BioSystems 44 (1997) 107–134

A code in the protein coding genes

Didier G. Arquès [a],*, Christian J. Michel [1,b]

[a] *Equipe de Biologie Théorique, Université de Marne la Vallée, Institut Gaspard Monge, 2 rue de la Butte Verte,*
93160 Noisy Le Grand, France
[b] *Equipe de Biologie Théorique, Université de Franche-Comté, Institut Universitaire de Technologie de Belfort-Montbéliard, BP 527,*
90016 Belfort, France

Received 9 May 1997

Abstract

A statistical analysis with 12 288 autocorrelation functions applied in protein (coding) genes of prokaryotes and eukaryotes identifies three subsets of trinucleotides in their three frames: $T_0 = X_0 \cup \{\text{AAA, TTT}\}$ with $X_0 = \{$AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC$\}$ in frame 0 (the reading frame established by the ATG start trinucleotide), $T_1 = X_1 \cup \{\text{CCC}\}$ in frame 1 and $T_2 = X_2 \cup \{\text{GGG}\}$ in frame 2 (the frames 1 and 2 being the frame 0 shifted by one and two nucleotides, respectively, to the right). These three subsets are identical in these two gene populations and have five important properties: (i) the property of maximal (20 trinucleotides) circular code for X_0 (resp. X_1, X_2) allowing to retrieve automatically the frame 0 (resp. 1, 2) in any region of the gene without start codon; (ii) the DNA complementarity property $\mathscr{C}$ (e.g. $\mathscr{C}(\text{AAC}) = \text{GTT}$): $\mathscr{C}(T_0) = T_0$, $\mathscr{C}(T_1) = T_2$ and $\mathscr{C}(T_2) = T_1$ allowing the two paired reading frames of a DNA double helix simultaneously to code for amino acids; (iii) the circular permutation property $\mathscr{P}$ (e.g. $\mathscr{P}(\text{AAC}) = \text{ACA}$): $\mathscr{P}(X_0) = X_1$ and $\mathscr{P}(X_1) = X_2$ implying that the two subsets X_1 and X_2 can be deduced from X_0; (iv) the rarity property with an occurrence probability of $X_0 = 6 \times 10^{-8}$; and (v) the concatenation properties in favour of an evolutionary code: a high frequency (27.5%) of misplaced trinucleotides in the shifted frames, a maximum (13 nucleotides) length of the minimal window to retrieve automatically the frame and an occurrence of the four types of nucleotides in the three trinucleotide sites. In Discussion, a simulation based on an independent mixing of the trinucleotides of T_0 allows to retrieve the two subsets T_1 and T_2. Then, the identified subsets T_0, T_1 and T_2 replaced in the 2-letter genetic alphabet {R, Y} (R = purine = A or G, Y = pyrimidine = C or T) allow to retrieve the RNY model (N = R or Y) and to explain previous works in the alphabet {R, Y}. Then, these three subsets are related to the genetic code. The trinucleotides of T_0 code for 13 amino acids: Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Lys, Phe, Thr, Tyr and Val. Finally, a strong correlation between the usage of the trinucleotides of T_0 in protein genes and the amino acid frequencies in proteins is observed as six among seven amino acids not coded by T_0, have as expected the lowest frequencies in proteins of both prokaryotes and eukaryotes. © 1997 Elsevier Science Ireland Ltd.

* Corresponding author. Tel.: +33 149329010; fax: +33 149329138; e-mail: arques@univ-mlv.fr

[1] Present address: Equipe de Biologie Théorique, Institut Polytechnique de Sévenans, Rue du Château, Sévenans, 90010 Belfort, France.

PII S0303-2647(97)00049-X

Keywords: Circular code; Complementarity; Protein coding gene; Autocorrelation function; Trinucleotide; Frame

1. Introduction

The concept of code without comma introduced by Crick et al. (1957) is a code readable in only one frame and without a start signal. Such a theoretical code 'without comma' is a set *X* of codons so that their concatenation (series of codons) leads to genes which have the interesting property to retrieve automatically the concatenation of codons of *X* without the usage of a start codon in the case of the trace of this initial concatenation is lost (the 'commas' dividing the series of nucleotides into groups of three for constituting the codons in the initial concatenation). Such a code was proposed in order to explain how the reading of a series of nucleotides in the protein (coding) genes could code for the amino acids constituting the proteins. The two problems stressed were: why are there more codons than amino acids and how to choose the reading frame? For example, a series of nucleotides ...AGTCCGTACGA... can be read in three frames: ...AGT, CCG, TAC, GA..., ...A, GTC, CGT, ACG, A... and ...AG, TCC, GTA, CGA, ... Crick et al. (1957) have then proposed that only 20 among 64 codons, code for the 20 amino acids. However, the determination of a set of 20 codons forming a code *X* without commas depends on a great number of constraints. For example, the four codons with identical nucleotides AAA, CCC, GGG and TTT must be excluded from such a code. Indeed, the concatenation of AAA, for example, with itself does not allow to retrieve the frame: ...AAA, AAA, AAA, ..., ...A, AAA, AAA, AA... and ...AA, AAA, AAA, A... Similarly, two codons related to circular permutation, e.g. AAC and ACA (or CAA), cannot belong at the same time to such a code. Indeed, the concatenation of AAC, for example, with itself leads to the concatenation of ACA (or CAA) with itself in another frame, making the frame determination impossible. Therefore, by excluding AAA, CCC, GGG and TTT and by gathering the 60 remaining codons in 20 classes of three codons so that, in each class, the three codons are deduced from each other by circular permutations, e.g. AAC, ACA and CAA, a code without commas has only one codon per class and therefore contains at most 20 codons. This codon number is identical to the amino acid number. This remark has naturally led to propose a code without commas assigning one codon per amino acid (Crick et al., 1957).

In contrast, Dounce (1952) has earlier proposed an evolutionary code associating several codons per amino acid. Such a flexibility can explain the variations in G + C composition observed in the actual protein genes (Jukes and Bhushan, 1986).

The two discoveries that the codon TTT, an 'excluded' codon in the concept of code without commas, codes for phenylalanine (Nirenberg and Matthaei, 1961) and that the protein genes are placed in the reading frame with a particular codon, namely the start codon ATG, have led to give up the concept of code without commas in the alphabet {A, C, G, T}. For several biological reasons, in particular the interaction between mRNA and tRNA, the concept of code without commas is resumed later in the alphabet {R, Y} (R = purine = A or G, Y = pyrimidine = C or T) with two codon models for the primitive protein genes: RRY (Crick et al., 1976) and RNY (N = R or Y) (Eigen and Schuster, 1978).

In order to understand the circular code identified here in protein genes of prokaryotes and eukaryotes in the alphabet {A, C, G, T}, the concept of circular code is introduced in the alphabet {R, Y} with the RNY codon model (Eigen and Schuster, 1978). If a sequence, e.g. a protein gene, is constructed by concatenating trinucleotides of the type RNY, i.e. RRY and RYY forming the frame 0 (reading frame), and if the frame of construction is lost, e.g. situation observed when a region of a protein gene without a start codon is sequenced, then the property of code assures that the constructed sequence can be decomposed into a series of RNY trinucleotides according to a unique way. This unique decomposition, and therefore the frame 0, can be retrieved using a window of nucleotides with a minimal length depending on RNY. In the actual protein genes, the frame 0 is determined by the start codon. The notion of circular added to the concept of code concerns the limit case with sequences of infinite length, i.e. without a beginning and an end.

