basic analysis to evaluate the statistical significance of a circular code periodicity. Section 2.4 describes the acquisition of introns from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, May 2023). The results Section 3 are presented in six parts. A circular code periodicity modulo 3 is identified in 5 subgroups of introns, like that observed in genes, a result that has never been reported to our knowledge (Section 3.1). A classical periodicity modulo 2 is retrieved in 6 subgroups of introns (Section 3.2). A mixed periodicity modulo 2 and 3 is found in the introns of insects and a striking subperiodicity modulo 6 can exist in these 3 classes of introns (Section 3.3). A theoretical study shows that the particular trinucleotides $N_1N_2N_1$ of a circular code, i.e. CTC and GAG in the circular code X(1.1), can be associated with a periodicity 1 modulo 2 (Section 3.4). When the particular trinucleotides $N_1N_2N_1$ of the circular code X(1.1) are not considered, the circular code periodicity modulo 3, hidden by the periodicity modulo 2, is now retrieved in 5 groups of introns: amphibians, fishes, other animals, reptiles and insects (Section 3.5). Section 3.6 shows that the circular code periodicity 0 modulo 3 also is also observed with the two permuted circular codes of X(1.1). Section 3.7 presents a few new properties and future research ideas on circular codes. The trinucleotides $N_1N_2N_1$ are studied in the 216 maximal C^3 self-complementary trinucleotide circular codes (Section 3.7.1). Hexanucleotide codes are proposed for modelling the periodicity modulo 6 (Section 3.7.2).

2. Method

2.1. Circular code autocorrelation function

We recall here the autocorrelation function applied to the circular code (Michel and Thompson, 2020). This approach that gives exact probabilities, to the nearest numerical approximations, is particularly adapted to identify periodicities in noisy sequences, such as the introns.

An intron family F consists of $|F|$ introns (genetic sequences) on the 4-nucleotide alphabet $\mathcal{B} := \{A, C, G, T\}$ where A stands for adenine, C stands for cytosine, G stands for guanine, and T stands for thymine. Let the sequence $s = N_1N_2 \cdots N_{|s|}$ be an intron of F of length of $|s|$ nucleotides, $N_i \in \mathcal{B}$ for $i \in \{1, \dots, |s|\}$. Let m and m' be 2 motifs of respective lengths $|m|$ and $|m'|$ on $\mathcal{B}$. Then, the correlation function $A_{m,m'}(i, s)$ in a sequence s is defined by

$$A_{m,m'}(i,s) = \frac{1}{l(s)} \sum_{p=1}^{l(s)} \delta_m(p) \cdot \delta_{m'}(p + |m| + i), \quad i = 0, \dots, n, \quad n \ll l(s), \tag{2.1}$$

with

$$\delta_m(p) = \begin{cases} 1 & \text{if } s[p..p + |m| - 1] = m \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

and $l(s) = |s| - (|m \cdot m'| + n) + 1$ with the length $|m \cdot m'| = |m| + |m'|$.

Table 1
Autocorrelation function $A_{R,R}(i, (RNY)^+)$.

i	R	N	Y	R	N	Y	Probability	Total
0	R	R					$1/3 \times 1 \times 1/2$	
		R	R				$1/3 \times 1/2 \times 0$	
			R	R			$1/3 \times 0 \times 1$	1/6
1	R		R				$1/3 \times 1 \times 0$	
		R		R			$1/3 \times 1/2 \times 1$	
			R		R		$1/3 \times 0 \times 1/2$	1/6
2	R			R			$1/3 \times 1 \times 1$	
		R			R		$1/3 \times 1/2 \times 1/2$	
			R			R	$1/3 \times 0 \times 0$	5/12

Remark 2.1. $s[p..p + |m| - 1] = m$ means that an occurrence of m starts at the left position p on s, or briefly that the motif m in s is (occurs) in position $p..p + |m| - 1$.

Remark 2.2. When $i = 0$, the motif m in position $p..p + |m| - 1$ and the motif m' in position $p + |m| + i..p + |m| + i + |m'| - 1$, i.e. $p + |m|..p + |m \cdot m'| - 1$, are consecutive.

This definition of $A_{m,m'}(i, s)$ can also be understood as follows. Let an i-motif mN^im' $(m \cdot N^i \cdot m')$ be 2 motifs m and m' separated by i, $i \in \{0, \dots, n\}$, any nucleotides $N \in \mathcal{B}$. Thus according to this convention, the 0-motif is mm' (m and m' are consecutive), the 1-motif is mNm' (m and m' are separated by any 1 nucleotide N), the 2-motif is $mNNm'$ (m and m' are separated by any 2 nucleotides N), etc. In order to count the occurrences of mN^im' in a sequence s of F under the same conditions for all $i \in \{0, \dots, n\}$, i.e. without probability bias, only the $l(s)$ first nucleotides of s are analysed (a few i-motifs at the end of s are thus not considered, since $l(s)$ is a function of the constant n and not of the variable i). Indeed, when $p = l(s)$ and $i = n$, then the motif m' in position $p + |m| + i = |s| - |m'| + 1$ has its last nucleotide $l_{|m'|}$ in the last position of s.

Remark 2.3. The definition of $A_{m,m'}(i, s)$ is a generalisation of the classical letter correlation function used in signal analysis when the motifs m and m' are letters N and N', respectively.

Remark 2.4. As a consequence of Eq. (2.1), the correlation function $A_{m,m'}(i, s)$ gives exact probabilities, to the nearest numerical approximations, that can be retrieved mathematically when the sequence s has a basic structure or a combination of basic structures, e.g. N^j, $(N_1N_2)^j$, $(N_1N_2)^j(N_1N_2N_3)^k$, etc., where j and k are positive integers (detailed in Appendix A in Michel and Thompson (2020)). However, only the function $A_{m,m'}(i, s)$ can compute real and noisy genetic sequences s.

For the convenience of the reader, we give a computation example of the correlation function $A_{m,m'}(i, s)$ (Eq. (2.1)).

Example 2.5. For sake of simplicity, we compute $A_{m,m'}(i, s)$ (Eq. (2.1)) over the 2-letter alphabet $\mathcal{B}' := \{R, Y\}$ with $N = \{R, Y\}$ (with equiprobability) and the one-letter motifs m and m' such that $m = m' = R$ (autocorrelation) on the sequence $(RNY)^+ = RNYRNY...$ As the sequence is a concatenation of a same trinucleotide, three probability cases must be considered (Table 1).

Thus, the autocorrelation function $A_{R,R}(i, (RNY)^+)$ is

$$A_{R,R}(i, (RNY)^+) = \begin{cases} \frac{1}{6} & \text{for } i \equiv 0 \pmod 3 \\ \frac{1}{6} & \text{for } i \equiv 1 \pmod 3 \\ \frac{5}{12} & \text{for } i \equiv 2 \pmod 3\,. \end{cases}$$

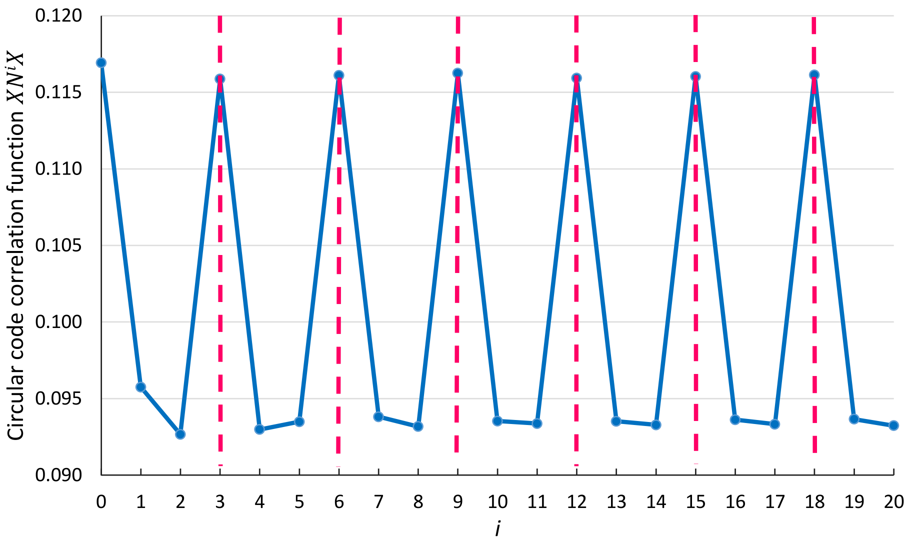


Fig. 1. (Fig. 2 in Michel and Thompson (2020)). Circular code periodicity 0 modulo 3 identified by the circular code autocorrelation function $A_{X,X}(i,F)$ (Eq. (2.4)) in bacterial (protein coding) genes (465,762 genes, 2,339,752,707 trinucleotides). The abscissa represents the number i of nucleotides N between X and itself, i varying from 0 to $n = 20$. The ordinate gives the occurrence probability $A_{X,X}(i,F)$ of XN^iX in bacterial genes.

In order to study the correlation function of the circular code X(1.1) based on 20 trinucleotides, we choose $|m| = |m'| = 3$ and extend Eq. (2.1) to a set of motifs. Let $\mathcal{B}^3 = \{AAA,\dots,TTT\}$ be the set of the 64 trinucleotides with the following partition into 2 classes $\{X,\overline{X} : X \cap \overline{X} = \emptyset, X \cup \overline{X} = \mathcal{B}^3\}$ with X defined in (1.1). Then, the circular code autocorrelation function $A_{X,X}(i,s)$ in a sequence s is defined by

$$A_{X,X}(i,s) = \sum_{m\in X}\sum_{m'\in X} A_{m,m'}(i,s), \quad i = 0,\dots,n, \tag{2.3}$$

with $A_{m,m'}(i,s)$ defined in Eq. (2.1). Eq. (2.3) easily extends to a sequence family. Thus, the circular code autocorrelation function $A_{m,m'}(i,F)$ in a family F is defined by

$$A_{X,X}(i,F) = \frac{1}{|F|}\sum_{s\in F} A_{X,X}(i,s), \quad i = 0,\dots,n, \tag{2.4}$$

with $A_{X,X}(i,s)$ defined in Eq. (2.3).

The function $i \longrightarrow A_{X,X}(i,F)$ giving the occurrence probability that the circular code X appears any i nucleotides N after X in the family F, is called the circular code autocorrelation function XN^iX (associated with the i-motif XN^iX based on the circular code X). It is represented by a curve with:

- on the abscissa, the number i of nucleotides N between X and itself, i varying from 0 to n, which is chosen to be $n = 20$ in the described results.

- on the ordinate, the occurrence probability $A_{X,X}(i,F)$ of XN^iX in F.

Remark 2.6. $\sum_{S\in\{X,\overline{X}\}}\sum_{S'\in\{X,\overline{X}\}} A_{S,S'}(i,F) = A_{X,X}(i,F)+A_{X,\overline{X}}(i,F)+A_{\overline{X},X}(i,F)+A_{\overline{X},\overline{X}}(i,F) = A_{\mathcal{B}^3,\mathcal{B}^3}(i,F) = 1$ for all i, $i \in \{0,\dots,n\}$, in F. The curve $A_{\mathcal{B}^3,\mathcal{B}^3}(i,F)$ is a horizontal line of value 1.

Remark 2.7. $A_{X,X}(i,F) = \frac{20\cdot 20}{64\cdot 64} = \frac{25}{256} \approx 0.0977$ for all i, $i \in \{0,\dots,n\}$, in a random family F, and in particular in a random sequence s. Note that s is a particular case where $|F| = 1$. Thus, in the random case, the curve $A_{X,X}(i,F)$ is a horizontal line of value 0.0977.

Remark 2.7 is particularly interesting as any correlation curve without horizontal line of value 0.0977 can be associated with a non-random family F or a non-random sequence s.

2.2. *Circular code periodicity 0 modulo 3 in protein coding genes*

The observation of a periodicity modulo 3 in (protein coding) genes of eukaryotes, bacteria, viruses, chloroplasts and mitochondria is classical and has been described in the past by several authors and by different statistical and signal methods, in particular at the sequence level by Shepherd (Shepherd, 1981a,b) and at the population level by Fickett (1982), Michel (Fig. 1 in Michel (1986)), and Arquès and Michel (Arquès and Michel, 1987a,b, 1990a,b). This periodicity modulo 3 in genes has mainly been assigned to the properties of the codon length of 3 nucleotides and the degeneracy of the genetic code.

In this line of research, a circular code periodicity 0 modulo 3 has recently been identified by applying the circular code autocorrelation function (Eq. (2.4)) to bacterial genes (465,762 genes, 2,339,752,707 trinucleotides). Fig. 1, which is a reminder of Fig. 2 in Michel and Thompson (2020), has higher values for multiples $(3i)$ than for multiples of $(3i+1)$ or $(3i+2)$ where $i \in \{0,1,\dots,\lceil\frac{n}{3}\rceil\}$. Note that the values in Fig. 1 are around 0.1, as expected by Remark 2.7. Thus, a circular code periodicity modulo 3 in genes has been assigned to the frame retrieval of genes.