RNY is a circular code. Indeed, the concatenation of two trinucleotides of RNY, ...RRYRRY..., ...RRYRYY..., ...RYYRRY... and ...RYYRYY... leads to only one decomposition over RNY. This because the eventual decomposition in frame 1 has always an R in the 3rd position but no trinucleotide of RNY ends with R and as the eventual decomposition in frame 2 has always an Y in the first position but no trinucleotide of RNY begins with Y. The RNY codon model leads to a protein gene formed by a series RNYRNY... of nucleotides so that there is one type of trinucleotide RNY in frame 0 (reading frame), one type of trinucleotide NYR in frame 1 and one type of trinucleotide YRN in frame 2 (frames 1 and 2 being the frame 0 shifted by one and two nucleotides, respectively, in the 5′–3′ direction). RNY is self-complementary and, NYR and YRN are complementary to each other. This property allows the two paired reading frames to simultaneously code for amino acids according to a purine/pyrimidine genetic code. Furthermore, NYR (resp. YRN) is obtained by one (resp. 2) circular permutation of RNY. This property allows NYR and YRN to be deduced from RNY. Finally, the length of the minimal window to automatically retrieve the frame 0 in a series RNYRNY... is obviously equal to three nucleotides. Indeed, two nucleotides are insufficient as RY is both in frame 1 of RRY (RNY with N = R) and in frame 0 of RYY (RNY with N = Y). This property allows to automatically retrieve the reading frame in any region of the gene (formed by a series of RNY codons) without a start codon.

Recently, the 64 autocorrelation functions analyzing in the protein genes of prokaryotes and eukaryotes, the occurrence probability of two trinucleotides obtained by specifying YRY on {A, C, G, T} (YRY leads to eight specified trinucleotides {CAC, CAT, ..., TGT}) and separated by any i bases N (N = A or C or G or T), have allowed to retrieve the classical periodicity 0 modulo 3 (maximal values of the function at i = 0, 3, 6, etc.) and unexpectedly, to identify shifted modulo 3 periodicities: 1 modulo 3 (maximal values of the function at i = 1, 4, 7, etc.) and 2 modulo 3 (maximal values of the function at i = 2, 5, 8, etc.) (Arquès et al., 1995). This result means that these eight trinucleotides are associated with the three frames of protein genes, some trinucleotides with the reading frame, and others with the reading frame shifted by one or two nucleotides in the 5′–3′ direction. However, as the definition used is based on the average of the three frames 0, 1 and 2, it is impossible to associate a particular frame with a given trinucleotide.

In order to extend this result to the 64 trinucleotides {AAA, ..., TTT} and to identify subsets of trinucleotides having a preferential occurrence frame, the previous approach is generalized here to the 12 288 autocorrelation functions analyzing the probability that a trinucleotide in any frame occurs any i bases N after a trinucleotide in a given frame of the protein genes. This statistical analysis identifies three subsets of trinucleotides per frame in protein genes of both prokaryotes and eukaryotes which have five important properties: maximal circular code, DNA complementarity, circular permutation, rarity and concatenation. Therefore, the code identified here in protein genes of prokaryotes and eukaryotes in the alphabet {A, C, G, T} retrieves the properties both of the code RNY in the reduced alphabet {R, Y} (Eigen and Schuster, 1978) and of the evolutionary code (Dounce, 1952).

2. Method

2.1. Definition of the autocorrelation function in frame

This definition generalizes the autocorrelation function definition in the alphabet {R, Y} and without considering the frame (Arquès and Michel, 1987) to the alphabet $\mathscr{B} = \{A, C, G, T\}$

and in frame. For a gene population F, two trinucleotides w and w′ and a frame p, the ‘autocorrelation function in frame’ noted $w^p(N)_i w'$ is defined as the function of i giving the occurrence probability in the gene population F that the trinucleotide w′ in any frame occurs any i bases N after the trinucleotide w read in frame p.

More precisely, let F be a gene population (language) with n(F) protein (coding) genes (words). Let s be a protein gene in F of length l(s). Let w be a trinucleotide (word of three letters) on $\mathscr{B}$, $w \in T = \{AAA, \ldots, TTT\}$ (64 trinucleotides). A trinucleotide w read in frame p in a protein gene is noted w^p, $p \in \{0, 1, 2\}$ ($64 \times 3 = 192$ trinucleotides in the three frames). Let the i-motif $m_i^p = w^p(N)_i w'$ be two trinucleotides w^p in frame p and w′ in any frame separated by any i bases N, with $w, w' \in T$, $p \in \{0, 1, 2\}$, $i \in [0, 99]$ and N = A, C, G or T, i.e. $m_i^p \in \{AAA^p(N)_i AAA, \ldots, TTT^p(N)_i TTT\}$ ($64^2 \times 3 = 12\,288$ i-motifs). For each protein gene s of F, the counter $c_i^p(s)$ counts the occurrences of m_i^p in s. The occurrence probability $o_i^p(s)$ of m_i^p in s, is then equal to the ratio of the counter by the total number t(s) of trinucleotides read in frame p, $o_i^p(s) = c_i^p(s)/t(s)$ with $t(s) = (l(s) - 104)/3$ $(104 = 99 + 2 \times 3 - 1)$. The occurrence probability $A^p_{w,w'}(i, F)$ of the i-motif m_i^p in a gene population F, is $A^p_{w,w'}(i, F) = [1/n(F)] \times \Sigma_{s \in F}\ o_i^p$.

For a gene population F, the autocorrelation function $w^p(N)_i w'$ is then the function $i \rightarrow A^p_{w,w'}(i,F)$ giving the occurrence probability that w′ in any frame occurs any i bases N after w^p in frame p and is represented as a curve as follows: (i) the abscissa shows the number i of bases N between w^p and w′, by varying i between 0 and 99 (for clarity reasons, figures are given with a lesser number of points); (ii) the ordinate gives the occurrence probability of $w^p(N)_i w'$ in a gene population F.

2.2. Data acquisition

The two gene populations F analyzed here with the 12 288 autocorrelation functions are the protein genes of prokaryotes F = CPRO (13 686 sequences, 14 167 kb) and (nuclear) eukaryotes F = CEUK (26 757 sequences, 34 227 kb). The sequences with a minimal length of 200 bases are obtained from the release 39 of the EMBL Nucleotide Sequence Data Library in the same way as described in previous papers (Arquès and Michel, 1990a,b for a description of data acquisition). These large populations allow to obtain significant statistical results.

3. Results

3.1. Identification of shifted periodicities in protein genes

The 12 288 autocorrelation functions applied in the protein genes of prokaryotes and eukaryotes are non-random. Indeed, almost all autocorrelation functions have a modulo 3 periodicity among the three following types:

(i) Type 0: a periodicity 0 modulo 3 (maximal values of the function at i = 0, 3, 6, etc.), three examples are given in Fig. 1(a)–(c) with F = CPRO;

(ii) Type 1: a periodicity 1 modulo 3 (maximal values of the function at i = 1, 4, 7, etc.), three examples are given in Fig. 1(d)–(f) with F = CPRO;

(iii) Type 2: a periodicity 2 modulo 3 (maximal values of the function at i = 2, 5, 8, etc.), three examples are given in Fig. 1(g)–(i) with F = CPRO.

The identification of shifted periodicities (types 1 and 2) in protein genes is new.