2.3. *A basic analysis to evaluate the statistical significance of a circular code periodicity*

The probability that $A_{X,X}(0,F) > A_{X,X}(1,F)$ is equal to $\frac{1}{2}$. The probability that $A_{X,X}(i,F) > A_{X,X}(i-1,F)$ and $A_{X,X}(i,F) > A_{X,X}(i+1,F)$ with $i \equiv 0 \pmod 3$ and $i > 0$ is equal to $\frac{1}{3}$. By assuming independence between the events, the probability of a periodicity 0 modulo 3 until $i = n$ is equal to $p = \frac{1}{2}\cdot\left(\frac{1}{3}\right)^{\lfloor\frac{n-1}{3}\rfloor}$. For example, when $n = 20$ then $p = \frac{1}{2}\cdot\left(\frac{1}{3}\right)^{6} \approx 7\times 10^{-4}$, a strong statistical significance observed in Fig. 1.

2.4. *Data*

Eukaryotic introns groups were obtained from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, May 2023) Introns with nucleotides different from $\mathcal{B}$ as well as less than 500 nucleotides are excluded from this statistical analysis. Groups of introns with less

Table 2
Groups and subgroups F of introns extracted from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, May 2023) with their symbol and their total numbers of chromosomes, introns and nucleotides.

Group	Subgroup	Total number		
		Chromosomes	Introns	Nucleotides
Animals **A**	Amphibians **AA**	116	1,739,367	3,483,784,343
	Birds **AB**	1,272	7,239,100	11,573,631,130
	Fishes **AF**	2,986	21,559,237	35,662,962,674
	Insects **AI**	1,453	4,553,961	7,284,977,520
	Mammals **AM**	2,056	22,532,010	42,974,235,162
	Other Animals **AOA**	526	3,123,784	4,844,433,483
	Reptiles **AR**	380	4,595,378	8,321,053,066
Fungi **F**	Ascomycetes **FA**	846	12,595	13,751,462
	Basidiomycetes **FB**	238	3,631	4,142,508
Plants **Pl**	Green Algae **PGA**	133	1,319	1,982,139
	Land Plants **PLP**	2,147	1,817,549	3,349,278,680
Protists **Pr**	Apicomplexans **PA**	263	40,880	28,972,306

than 100 chromosomes (identifier NC) are excluded. Table 2 gives some basic information about the studied groups of introns.

3. Results

A statistical analysis of introns is carried out on a very large scale, including 4 groups: animals **A**, fungi **F**, plants **Pl** and protists **Pr** according to the taxonomy of the GenBank database (Table 2). The animals group analysed is divided into 7 subgroups: amphibians **AA**, birds **AB**, fishes **AF**, insects **AI**, mammals **AM**, other animals **AOA** and reptiles **AR**. The fungi group has 2 subgroups: ascomycetes **FA** and basidiomycetes **FB**. The plants group contains 2 subgroups: green algae **PGA** and land plants **PLP**. The protists groups is represented by the subgroup apicomplexans **PA**.

By applying the circular code autocorrelation function $A_{X,X}(i,F)$ (Eq. (2.4)), these 12 subgroups of introns can be classified into 3 classes according to their identified periodicities.

Remark 3.1. For simplicity of writing in the following, a periodicity 0 modulo 3 is briefly named by modulo 3, and a periodicity 1 modulo 2, by modulo 2.

3.1. *Circular code periodicity 0 modulo 3 in introns*

Very surprisingly, 5 subgroups of introns have a circular code periodicity 0 modulo 3, as in the (protein coding) genes (Fig. 2): birds **AB**, ascomycetes **FA**, basidiomycetes **FB**, green algae **PGA** and land plants **PLP**. Therefore, the (complete) groups of fungi and plants have a circular code periodicity. In a strange way, 1 subgroup of animals, the birds, has also a circular code periodicity. Note that the value $A_{X,X}(0,F)$ at $i=0$ is very low, compared to the remaining values, in $F=\mathbf{AB}$ (flattened periodicity in Fig. 2(A)) and $F=\mathbf{PLP}$. The circular code periodicity in introns is less regular than in genes (compare Figs. 1 and 2).

3.2. *Periodicity 1 modulo 2 in introns*

As is the norm (Michel, 1986; Arquès and Michel, 1987c; Konopka and Smythers, 1987), 6 subgroups of introns have a periodicity 1 modulo 2 (Fig. 3): amphibians **AA**, fishes **AF**, mammals **AM**, other animals **AOA**, reptiles **AR** and apicomplexans **PA**. The group of animals, except the birds with a periodicity modulo 3 (see Section 3.1) and the insects with a mixed periodicity (see below Section 3.3), has a periodicity modulo 2. Note that the value $A_{X,X}(0,F)$ at $i=0$ is very low, compared to the remaining values, in almost all subgroups, except for $F=\mathbf{AOA}$ and $F=\mathbf{PA}$. The periodicity modulo 2 is short in $F=\mathbf{AA}$ and $F=\mathbf{AR}$ as it disappears for $i \geq 11$.

Homo sapiens (24 chromosomes, 593,398 introns, 1,182,410,148 nucleotides) has a short periodicity 1 modulo 2 (Fig. 4) as it disappears for $i \geq 7$.

3.3. *Mixed periodicity in introns*

The case of introns for the animals insects **AI** is strange and interesting. Indeed, Fig. 5 shows that the periodicity 0 modulo 3 is incomplete as there are no peaks at $i = 6, 12, 18$. Analogously, the periodicity 1 modulo 2 is also incomplete as there are no peaks at $i = 1, 11, 19$. This observation may suggest that the introns of **AI** have a mixed periodicity modulo 2 and 3. Furthermore, a subperiodicity 3 modulo 6 can be observed: peaks at $i = 3, 9, 15$ in Fig. 5. Very surprisingly, by re-examining the previous figures, a subperiodicity 3 modulo 6 also exists in some introns with a complete circular code periodicity 0 modulo 3, precisely in fungi ascomycetes **FA** (peak at $i = 15$ not significant), plants green algae **PGA** and plants land plants **PLP** (Fig. 2(B)(D)(E), respectively). Interestingly, the introns of fungi ascomycetes **FA** and plants green algae **PGA** (Fig. 2(B)(D), respectively) have an additional subperiodicity 0 modulo 6 with peaks at $i = 0, 6, 12, 18$. It is also astonishing to identify this subperiodicity 3 modulo 6 in the introns with a complete periodicity 1 modulo 2 of animals fishes **AF** (peak at $i = 15$ not significant), mammals **AM**, other animals **AOA** and reptiles **AR** (peak at $i = 15$ not significant), and of protists apicomplexans **PA** (peak at $i = 15$ not significant) (Fig. 3(B)(C)(D)(E)(F)). Note that the subperiodicity 0 modulo 6 obviously cannot be present in introns with a complete periodicity 1 modulo 2.