The method to classify the autocorrelation function $w^p(N)_i w'$, w, w′ in T and p in {0, 1, 2}, consists in determining the number of points of each modulo 3 type which are greater than their two adjacent points. The maximal number is 33 points for a curve with 100 points ($i \in [0, 99]$). If the number of points for the type p is > 22 (corresponding to a statistical level of significance $< 10^{-4}$), then the autocorrelation function is said to be classifiable according to this type p. With the statistical level chosen, an autocorrelation function can only be classifiable in one type. A few autocorrelation functions cannot be classifiable, in particular an autocorrelation function with an i-motif containing a stop trinucleotide is

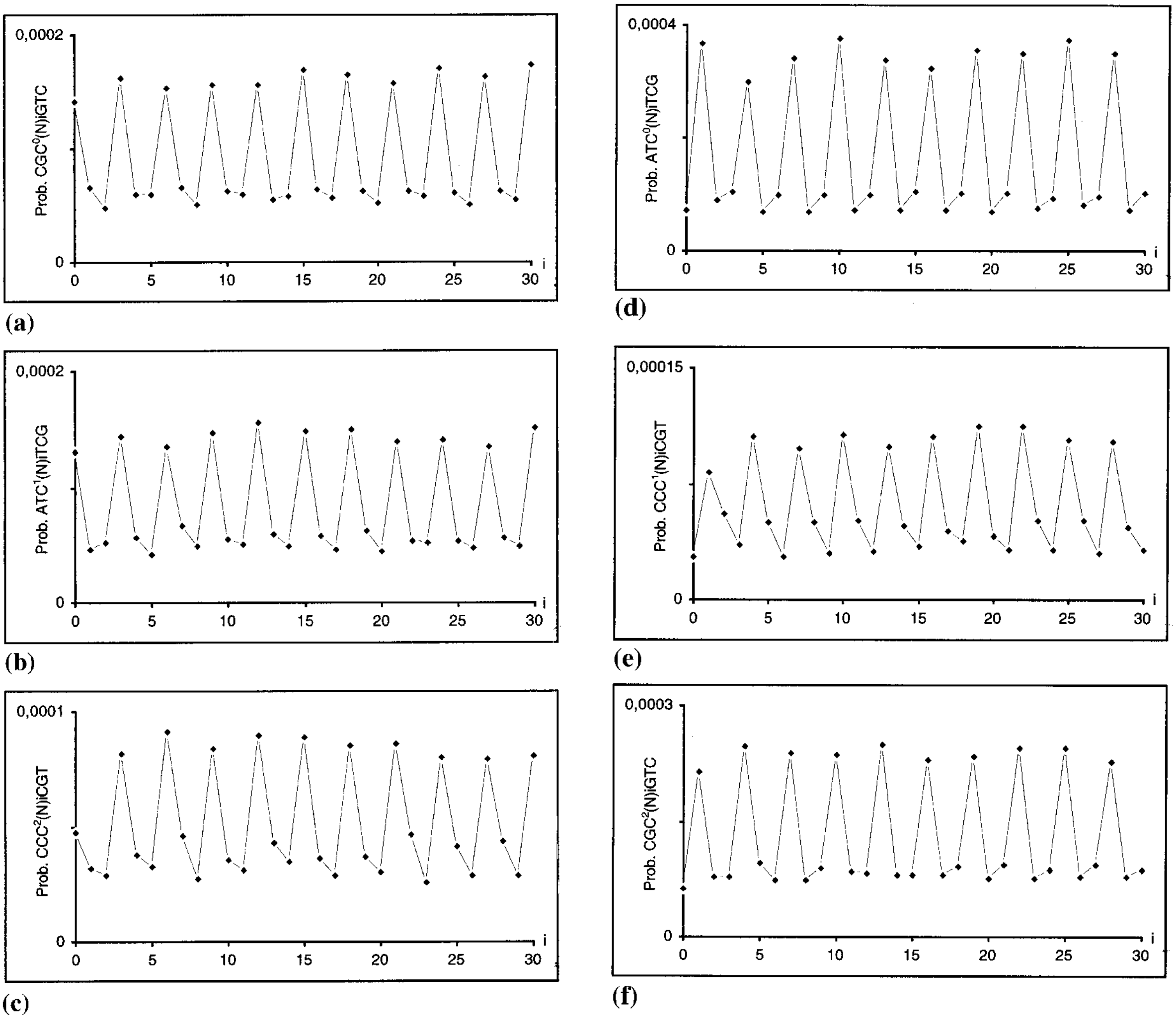
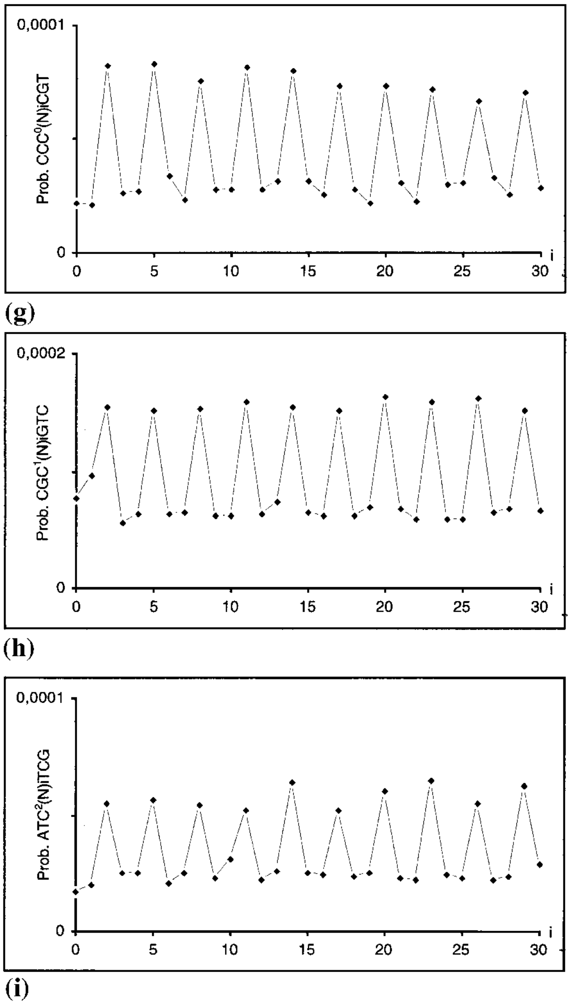


Fig. 1. (a) Periodicity 0 modulo 3 in frame 0 of the prokaryotic protein coding genes. The horizontal axis represents the number i, $i \in [0, 30]$, of any bases N between CGC^0 in frame 0 and GTC in any frame. The vertical axis represents the autocorrelation function $CGC^0(N)_iGTC$; (b) periodicity 0 modulo 3 in frame 1 of the prokaryotic protein coding genes. The horizontal axis represents the number i, $i \in [0, 30]$, of any bases N between ATC^1 in frame 1 and TCG in any frame. The vertical axis represents the autocorrelation function $ATC^1(N)_iTCG$; (c) periodicity 0 modulo 3 in frame 2 of the prokaryotic protein coding genes. The horizontal axis represents the number i, $i \in [0, 30]$, of any bases N between CCC^2 in frame 2 and CGT in any frame. The vertical axis represents the autocorrelation function $CCC^2(N)_iCGT$; (d) periodicity 1 modulo 3 in frame 0 of the prokaryotic protein coding genes. The horizontal axis represents the number i, $i \in [0, 30]$, of any bases N between ATC^0 in frame 0 and TCG in any frame. The vertical axis represents the autocorrelation function $ATC^0(N)_iTCG$; (e) periodicity 1 modulo 3 in frame 1 of the prokaryotic protein coding genes. The horizontal axis represents the number i, $i \in [0, 30]$, of any bases N between CCC^1 in frame 1 and CGT in any frame. The vertical axis represents the autocorrelation function $CCC^1(N)_iCGT$; (f) periodicity 1 modulo 3 in frame 2 of the prokaryotic protein coding genes. The horizontal axis represents the number i, $i \in [0, 30]$, of any bases N between CGC^2 in frame 2 and GTC in any frame. The vertical axis represents the autocorrelation function $CGC^2(N)_iGTC$; (g) periodicity 2 modulo 3 in frame 0 of the prokaryotic protein coding genes. The horizontal axis represents the number i, $i \in [0, 30]$, of any bases N between CCC^0 in frame 0 and CGT in any frame. The vertical axis represents the autocorrelation function $CCC^0(N)_iCGT$; (h) periodicity 2 modulo 3 in frame 1 of the prokaryotic protein coding genes. The horizontal axis represents the number i, $i \in [0, 30]$, of any bases N between CGC^1 in frame 1 and GTC in any frame. The vertical axis represents the autocorrelation function $CGC^1(N)_iGTC$; (i) periodicity 2 modulo 3 in frame 2 of the prokaryotic protein coding genes. The horizontal axis represents the number i, $i \in [0, 30]$, of any bases N between ATC^2 in frame 2 and TCG in any frame. The vertical axis represents the autocorrelation function $ATC^2(N)_iTCG$.



(g)

(h)

(i)

Fig. 1. (*Continued*)

equal to 0 (TAA, TAG or TGA do not occur in frame 0).

These three types of periodicities exist whatever the frame p of the trinucleotide w^p. With the examples given previously, the periodicity 0 modulo 3 can be observed in frame 0 (e.g. the autocorrelation function $CGC^0(N)_iGTC$: Fig. 1(a)), in frame 1 (e.g. the autocorrelation function $ATC^1(N)_iTCG$: Fig. 1(b)) and in frame 2 (e.g. the autocorrelation function $CCC^2(N)_iCGT$: Fig. 1(c)). Similarly, the periodicity 1 (resp. 2) modulo 3 can be observed in frame 0 (e.g. resp. the autocorrelation functions $ATC^0(N)_iTCG$: Fig. 1(d) and $CCC^0(N)_iCGT$: Fig. 1(g)), in frame 1 (e.g. resp. the autocorrelation functions $CCC^1(N)_iCGT$: Fig. 1(e) and $CGC^1(N)_iGTC$: Fig. 1(h)) and in frame 2 (e.g. resp. the autocorrelation functions $CGC^2(N)_iGTC$: Fig. 1(f) and $ATC^2(N)_iTCG$: Fig. 1(i)).

If the autocorrelation function $w^p(N)_iw'$, w, w′ in *T* and p in {0, 1, 2}, has a periodicity j modulo 3 then the trinucleotide w′ preferentially occurs in frame $q(w, p) = p + j$ modulo 3 after w in frame p. Table 1 illustrates this remark with the previous examples (Fig. 1(a)–(i)). Table 1 also serves to introduce the next section by showing with these

Table 1
Frame $q=p+j$ of w' deduced from the frame p of w and the type of modulo 3 periodicity identified with the autocorrelation functions $w^p(N)_iw'$ given in Fig. 1(a)–(i)

Autocorrelation function $w^p(N)_iw'$	Periodicity $j=0$ modulo 3	Periodicity $j=1$ modulo 3	Periodicity $j=2$ modulo 3	w' in frame $q=p+j$ modulo 3
$ATC^{p=0}(N)_iTCG$		Fig. 1(d)		TCG in frame $p+j=0+1=1$ modulo 3
$ATC^{p=1}(N)_iTCG$	Fig. 1(b)			TCG in frame $p+j=1+0=1$ modulo 3
$ATC^{p=2}(N)_iTCG$			Fig. 1(i)	TCG in frame $p+j=2+2=1$ modulo 3
$CCC^{p=0}(N)_iCGT$			Fig. 1(g)	CGT in frame $p+j=0+2=2$ modulo 3
$CCC^{p=1}(N)_iCGT$		Fig. 1(e)		CGT in frame $p+j=1+1=2$ modulo 3
$CCC^{p=2}(N)_iCGT$	Fig. 1(c)			CGT in frame $p+j=2+0=2$ modulo 3
$CGC^{p=0}(N)_iGTC$	Fig. 1(a)			GTC in frame $p+j=0+0=0$ modulo 3
$CGC^{p=1}(N)_iGTC$			Fig. 1(h)	GTC in frame $p+j=1+2=0$ modulo 3
$CGC^{p=2}(N)_iGTC$		Fig. 1(f)		GTC in frame $p+j=2+1=0$ modulo 3

examples that the three trinucleotides $w'=$ CGT, GTC and TCG have a constant preferential occurrence frame 2, 0 and 1, respectively, independent both of the three trinucleotides $w=$ ATC, CCC and CGC and their frame p.

3.2. Preferential occurrence frame of a trinucleotide

Property 1: for each trinucleotide w' in T (with a few exceptions given below), the autocorrelation functions $w^p(N)_iw'$ obtained by varying w in T and p in {0, 1, 2}, have a periodicity j modulo 3 implying a constant preferential occurrence frame q(w, p) in {0, 1, 2} ($=p+j$ modulo 3) for w' independent of w and p.

For a trinucleotide w', the classifiable autocorrelation function $w^p(N)_iw'$ implies a frame q for w'. A maximum of $64\times 3=192$ autocorrelation functions are classifiable. As the three stop trinucleotides do not occur in frame 0, only $192-3=189$ autocorrelation functions are classifiable in protein genes. For each trinucleotide w' and for each frame q in the protein genes of prokaryotes (resp. eucaryotes), Table 2(a) (resp. Table 2(b)) gives the number of classifiable autocorrelation functions $w^p(N)_iw'$ implying the frame q for w'. Very unexpectedly, these two Table 2(a,b) show that the trinucleotides w' have a constant preferential occurrence frame and can easily be classified in three subsets of trinucleotides according to the frame (Table 3(a)). The 22 trinucleotides in frame 0 form the subset $T_0=$ {AAA, AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC, TTT} and the 21 trinucleotides in each of the frames 1 and 2, the subsets T_1 and T_2, respectively, $T=T_0\cup T_1\cup T_2$ with T_1 and T_2 defined in Table 3(a). By considering the four trinucleotides with identical nucleotides, three subsets X_0, X_1 and X_2 of 20 trinucleotides can be defined from T_0, T_1 and T_2: $X_0=T_0-$ {AAA, TTT}, $X_1=T_1-$ {CCC} and $X_2=T_2-$ {GGG}. The same three subsets T_0, T_1 and T_2 are retrieved for the two populations. The few trinucleotides, classified into two frames or misclassified (two for the prokaryotes and five for the eukaryotes), have been assigned to the frame according to the properties identified with the other trinucleotides, both in prokaryotes and eukaryotes.

Table 2
For each trinucleotide w′ and for each frame q in the prokaryotic protein coding genes (a) and in the eukaryotic protein coding genes (b) the number of classifiable autocorrelation functions $w^p(N)_i w'$ implying the frame q for w′ is given

w′ in q = 0	Number	w′ in q = 1	Number	w′ in q = 2	Number	Total
(a) *Prokaryotic protein coding genes*						
AAA	189	AAA	0	AAA	0	189
AAC	161	AAC	3	AAC	5	169
AAG	2	**AAG**	187	AAG	0	189
AAT	189	AAT	0	AAT	0	189
ACA	19	**ACA**	161	ACA	0	180
ACC	188	ACC	0	ACC	0	188
ACG	0	**ACG**	189	ACG	0	189
ACT	75	**ACT**	79	ACT	17	171
AGA	0	AGA	29	**AGA**	154	183
AGC	0	**AGC**	99	AGC	87	186
AGG	0	**AGG**	127	AGG	54	181
AGT	0	AGT	53	**AGT**	129	182
ATA	0	**ATA**	189	ATA	0	189
ATC	185	ATC	1	ATC	0	186
ATG	17	**ATG**	163	ATG	0	180
ATT	189	ATT	0	ATT	0	189
CAA	7	CAA	4	**CAA**	165	176
CAC	0	CAC	4	**CAC**	183	187
CAG	110	CAG	78	CAG	0	188
CAT	0	CAT	0	**CAT**	189	189
CCA	5	**CCA**	179	CCA	0	184
CCC	0	**CCC**	146	CCC	20	166
CCG	0	**CCG**	189	CCG	0	189
CCT	22	CCT	0	**CCT**	155	177
CGA	0	CGA	0	**CGA**	189	189
CGC	0	CGC	0	**CGC**	189	189
CGG	0	CGG	0	**CGG**	188	188
CGT	4	CGT	0	**CGT**	178	182
CTA	0	CTA	97	*CTA*	92	189
CTC	123	CTC	53	CTC	0	176
CTG	141	CTG	42	CTG	0	183
CTT	30	CTT	0	**CTT**	136	166
GAA	189	GAA	0	GAA	0	189
GAC	189	GAC	0	GAC	0	189
GAG	189	GAG	0	GAG	0	189
GAT	149	GAT	0	GAT	31	180
GCA	94	GCA	37	*GCA*	39	170
GCC	189	GCC	0	GCC	0	189
GCG	59	**GCG**	81	GCG	0	140
GCT	74	GCT	0	**GCT**	108	182
GGA	4	GGA	0	**GGA**	183	187
GGC	184	GGC	0	GGC	4	188
GGG	0	GGG	0	**GGG**	186	186
GGT	98	GGT	0	GGT	88	186
GTA	70	GTA	82	GTA	0	152
GTC	189	GTC	0	GTC	0	189
GTG	110	*GTG*	72	GTG	0	182
GTT	115	GTT	0	GTT	70	185
TAA	0	TAA	0	**TAA**	189	189
TAC	98	TAC	0	TAC	88	186
TAG	0	**TAG**	189	TAG	0	189