3.4. *Trinucleotide circular code associated with a periodicity modulo 2*

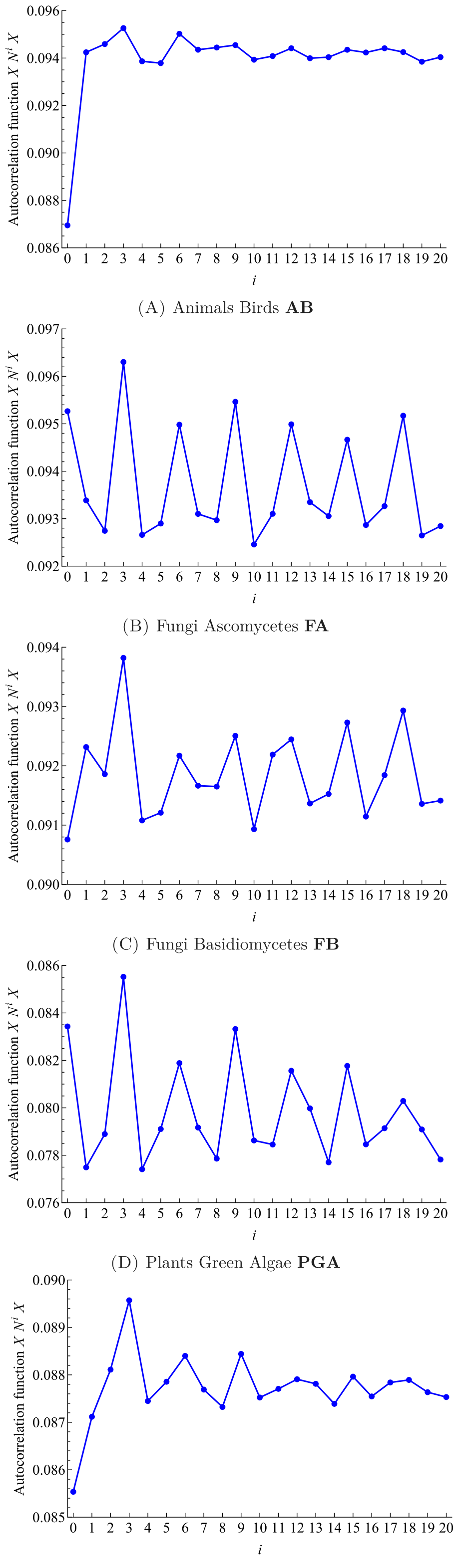
The observation that a trinucleotide circular code, such that X(1.1), can be associated with a periodicity 1 modulo 2, is not obvious and needs some investigation. A periodicity modulo 2 is generated by the sequence $(N_1N_2)^+ = N_1N_2N_1N_2 \ldots$, with $N_1, N_2 \in \mathcal{B}$ and $N_1 \neq N_2$. The code $\{N_1N_2\}$ is circular.

Remark 3.2. If $N_1 = N_2$ then the sequence $(N_1N_1)^+ = N_1^+$ cannot generated a periodicity modulo 2.

The study of trinucleotides involving in a periodicity modulo 2 can be carried out for a code in a general context. It is considered here for a code that is circular.

A first hypothesis is based on the fact that a sequence $(N_1N_2)^+$ can be constructed by a series of 2 trinucleotides: $(N_1N_2)^+ = (N_1N_2N_1 \cdot N_2N_1N_2)^+$. However, a circular code cannot contain at the same time the 2 trinucleotides $N_1N_2N_1$ and $N_2N_1N_2$ as some sequences have 2 (even 3) decompositions on a circle: $N_1N_2N_1 \cdot N_2N_1N_2$ and $N_1 \cdot N_2N_1N_2 \cdot N_1N_2$. Thus, this hypothesis must be rejected, noting that such a complex process of constructing periodicities modulo 2 also seems hardly appropriate for a primitive stage of life.

A second hypothesis is to observe that a sequence $(N_1N_2)^+$ contains an unique trinucleotide $N_1N_2N_3$ that overlaps at the 3rd codon site: $\overline{N_1N_2}\overline{\underline{N_1}}\underline{N_2}\overline{\underline{N_1}}\overline{N_2}\overline{\underline{N_1}}\underline{N_2N_1} \cdots$. Such "modulo 2" trinucleotides allowing a reading by 2 nucleotides, only exist if $N_1 = N_3$. The circular



(A) Animals Birds **AB**

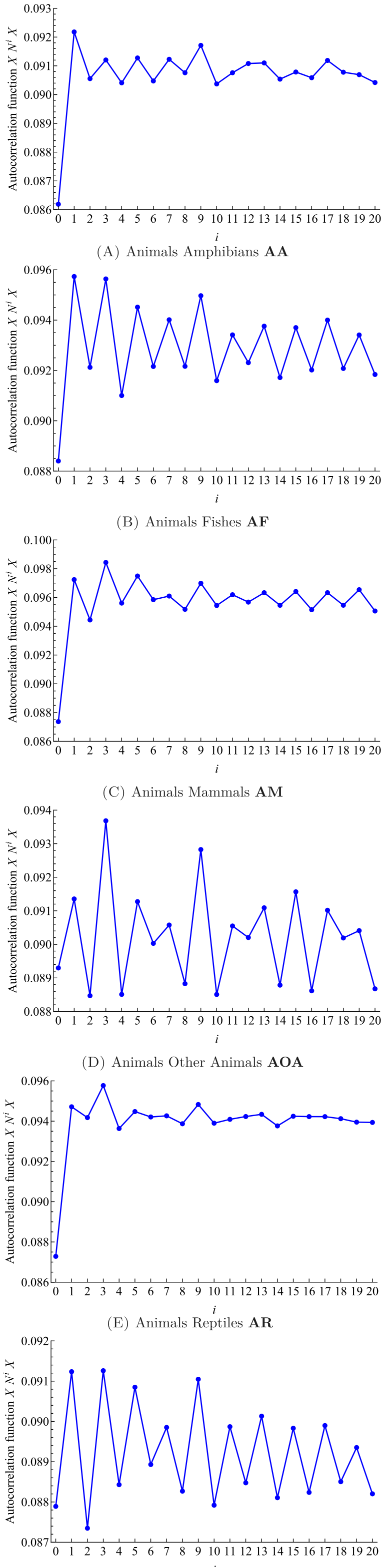
(B) Fungi Ascomycetes **FA**

(C) Fungi Basidiomycetes **FB**

(D) Plants Green Algae **PGA**

(E) Plants Land Plants **PLP**

Fig. 2. Circular code periodicity 0 modulo 3 identified by the circular code autocorrelation function $A_{X,X}(i,F)$ (Eq. (2.4)) in 5 subgroups of introns (Table 2). The abscissa represents the number i of nucleotides N between X and itself, i varying from 0 to $n = 20$. The ordinate gives the occurrence probability $A_{X,X}(i,F)$ of XN^iX in introns of: (A) Animals Birds **AB** (until $i < 17$). (B) Fungi Ascomycetes **FA**. (C) Fungi Basidiomycetes **FB**. (D) Plants Green Algae **PGA**. (E) Plants Land Plants **PLP**.