Table 2 (continued)

w′ in q = 0	Number	w′ in q = 1	Number	w′ in q = 2	Number	Total
TAT	1	TAT	0	**TAT**	185	186
TCA	0	**TCA**	156	TCA	14	170
TCC	0	**TCC**	105	TCC	78	183
TCG	0	**TCG**	189	TCG	0	189
TCT	23	*TCT*	51	TCT	80	154
TGA	0	TGA	30	**TGA**	154	184
TGC	0	*TGC*	91	TGC	95	186
TGG	0	TGG	170	*TGG*	12	182
TGT	0	TGT	2	**TGT**	184	186
TTA	2	**TTA**	180	TTA	0	182
TTC	120	TTC	16	TTC	16	152
TTG	0	**TTG**	189	TTG	0	189
TTT	69	TTT	0	TTT	78	147
(b) *Eukaryotic protein coding genes*						
AAA	187	AAA	0	AAA	0	187
AAC	187	AAC	2	AAC	0	189
AAG	183	*AAG*	6	AAG	0	189
AAT	189	AAT	0	AAT	0	189
ACA	0	**ACA**	189	ACA	0	189
ACC	108	ACC	60	ACC	3	171
ACG	0	**ACG**	189	ACG	0	189
ACT	68	**ACT**	118	ACT	0	186
AGA	0	AGA	109	*AGA*	77	186
AGC	0	**AGC**	166	AGC	19	185
AGG	0	**AGG**	189	AGG	0	189
AGT	2	AGT	105	*AGT*	74	181
ATA	0	**ATA**	189	ATA	0	189
ATC	189	ATC	0	ATC	0	189
ATG	0	**ATG**	189	ATG	0	189
ATT	189	ATT	0	ATT	0	189
CAA	1	CAA	0	**CAA**	188	189
CAC	0	CAC	2	**CAC**	187	189
CAG	124	CAG	61	CAG	1	186
CAT	0	CAT	0	**CAT**	189	189
CCA	3	**CCA**	186	CCA	0	189
CCC	15	**CCC**	75	CCC	55	145
CCG	0	**CCG**	189	CCG	0	189
CCT	27	CCT	0	**CCT**	155	182
CGA	0	CGA	0	**CGA**	189	189
CGC	7	CGC	0	**CGC**	182	189
CGG	0	CGG	1	**CGG**	188	189
CGT	0	CGT	0	**CGT**	189	189
CTA	0	CTA	97	*CTA*	92	189
CTC	61	CTC	60	CTC	39	160
CTG	95	CTG	94	CTG	0	189
CTT	0	CTT	0	**CTT**	189	189
GAA	96	GAA	0	GAA	93	189
GAC	189	GAC	0	GAC	0	189
GAG	189	GAG	0	GAG	0	189
GAT	188	GAT	0	GAT	1	189
GCA	48	GCA	121	*GCA*	8	177
GCC	189	GCC	0	GCC	0	189

Table 2 (continued)

w′ in q = 0	Number	w′ in q = 1	Number	w′ in q = 2	Number	Total
GCG	0	**GCG**	189	GCG	0	189
GCT	118	GCT	3	*GCT*	65	186
GGA	1	GGA	0	**GGA**	188	189
GGC	108	GGC	0	GGC	67	175
GGG	0	GGG	0	**GGG**	189	189
GGT	101	GGT	0	GGT	88	189
GTA	4	GTA	130	GTA	53	187
GTC	189	GTC	0	GTC	0	189
GTG	117	*GTG*	72	GTG	0	189
GTT	131	GTT	0	GTT	54	185
TAA	0	TAA	0	**TAA**	189	189
TAC	188	TAC	0	TAC	1	189
TAG	0	**TAG**	189	TAG	0	189
TAT	94	TAT	0	*TAT*	84	178
TCA	0	**TCA**	189	TCA	0	189
TCC	6	**TCC**	131	TCC	40	177
TCG	0	**TCG**	189	TCG	0	189
TCT	70	**TCT**	101	TCT	0	171
TGA	0	TGA	0	**TGA**	189	189
TGC	2	*TGC*	45	TGC	140	187
TGG	0	TGG	141	*TGG*	47	188
TGT	0	TGT	0	**TGT**	189	189
TTA	1	**TTA**	188	TTA	0	189
TTC	189	TTC	0	TTC	0	189
TTG	0	**TTG**	189	TTG	0	189
TTT	106	TTT	0	TTT	73	179

The last column gives the total number of classifiable autocorrelation functions. The trinucleotides in bold have a preferential occurrence frame.

(a) The trinucleotides in italics, classified into two frames (CTA, GTA, GTG, TCT, TGC, TTT) or misclassified (<50: GCA, TGG), have been assigned to the frame according to the properties identified with the other trinucleotides, both in prokaryotes and eukaryotes.

(b) The trinucleotides in italics, classified into two frames (AGA, AGT, CTA, GCT, GTG, TAT) or misclassified (<50: AAG, GCA, GTA, TGC, TGG), have been assigned to the frame according to the properties identified with the other trinucleotides, both in prokaryotes and eukaryotes.

3.3. Complementarity property

Recall of the DNA complementarity rule (Watson and Crick, 1953): (i) the DNA double helix consists of two nucleotide sequences s_1 and s_2 connected with the nucleotide pairing (hydrogen bonds) according to the complementarity rule $\mathscr{C}$: the nucleotide A (resp. C, G, T) in s_1 pairs with the complementary nucleotide $\mathscr{C}(A) = T$ (resp. $\mathscr{C}(C) = G$, $\mathscr{C}(G) = C$, $\mathscr{C}(T) = A$) in s_2; (ii) the two nucleotide sequences s_1 and s_2 run in opposite directions (called antiparallel) in the DNA double helix: the trinucleotide $w = l_1l_2l_3$, $l_1,l_2,l_3 \in \{A, C, G, T\}$, in s_1 pairs with the complementary trinucleotide $\mathscr{C}(w) = \mathscr{C}(l_3)\mathscr{C}(l_2)\mathscr{C}(l_1)$ in s_2.

Property 2: $\mathscr{C}(T_0) = T_0$, $\mathscr{C}(T_1) = T_2$ and $\mathscr{C}(T_2) = T_1$ (Table 3(b)). T_0 is self-complementary (11 trinucleotides of T_0 are complementary to 11 other trinucleotides of T_0) and, T_1 and T_2 are complementary to each other (the 21 trinucleotides of T_1 are complementary to the 21 trinucleotides of T_2). Note also that $\mathscr{C}(X_0) = X_0$, $\mathscr{C}(X_1) = X_2$ and $\mathscr{C}(X_2) = X_1$.

Biological consequence (detailed in the discussion): the two paired reading frames may simultaneously code for amino acids.