(A) Animals Amphibians **AA**

(B) Animals Fishes **AF**

(C) Animals Mammals **AM**

(D) Animals Other Animals **AOA**

(E) Animals Reptiles **AR**

(F) Protists Apicomplexans **PA**

Fig. 3. Periodicity 1 modulo 2 identified by the circular code autocorrelation function $A_{X,X}(i,F)$ (Eq. (2.4)) in 6 subgroups of introns (Table 2). The abscissa represents the number i of nucleotides N between X and itself, i varying from 0 to $n = 20$. The ordinate gives the occurrence probability $A_{X,X}(i,F)$ of XN^iX in introns of: (A) Animals Amphibians **AA** (until $i < 11$). (B) Animals Fishes **AF**. (C) Animals Mammals **AM**. (D) Animals Other Animals **AOA**. (E) Animals Reptiles **AR** (until $i < 11$). (F) Protists Apicomplexans **PA**.

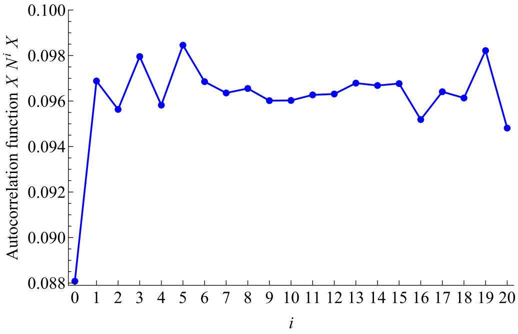


Fig. 4. Short periodicity 1 modulo 2 (until $i < 7$) identified by the circular code autocorrelation function $A_{X,X}(i,F)$ (Eq. (2.4)) in introns of *Homo sapiens* (24 chromosomes, 593,398 introns, 1,182,410,148 nucleotides). The abscissa represents the number i of nucleotides N between X and itself, i varying from 0 to $n = 20$. The ordinate gives the occurrence probability $A_{X,X}(i,F)$ of XN^iX in introns of *Homo sapiens*.

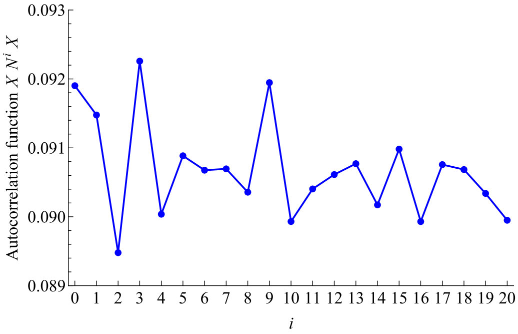


Fig. 5. Mixed periodicity modulo 2 and 3 identified by the circular code autocorrelation function $A_{X,X}(i,F)$ (Eq. (2.4)) in introns of Animals Insects **AI** (Table 2). The periodicity 0 modulo 3 is incomplete as there are no peaks at $i = 6, 12, 18$. Analogously, the periodicity 1 modulo 2 is also incomplete as there are no peaks at $i = 1, 11, 19$. A subperiodicity 3 modulo 6 can be observed: peaks at $i = 3, 9, 15$. The abscissa represents the number i of nucleotides N between X and itself, i varying from 0 to $n = 20$. The ordinate gives the occurrence probability $A_{X,X}(i,F)$ of XN^iX in introns of Animals Insects **AI**.

code X contains 2 such modulo 2 trinucleotides $N_1N_2N_1$: CTC and GAG(1.1). However, the construction of a sequence $(N_1N_2)^+$ involving a trinucleotide $N_1N_2N_1$ is an open problem that requires further investigation. For example, if $(N_1N_2)^+ = (N_1N_2N_1 \cdot N_2)^+$ then the tetranucleotide code $\{N_1N_2N_1N_2\}$ or the mixed code $\{N_2, N_1N_2N_1\}$ are both not circular. A hexanucleotide code could also be an area for research (see below Section 3.7.2).

Remark 3.3. In a self-complementary code, in particular the circular code X(1.1), trinucleotides $N_1N_2N_1$ always occur by complementary pairs: $N_1N_2N_1$ and $C(N_1)C(N_2)C(N_1)$ where C is the classical complementary map, e.g. CTC and GAG in X. Note that self-complementarity C preserves the property $N_1 = N_3$.

3.5. Circular code periodicity 0 modulo 3 hidden by the periodicity 1 modulo 2 in introns

In order to test if the circular code periodicity 0 modulo 3 is hidden by the periodicity 1 modulo 2 in introns of Animals Amphibians **AA**, Animals Fishes **AF**, Animals Mammals **AM**, Animals Other Animals **AOA**, Animals Reptiles **AR**, Protists Apicomplexans **PA** (see Section 3.2) and by the mixed periodicity in introns of Animals Insects **AI** (see Section 3.3), the series $(CT)^+$ and $(GA)^+$, revealed by the 2 trinucleotides CTC and GAG of the circular code X (involved in the circular code autocorrelation function $A_{X,X}(i,F)$ (Eq. (2.4))), are replaced in these introns by A^+, a nucleotide series that cannot be associated with a circular code (a periodic trinucleotide, such as AAA, cannot be a word of a circular code).

Very interestingly, when the series $(CT)^+$ and $(GA)^+$ are not considered in the introns with a periodicity 1 modulo 2, the circular code periodicity 0 modulo 3 is now retrieved in 5 groups of introns (Fig. 6): amphibians **AA**, fishes **AF**, other animals **AOA**, reptiles **AR** and insects **AI**. This approach has failed to retrieve the circular code periodicity 0 modulo 3 in introns of mammals **AM** and apicomplexans **PA** and needs investigation in the future.

3.6. Circular code periodicity 0 modulo 3 with the two permuted circular codes of X

The maximal C^3 self-complementary trinucleotide circular code X(1.1) has 2 permuted circular codes X_1 (shifted by 1 nucleotide in the $5' - 3'$ direction, i.e. to the right) and X_2 (shifted by 2 nucleotides in the $5' - 3'$ direction) (Arquès and Michel, 1996) where

$$X_1 = \{AAG, ACA, ACG, ACT, AGC, AGG, ATA, ATG, CCA, CCG,\\ GCG, GTG, TAG, TCA, TCC, TCG, TCT, TGC, TTA, TTG\} \tag{3.1}$$

and

$$X_2 = \{AGA, AGT, CAA, CAC, CAT, CCT, CGA, CGC, CGG, CGT,\\ CTA, CTT, GCA, GCT, GGA, TAA, TAT, TGA, TGG, TGT\}. \tag{3.2}$$

We extend the statistical study by applying the circular code autocorrelation functions $A_{X_1,X_1}(i,F)$ (Eq. (2.4) with X_1 defined in (3.1)) and $A_{X_2,X_2}(i,F)$ (Eq. (2.4) with X_2 defined in (3.2)) to the 2 subgroups of introns Fungi Ascomycetes **FA** and Plants Green Algae **PGA** where the circular code periodicity 0 modulo 3 is well observed (Fig. 2(B) and 2(D)).