Table 3

(a)																						
T_0	AAA	AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC	TTT
T_1	AAG	ACA	ACG	ACT	AGC	AGG	ATA	ATG	CCA	CCC	CCG	GCG	GTG	TAG	TCA	TCC	TCG	TCT	TGC	TTA	TTG	
T_2	AGA	AGT	CAA	CAC	CAT	CCT	CGA	CGC	CGG	CGT	CTA	CTT	GCA	GCT	GGA	GGG	TAA	TAT	TGA	TGG	TGT	
(b)																						
T_0	AAA	AAC	AAT	ACC	ATC	CAG	CTC	GAA	GAC	GCC	GTA											
T_0	TTT	GTT	ATT	GGT	GAT	CTG	GAG	TTC	GTC	GGC	TAC											
T_1	AAG	ACA	ACG	ACT	AGC	AGG	ATA	ATG	CCA	CCC	CCG	GCG	GTG	TAG	TCA	TCC	TCG	TCT	TGC	TTA	TTG	
T_2	CTT	TGT	CGT	AGT	GCT	CCT	TAT	CAT	TGG	GGG	CGG	CGC	CAC	CTA	TGA	GGA	CGA	AGA	GCA	TAA	CAA	
(c)																						
X_0	AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC		
X_1	ACA	ATA	CCA	TCA	TTA	AGC	TCC	TGC	AAG	ACG	AGG	ATG	CCG	GCG	GTG	TAG	TCG	TTG	ACT	TCT		
X_2	CAA	TAA	CAC	CAT	TAT	GCA	CCT	GCT	AGA	CGA	GGA	TGA	CGC	CGG	TGG	AGT	CGT	TGT	CTA	CTT		

(d)

Alphabet	{R, Y}	{A, C, G, T}
Cardinal b of the alphabet	2	4
Cardinal b^3 of the trinucleotides	$2^3 = 8$	$4^3 = 64$
Number $(b^3 - b)/3$ of classes invariant by circular permutation	$(8-2)/3 = 2$	$(64-4)/3 = 20$
Number $3^{(b^3-b)/3}$ of potential (maximal) circular codes	$3^2 = 9$	$3^{20} = 3\,486\,784\,401$
Number of (maximal) circular codes	8	12 964 440
Probability of (maximal) circular codes	0.89	3.7×10^{-3}
Number of (maximal) complementary circular codes	2	528
Probability of (maximal) complementary circular codes	0.22	1.5×10^{-7}
Number of (maximal) complementary circular codes with two permutated circular codes (C^3 codes)	2	216
Probability of C^3 codes	0.22	6.2×10^{-8}

(a) List of the trinucleotides per frame in lexicographical order deduced from the Table 2(a–b). Three subsets of trinucleotides can be identified: $T_0 = X_0 \cup \{AAA, TTT\}$ in frame 0, $T_1 = X_1 \cup \{CCC\}$ in frame 1 and $T_2 = X_2 \cup \{GGG\}$ in frame 2.
(b) Complementarity property with the three subsets T_0, T_1 and T_2 of trinucleotides identified in Table 3(a).
(c) Circularity property with the three subsets X_0, X_1 and X_2 of trinucleotides identified in Table 3(a).
(d) Circular code statistics in the alphabets {R, Y} and {A, C, G, T}.

3.4. Circularity property

Definition of the trinucleotide circular permutation: the circular permutation $\mathcal{P}$ of the trinucleotide $w = l_1l_2l_3$, $l_1,l_2,l_3 \in \{A, C, G, T\}$, is the permutated trinucleotide $\mathcal{P}(w) = l_2l_3l_1$.

Property 3: $\mathcal{P}(X_0) = X_1$ and $\mathcal{P}(X_1) = X_2$ (Table 3(c)). X_0 generates X_1 by one circular permutation and X_2 by another circular permutation (one and two circular permutations with each trinucleotide of X_0 lead to the trinucleotides of X_1 and X_2, respectively).

Biological consequence (detailed in the discussion): the two subsets X_1 and X_2 can be deduced from X_0.

3.5. Circular code property

3.5.1. Definition of a circular code

Recall of a few notations: let $\mathcal{B}$ be a genetic alphabet, $\mathcal{B}_2 = \{R, Y\}$ and $\mathcal{B}_4 = \{A, C, G, T\}$. $\mathcal{B}^*$ denotes the words on $\mathcal{B}$ of finite length including the empty word of length 0. $\mathcal{B}^+$ denotes the words on $\mathcal{B}$ of finite length ≥ 1. Let w_1w_2 be the concatenation of the two words w_1 and w_2.

A subset X of $\mathcal{B}^+$ is a circular code if for all $n, m \geq 1$ and $x_1, x_2, ..., x_n \in X$, $y_1, y_2, ..., y_m \in X$ and $p \in \mathcal{B}^*$, $s \in \mathcal{B}^+$, the equalities $sx_2x_3...x_np = y_1y_2...y_m$ and $x_1 = ps$ imply $n = m$, $p = 1$ and $x_i = y_i$, $1 \leq i \leq n$ (Béal, 1993; Berstel and Perrin, 1985) (Fig. 2(a)). In other terms, every word on $\mathcal{B}$ 'written on a circle' has at most one factorization (decomposition) over X.

Remark: In the following, X will be a set of words of length three as a protein gene is a concatenation of trinucleotides.

3.5.2. Complete study of circular codes with trinucleotides on the alphabet $\mathcal{B}_2 = \{R, Y\}$

Such a complete study has not been presented so far and allows to introduce simply the differents concepts and properties of circular codes which are complex in the alphabet $\mathcal{B}_4 = \{A, C, G, T\}$ and cannot be analyzed by hand.

If b is the cardinal of the alphabet $\mathcal{B}$, then X contains at most b^3 trinucleotides (Table 3(d)). Therefore, on $\mathcal{B}_2$, X is a subset of {RRR, RRY, RYR, RYY, YRR, YRY, YYR, YYY}. There

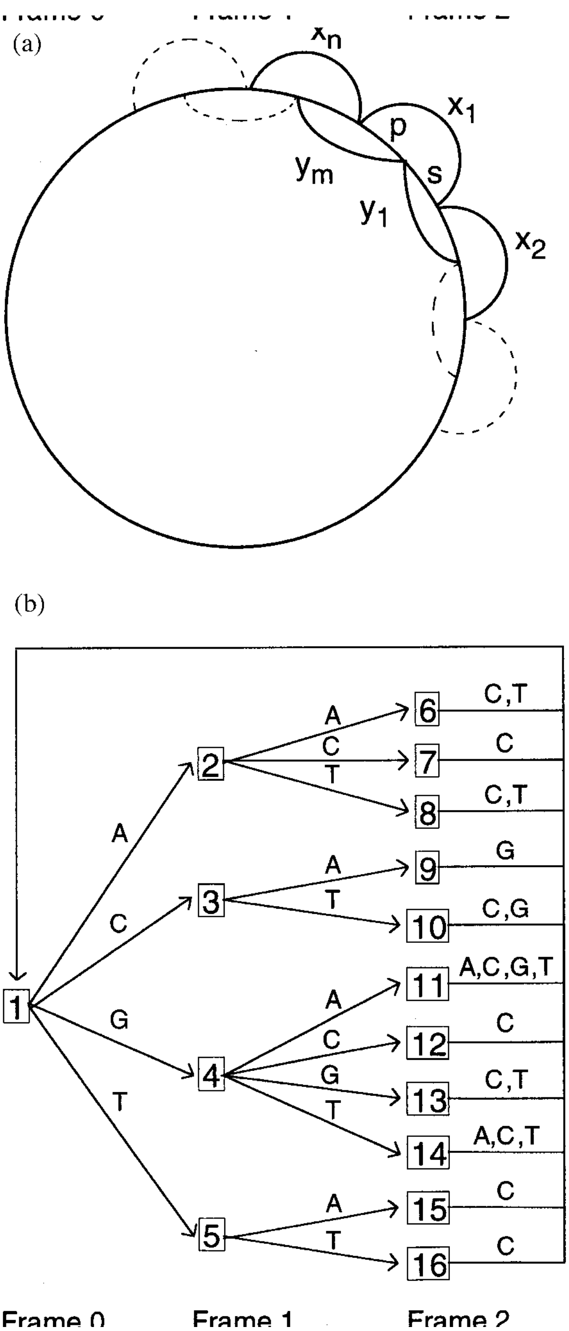


Fig. 2. (a) A representation of the definition of a circular code; (b) flower automaton $\mathcal{F}(X_0)$ associated with the circular code X_0.

are two obvious constraints so that X is a circular code:

(i) X cannot have the trinucleotides $w = lll$, $l \in \mathscr{B}$. For example, if X contains RRR then the word ...RRRRRR... has three factorizations over X: ...RRR, RRR..., ...R, RRR, RR... and ...RR, RRR, R... Therefore, X will be a subset of $\mathscr{B}'_2 =$ {RRY, RYR, RYY, YRR, YRY, YYR}. The cardinal of $\mathscr{B}'$ is $b^3 - b$, i.e. six for $\mathscr{B}'_2$.

(ii) X cannot have at the same time two trinucleotides deduced from each other by circular permutation. For example, if X contains RRY and RYR (RYR is one circular permutation of RRY) then the word ...RRYRRYRRY... has two factorizations over X: ...RRY, RRY, RRY... and ...R, RYR, RYR, RY... Therefore, by gathering the six trinucleotides of $\mathscr{B}'_2$ in two classes of three codons so that the three codons are deduced from each other by circular permutations {RRY, RYR, YRR} and {RYY, YYR, YRY}, X has at most one trinucleotide in each class. Therefore, X contains at most two trinucleotides and $3^2 = 9$ sets X are potential (maximal) circular codes. The number of classes invariant by circular permutation is $\mathrm{Card}(\mathscr{B}')/3 = (b^3 - b)/3$, i.e. two on $\mathscr{B}_2$ (Table 3(d)). The number of potential (maximal) circular codes is $3^{(b^3-b)/3}$, i.e. $3^2 = 9$ on $\mathscr{B}_2$ (Table 3(d)).

$X_a =$ {RRY, RYY} = RNY is a circular code (demonstrated in the introduction). X_a is a maximal (two trinucleotides) circular code and corresponds to the RNY model (Eigen and Schuster, 1978). X_a is self-complementary (complementary circular code), i.e. $\mathscr{C}(X_a) = X_a$, as RRY and RYY are complementary. Any subset of X_a is also a circular code but not maximal. For example, the subset RRY is a non-maximal circular code and corresponds to the RRY model (Crick et al., 1976). The two subsets $X_b = \mathscr{P}(X_a) =$ {RYR, YYR} and $X_c = \mathscr{P}(X_b) =$ {YRR, YRY} obtained by circular permutations of X_a are also maximal circular codes (identical proof). X_b and X_c are complementary to each other, i.e. $\mathscr{C}(X_b) = X_c$ and $\mathscr{C}(X_c) = X_b$, as RYR (resp. YYR) and YRY (resp. YRR) are complementary.

The previous results remain unchanged by substituting R by Y and reciprocally. Therefore, $X_d =$ {YRR, YYR} is a maximal complementary circular code ($\mathscr{C}(X_d) = X_d$) whose two subsets $X_e = \mathscr{P}(X_d) =$ {RRY, YRY} and $X_f = \mathscr{P}(X_e) =$ {RYR, RYY} obtained by circular permutations of X_d are also maximal circular codes and complementary to each other ($\mathscr{C}(X_e) = X_f$ and $\mathscr{C}(X_f) = X_e$).

The three remaining sets X are $X_g =$ {RYY, YRR} and $X_h = \mathscr{P}(X_g) = \mathscr{C}(X_g) =$ {RRY, YYR} which are circular codes and $X_i = \mathscr{P}(X_h) =$ {RYR, YRY} which is not a circular code as the word ...RYRYRYRYR... has two factorizations over X_i: ...RYR, YRY, RYR..., and ...R, YRY, RYR, YR...

In summary, on $\mathscr{B}_2$, eight among nine sets X are circular codes and two sets $X_a =$ {RRY, RYY} = RNY and $X_d =$ {YRR, YYR} = YNR are complementary Circular codes with two permutated Circular codes (called C^3 codes) (Table 3(d)).

3.5.3. Identification of circular codes with trinucleotides on the alphabet $\mathscr{B}_4 = \{A, C, G, T\}$

The study of circular codes on $\mathscr{B}_4$ is obviously more complex. For example, the search of a unique factorization needs the introduction of some classical definitions and results in coding theory, e.g. the flower automaton (Béal, 1993; Berstel and Perrin, 1985). The results obtained on {A, C, G, T} are new. From a biological point of view, no circular code (code without commas) on {A, C, G, T} has been identified with a theoretical, statistical or experimental approach. From a computational point of view, the classical circular codes determined by the methods of automatic construction of circular codes impose constraints on the choice of letters in the sites of the words, e.g. absence of a given letter in a given site of all words.

Property 4: the subset X_0 is a maximal (20 trinucleotides) circular code. The subsets X_1 and X_2 are also maximal circular codes. Remark: The property that X_0 is a circular code does not necessarily imply that X_1 and X_2 are also circular codes.

Proofs. (i) Proof of the maximum cardinal of a circular code. The 60 words of $\mathscr{B}'_4 =$ {AAA, ..., TTT}-{AAA, CCC, GGG, TTT} are gathered in $\mathrm{Card}(\mathscr{B}'_4)/3 = 20$ classes invariant by circular permutation (Table 3(d)). A circular code with words of length three on $\mathscr{B}_4$ has at most one word in each class and then contains at most 20 words.

(ii) Proof that X_0 is a circular code. As on $\mathscr{B}_4$ there are $3^{(b^3-b)/3}=3^{20}=3\,486\,784\,401$ potential circular codes (Table 3(d)), the use of some theorems in coding theory are necessary for the development of algorithms for automatically determining the circular codes. These basic theorems are given, the algorithms written in Pascal will be described elsewhere.

Definition 1: a deterministic finite state automaton $\mathscr{A}$ is said to be local if an integer n exists so that any two paths in $\mathscr{A}$ of the same length n and of the same associated word, have the same terminal state.

Lemma 1: If $\mathscr{A}$ is a strongly connected automaton, the two following properties are equivalent: $\mathscr{A}$ is local and $\mathscr{A}$ does not contain two cycles labelled with the same word (Béal, 1993).

Definition 2: the flower automaton $\mathscr{F}(X)$ associated with a subset X of $\mathscr{B}^+$ has a particular state (labelled 1 in Fig. 2(b)) and cycles issued from this state 1 and labelled by words of X.

Lemma 2: a finite subset X of $\mathscr{B}^+$ is a finite circular code if and only if the flower automaton $\mathscr{F}(X)$ is a local automaton (Béal, 1993).

Fig. 2(b) gives the flower automaton $\mathscr{F}(X_0)$ associated with X_0. To prove that 'X_0 is a circular code' is equivalent to prove that '$\mathscr{F}(X_0)$ is local', i.e. $\mathscr{F}(X_0)$ does not contain two cycles labelled with the same word. This proof can be done by hand (rather tedious and not explained here) or by algorithm. The algorithm developed identifies automatically all possible subsets X of $\mathscr{B}^+$ verifying the condition of circular code and allows statistics with circular codes (Section 3.7).

(iii) Proof that X_1 and X_2 are circular codes: similar to (ii) by constructing the flower automata $\mathscr{F}(X_1)$ and $\mathscr{F}(X_2)$ associated with X_1 and X_2, respectively.

In summary, the properties 1, 2 and 3 imply that X_0 is a complementary Circular code with two permutated Circular codes (called C^3 code).

3.6. Automatic frame determination property

Property 5: the length of the minimal window to automatically retrieve the frame with X_0 is equal to 13 nucleotides.

The problem to automatically determine the decomposition of a word into trinucleotides of X_0 is introduced with an example. Indeed, the unicity of such a decomposition is not obvious. For example, the word w′ = AGGTAATTACCA of length 12 can be decomposed into trinucleotides of X_0 in two ways: AG, GTA, ATT, ACC, A (GTA, ATT, ACC$\in X_0$, Table 3(a)) or A, GGT, AAT, TAC, CA (GGT, AAT, TAC$\in X_0$, Table 3(a)). In fact, the automatic frame determination property is a consequence of the circular code property. If a word is constructed by concatenating words of X_0 and if the frame of construction is lost, then the property of code assures that the frame can be retrieved according to a unique way. Such a decomposition is called in the following the reading frame of the word according to the code X_0.