The function $A_{X_1,X_1}(i,F)$ identifies a circular code periodicity 0 modulo 3 in introns of Fungi Ascomycetes **FA** (Fig. 7(A) as Fig. 2(B)) but surprisingly a periodicity 1 modulo 2 in introns of Plants Green Algae **PGA** (Fig. 7(B) in contrast to Fig. 2(D)).

The same periodicity result is found with the function $A_{X_2,X_2}(i,F)$ (Figs. 8(A) and 8(B)). Note that the Figs. 7(B) and 8(B) in introns of **PGA** are almost identical.

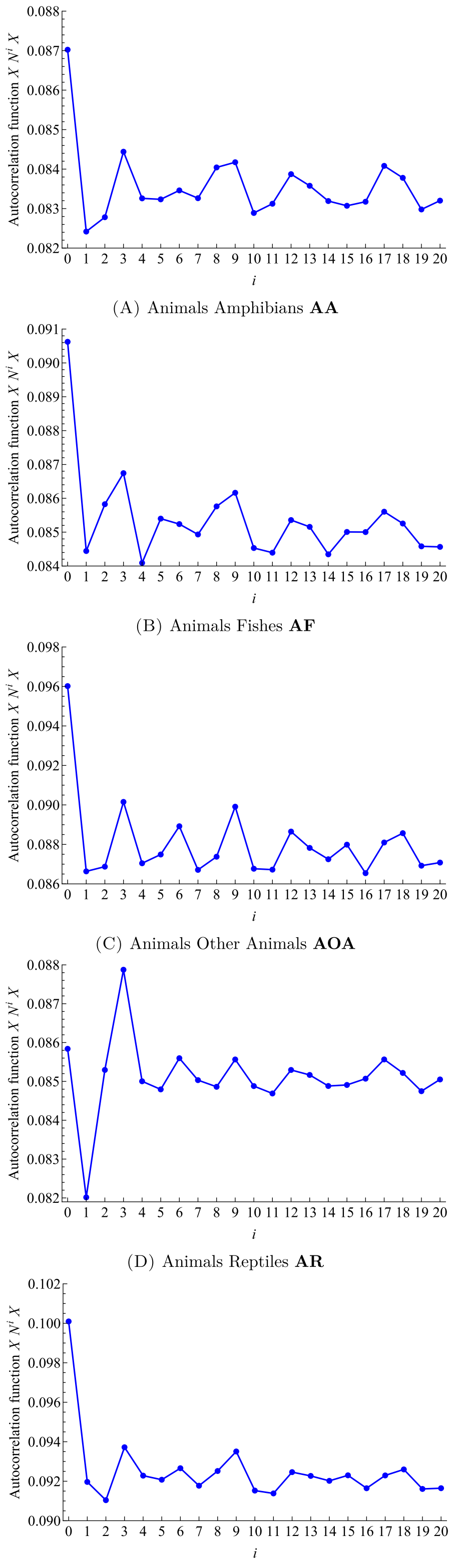
As in Section 3.5, the circular code periodicity 0 modulo 3 can be hidden by the periodicity 1 modulo 2 in introns of Plants Green Algae **PGA**. Thus, by applying the same approach, we identify the particular trinucleotides $N_1N_2N_1$ in X_1(3.1) and X_2(3.2). There are 5 trinucleotides $N_1N_2N_1 = \{ACA, ATA, GCG, GTG, TCT\}$ in X_1. There are (obviously) also 5 trinucleotides $N_1N_2N_1 = \{AGA, CAC, CGC, TAT, TGT\}$ in X_2. Thus, when computing the function $A_{X_1,X_1}(i,F)$, the 5 series $(AC)^+$, $(AT)^+$, $(GC)^+$, $(GT)^+$ and $(TC)^+$ are replaced in introns of **PGA** by A^+. Similarly, when computing the function $A_{X_2,X_2}(i,F)$, the 5 series $(AG)^+$, $(CA)^+$, $(CG)^+$, $(TA)^+$ and $(TG)^+$ are replaced in introns of **PGA** by A^+. Very interestingly, when these series are not considered, the circular code periodicity 0 modulo 3 is retrieved in introns of Plants Green Algae **PGA** with the functions $A_{X_1,X_1}(i,F)$ (Fig. 9(A)) and $A_{X_2,X_2}(i,F)$ (Fig. 9(B)).

In the next section, we study these particular trinucleotides $N_1N_2N_1$ in the 216 maximal C^3 self-complementary trinucleotide circular codes (Arquès and Michel, 1996), noted $C216$.

3.7. A few new properties on circular codes

3.7.1. Trinucleotides $N_1N_2N_1$ in the 216 maximal C^3 self-complementary trinucleotide circular codes

In the 216 maximal C^3 self-complementary trinucleotide circular codes $C216$, 24 codes $C216$ have no pair of trinucleotides $\{N_1N_2N_1, C(N_1)C(N_2)C(N_1)\}$. Four classes of 24 codes $C216$, thus 96 codes in



(A) Animals Amphibians **AA**

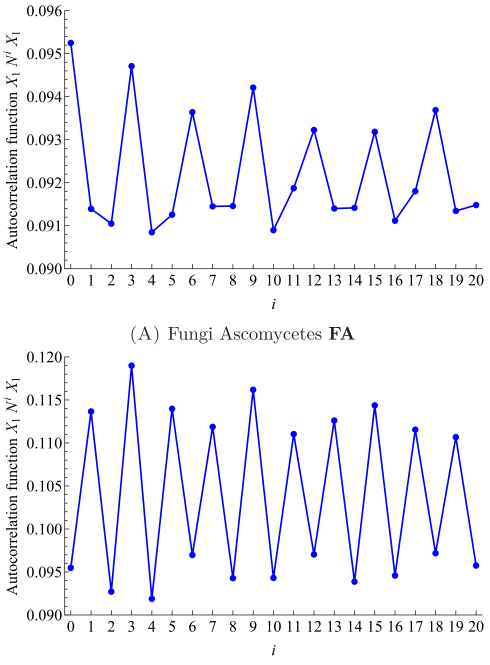
(B) Animals Fishes **AF**

(C) Animals Other Animals **AOA**

(D) Animals Reptiles **AR**

(E) Animals Insects **AI**

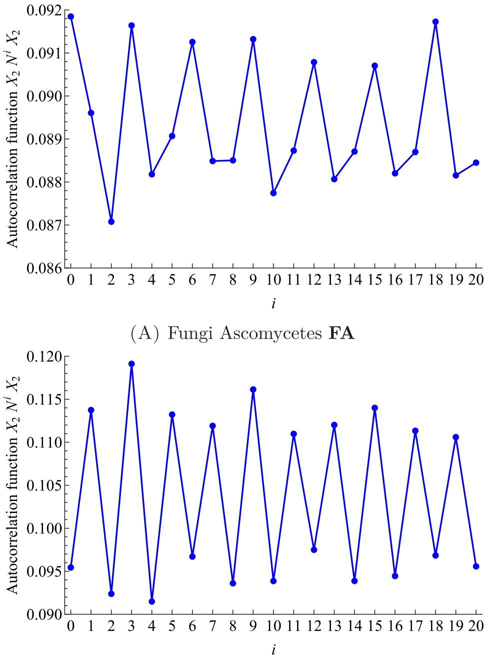
Fig. 6. Circular code periodicity hidden by the periodicity modulo 2 (1 modulo 2) and the mixed periodicity, now identified by the circular code autocorrelation function $A_{X,X}(i,F)$ (Eq. (2.4)) in 5 subgroups of introns (Table 2). The abscissa represents the number i of nucleotides N between X and itself, i varying from 0 to $n = 20$. The ordinate gives the occurrence probability $A_{X,X}(i,F)$ of XN^iX in introns of: (A) Animals Amphibians **AA** (until $i < 14$). (B) Animals Fishes **AF** (until $i < 17$, except $i = 6$). (C) Animals Other Animals **AOA**. (D) Animals Reptiles **AR** (until $i < 14$). (E) Animals Insects **AI**.



(A) Fungi Ascomycetes **FA**

(B) Plants Green Algae **PGA**

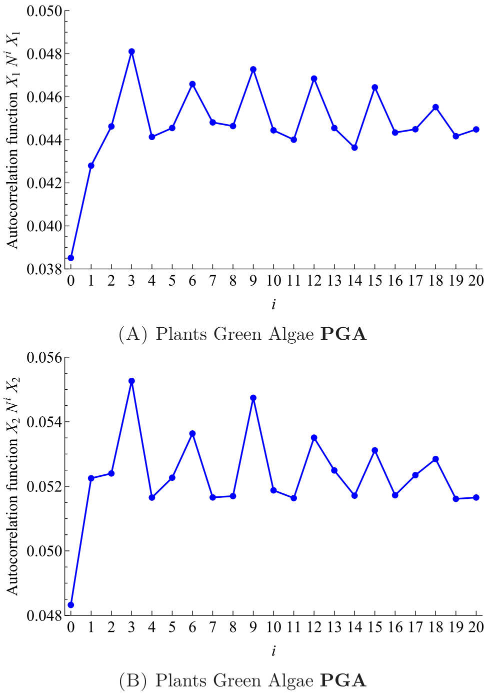
Fig. 7. Periodicities identified by the circular code autocorrelation function $A_{X_1,X_1}(i,F)$ (Eq. (2.4) with X_1 defined in (3.1)) in 2 subgroups of introns (Table 2). The abscissa represents the number i of nucleotides N between X_1 and itself, i varying from 0 to $n = 20$. The ordinate gives the occurrence probability $A_{X_1,X_1}(i,F)$ of $X_1N^iX_1$: (A) Circular code periodicity 0 modulo 3 in introns of Fungi Ascomycetes **FA**. (B) Periodicity 1 modulo 2 in introns of Plants Green Algae **PGA**.



(A) Fungi Ascomycetes **FA**

(B) Plants Green Algae **PGA**

Fig. 8. Periodicities identified by the circular code autocorrelation function $A_{X_2,X_2}(i,F)$ (Eq. (2.4) with X_2 defined in (3.2)) in 2 subgroups of introns (Table 2). The abscissa represents the number i of nucleotides N between X_2 and itself, i varying from 0 to $n = 20$. The ordinate gives the occurrence probability $A_{X_2,X_2}(i,F)$ of $X_2N^iX_2$: (A) Circular code periodicity 0 modulo 3 in introns of Fungi Ascomycetes **FA**. (B) Periodicity 1 modulo 2 in introns of Plants Green Algae **PGA**.



(A) Plants Green Algae **PGA**

(B) Plants Green Algae **PGA**

Fig. 9. Circular code periodicity hidden by the periodicity modulo 2 (1 modulo 2) in introns of Plants Green Algae **PGA** (Table 2) identified by the circular code autocorrelation functions $A_{X_1,X_1}(i,F)$ (Eq. (2.4) with X_1 defined in (3.1)) and $A_{X_2,X_2}(i,F)$ (Eq. (2.4) with X_2 defined in (3.2)). The abscissa represents the number i of nucleotides N between: (A) X_1 and itself; (B) X_2 and itself, i varying from 0 to $n = 20$. The ordinate gives the occurrence probability: (A) $A_{X_1,X_1}(i,F)$ of $X_1 N^i X_1$. (B) $A_{X_2,X_2}(i,F)$ of $X_2 N^i X_2$.

total, have one trinucleotide pair: $\{ACA, TGT\}$ or $\{AGA, TCT\}$ or $\{CAC, GTG\}$ or $\{CTC, GAG\}$. Two classes of 8 codes $C216$, thus 16 codes in total, have two trinucleotide pairs: $\{ACA, CTC, GAG, TGT\}$ or $\{AGA, CAC, GTG, TCT\}$. Two classes of 40 codes $C216$, thus 80 codes in total, have two trinucleotide pairs: $\{ACA, AGA, TCT, TGT\}$ or $\{CAC, CTC, GAG, GTG\}$. Obviously, by definition of code circularity, the 2 trinucleotide pairs $\{ATA, TAT\}$ and $\{CGC, GCG\}$ (case where $N_1 = C(N_2)$) cannot exist in the 216 codes $C216$.

3.7.2. Hexanucleotide codes modelling the periodicity 3 modulo 6

A striking property links the 3 classes of introns (modulo 2, modulo 3, modulo 2 and 3) (Section 3.3): a subperiodicity 3 modulo 6 (a multiple of 2 and 3). We propose a hexanucleotide code (words of 6 nucleotide length) to explain the periodicity 3 modulo 6. The code $\{X \cdot \overline{X}\}$ where $\overline{X} \subset B^3 \backslash X$ can generate the following sequence $x_0^1x_1^1x_2^1\overline{x}_0^1\overline{x}_1^1\overline{x}_2^1 \mid x_0^2x_1^2x_2^2\overline{x}_0^2\overline{x}_1^2\overline{x}_2^2 \mid x_0^3x_1^3x_2^3\overline{x}_0^3\overline{x}_1^3\overline{x}_2^3 \mid x_0^4x_1^4x_2^4\overline{x}_0^4\overline{x}_1^4\overline{x}_2^4 \mid \ldots$ where $x_0x_1x_2 \in X$ and $\overline{x}_0\overline{x}_1\overline{x}_2 \in \overline{X}$. Then $x_0^1x_1^1x_2^1 \in X$ and $x_0^2x_1^2x_2^2 \in X$ are separated by 3 nucleotides, $x_0^1x_1^1x_2^1 \in X$ and $x_0^3x_1^3x_2^3 \in X$, by 9 nucleotides, $x_0^1x_1^1x_2^1 \in X$ and $x_0^4x_1^4x_2^4 \in X$, by 15 nucleotides, $x_0^2x_1^2x_2^2 \in X$ and $x_0^3x_1^3x_2^3 \in X$, by 3 nucleotides, etc., thus generating a periodicity 3 modulo 6 with the circular code X. Such a hexanucleotide code could be a trace of more general circular codes at the origin of the circular code X(1.1) observed in genes. From a theoretical point of view, the definition and the combinatorial results of hexanucleotide codes with the properties of circularity, self-complementarity and of inclusion of the circular code X, are an open problem.

4. Conclusion

A circular code periodicity modulo 3 is identified in 5 subgroups of introns: birds, ascomycetes, basidiomycetes, green algae and land plants. This circular code periodicity, which is a property of retrieving the reading frame in (protein coding) genes, may suggest that these introns have a coding property and could be extinct genes, i.e. non functional genes. To our knowledge, such a property of introns has never been identified. It should refine the evolutionary hypotheses about introns.

A well-known periodicity modulo 2 is observed in 6 subgroups of introns: amphibians, fishes, mammals, other animals, reptiles and apicomplexans, thus almost the complete group of animals, except the birds and the insects. A mixed periodicity modulo 2 and 3 is found in the introns of insects. These statistical observations raise the following problem: is the circular code periodicity modulo 3 hidden by the periodicity modulo 2 and the mixed periodicity in these introns?

Very interestingly, when the particular trinucleotides $N_1N_2N_1$ of the circular code X are not considered, i.e. CTC and GAG, the circular code periodicity modulo 3, hidden by the periodicity modulo 2, is now retrieved in 5 groups of introns: amphibians, fishes, other animals, reptiles and insects (except for the mammals and apicomplexans), although noisier than in the genes. In summary, 10 groups of introns, taxonomically different, out of 12 have a coding property related to the reading frame retrieval.

The trinucleotides $N_1N_2N_1$ are analysed in the 216 maximal C^3 self-complementary trinucleotide circular codes. This class of trinucleotides which is associated with a periodicity modulo 2, deserves special attention in the theory of circular codes. Finally, a hexanucleotide code (words of 6 letters) is proposed to explain the periodicity 3 modulo 6. It could be a trace of more general circular codes at the origin of the circular code X observed in genes.

Memorandum

In memory of my parents Doctor Gilbert Pierre MICHEL (1931–1993) and Marie Denise MICHEL born BESCH (1930–2023), and my uncle Professor Jacques STREITH (1933–2023).

CRediT authorship contribution statement

Christian J. Michel: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Arquès, D.G., Michel, C.J., 1987a. Study of a perturbation in the coding periodicity. Math. Biosci. 86, 1–14.

Arquès, D.G., Michel, C.J., 1987b. A purine-pyrimidine motif verifying an identical presence in almost all gene taxonomic groups. J. Theoret. Biol. 128, 457–461.

Arquès, D.G., Michel, C.J., 1987c. Periodicities in introns. Nucleic Acids Res. 15, 7581–7592.

Arquès, D.G., Michel, C.J., 1990a. Periodicities in coding and noncoding regions of the genes. J. Theoret. Biol. 143, 307–318.

Arquès, D.G., Michel, C.J., 1990b. A model of DNA sequence evolution. Part 1: Statistical features and classification of gene populations, 743-753. Part 2: Simulation model, 753-766. Part 3: Return of the model to the reality, 766-770. Bull. Math. Biol. 52, 741–772.

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

Berget, S.M., Moore, C., Sharp, P.A., 1977. Spliced segments at the 5'-terminus of adenovirus 2 late mRNA. Proc. Natl. Acad. Sci. U.S.A. 74, 3171–3175.

Chow, L.T., Gelinas, R.E., Broker, T.R., Roberts, R.J., 1977. An amazing sequence arrangement at the 5'-ends of adenovirus 2 messenger RNA. Cell 12, 1–8.

Crick, F.H., Brenner, S., Klug, A., Pieczenik, G., 1976. A speculation on the origin of protein synthesis. Origins Life 7, 389–397.

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

Fickett, J.W., 1982. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. 10, 5303–5318.

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

Fimmel, E., Michel, C.J., Pirot, F., Sereni, J.-S., Strüngmann, L., 2019. Mixed circular codes. Math. Biosci. 317, 1–14, 108231.

Fimmel, E., Michel, C.J., Pirot, F., Sereni, J.-S., Strüngmann, L., 2023a. Diletter and triletter comma-free codes over finite alphabets. Australas. J. Combin. 86, 233–270, 2.

Fimmel, E., Michel, C.J., Strüngmann, L., 2023b. Circular mixed sets. Biosystems 229, 1–11, 104906.

Fimmel, E., Strüngmann, L., 2018. Mathematical fundamentals for the noise immunity of the genetic code. Biosystems 164, 186–198.

Gilbert, W., 1978. Why genes in pieces. Nature 271, 501.

Konopka, A.K., Smythers, G.W., 1987. DISTAN - A program which detects significant distances between short oligonucleotides. Bioinformatics 3, 193–201.

Michel, C.J., 1986. New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its evaluation. J. Theoret. Biol. 120, 223–236.

Michel, C.J., 2008. A 2006 review of circular codes in genes. Comput. Math. Appl. 55, 984–988.

Michel, C.J., 2015. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. J. Theor. Biol. 380, 156–177.

Michel, C.J., 2017. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, archaea, eukaryotes, plasmids and viruses. Life 7, 1–16, 2.

Michel, C.J., 2020. The maximality of circular codes in genes statistically verified. Biosystems 197, 1–7, 104201.

Michel, C.J., Mouillon, B., Sereni, J.-S., 2022. Trinucleotide k-circular codes I: Theory. Biosystems 217, 1–11, 104667.

Michel, C.J., Sereni, J.-S., 2022. Trinucleotide k-circular codes II: Biology. Biosystems 217, 1–18, 104668.

Michel, C.J., Sereni, J.-S., 2023. Reading frame retrieval of genes: A new parameter of codon usage based on the circular code theory. Bull. Math. Biol. 85, 1–21, 24.

Michel, C.J., Thompson, J.D., 2020. Identification of a circular code periodicity in the bacterial ribosome: origin of codon periodicity in genes? RNA Biol. 17, 571–583.

Shepherd, J.C.W., 1981a. Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. J. Mol. Evol. 17, 94–102.

Shepherd, J.C.W., 1981b. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci. U.S.A. 78, 1596–1600.