The unicity of such a decomposition is proved by using the properties of the associated flower automaton $\mathscr{F}(X_0)$. As all words of X_0 have a length of three (Fig. 2(b)), the states of the automaton $\mathscr{F}(X_0)$ can be associated with frames, state 1 with frame 0, states 2–5 with frame 1 and states 6–16 with frame 2. If any letter of a word obtained by a concatenation of trinucleotides of X_0, can be associated with a unique state of the automaton $\mathscr{F}(X_0)$, then a frame can be deduced for this letter because the associated unique state of $\mathscr{F}(X_0)$ is related to a given frame. As a consequence, the word can be decomposed into trinucleotides of X_0: its reading frame according to X_0 is then retrieved. Then, the problem consists in identifying such a unique state for a letter of the word. Such a unicity is not obvious. In the previous example, the factor w′ = AGGTAATTACCA of length 12 can be attributed to two reading frames according to X_0: frame 1 (initial state 3 or 4 of w′ in $\mathscr{F}(X_0)$) or frame 2 (initial state 11 or 14 of w′ in $\mathscr{F}(X_0)$) (Fig. 3(a)).

In the case of a local automaton $\mathscr{A}$, definition 1 asserts that there exists n so that for a word of length $\geq$ n, all paths associated with this word have the same terminal state and thus, the reading frame of the word according to X_0 can be determined. For $\mathscr{A}=\mathscr{F}(X_0)$, n = 13 and $\mathscr{F}(X_0)$ is called a 13-local automaton. In other words, the length of the minimal window to retrieve always

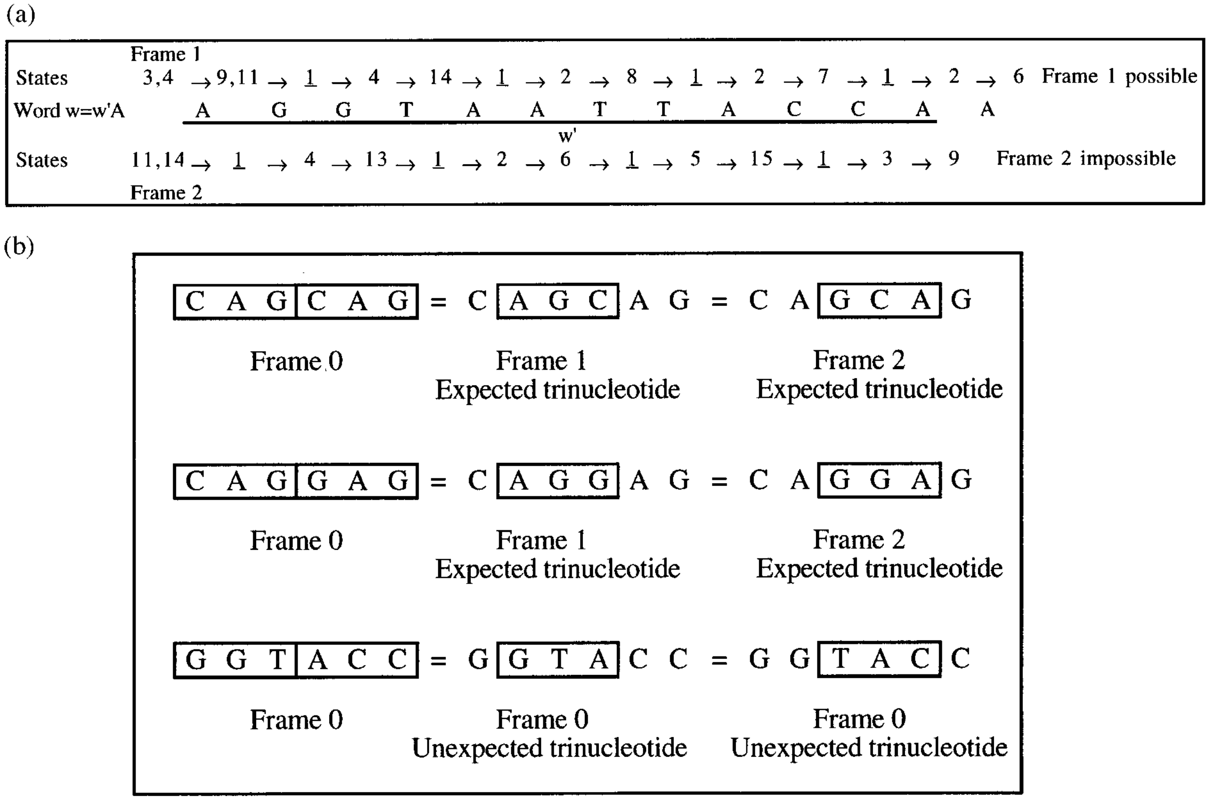

(a)
Frame 1
States 3,4 → 9,11 → 1 → 4 → 14 → 1 → 2 → 8 → 1 → 2 → 7 → 1 → 2 → 6 Frame 1 possible
Word w=w'A A G G T A A T T A C C A A
w'
States 11,14 → 1 → 4 → 13 → 1 → 2 → 6 → 1 → 5 → 15 → 1 → 3 → 9 Frame 2 impossible
Frame 2
(b)
CAG CAG = C AGC AG = CA GCA G
Frame 0
Frame 1
Expected trinucleotide
Frame 2
Expected trinucleotide
CAG GAG = C AGG AG = CA GGA G
Frame 0
Frame 1
Expected trinucleotide
Frame 2
Expected trinucleotide
GGT ACC = G GTA CC = GG TAC C
Frame 0
Frame 0
Unexpected trinucleotide
Frame 0
Unexpected trinucleotide

Fig. 3. (a) An example of the automatic frame determination of a word of length ≥ 13 with the flower automaton $\mathcal{F}(X_0)$ (Fig. 2(b)). The states 1 are associated with the position of the commas. The word w′ can be decomposed into trinucleotides of X_0 in two ways. The word w = w′A has a unique decomposition into trinucleotides of X_0; (b) examples of three concatenations of two trinucleotides of the C^3 code X_0 generating expected and unexpected trinucleotides in the shifted frames 1 and 2.

the frame is 13 letters with X_0. In the previous example, the first letter of the word w = w′A of length 13 is attributed to the unique frame, the frame 1, as there is no edge labelled A leaving the state 9 of $\mathcal{F}(X_0)$ (Fig. 2(b) and Fig. 3(a)). Then, the unique decomposition of w according to X_0 is AG, GTA, ATT, ACC, AA (Fig. 3(a)).

The length n of the minimal window to automatically retrieve the reading frame of a word according to X_0 can be determined by hand (rather tedious) or by algorithm testing all possible paths in the automaton. This computational approach allows statistics with the different C^3 codes (Section 3.8).

Remark: The lengths of the minimal windows to automatically retrieve the frames 1 and 2 with X_1 and X_2, respectively, are equal to 13 nucleotides (similar proof).

Biological consequence (detailed in the discussion): the code X_0 can retrieve automatically the frame 0 in any region of a protein gene (formed by a series of trinucleotides of X_0) without a start codon.

3.7. Rarity property

Property 6: the occurrence probability of $X_0 = 6.2 \times 10^{-8}$.

Statistics concerning circular codes with trinucleotides on the 4-letter alphabet {A, C, G, T} allow to determine the occurrence probability of the code X_0. There are $3^{20} = 3\,486\,784\,401$ potential circular codes (Section 3.5). The number of circular codes computed with an algorithm verifying the definition of circular code among the 3^{20} circular codes, is 12 964 440. The computed number of circular codes having the complementarity property (complementary circular codes), is 528. The computed number of complementary Circular codes with two permutated Circular codes (C^3 codes), is 216. Therefore, the probability to have a C^3 code, e.g. X_0, is $216/3^{20} = 6.2 \times 10^{-8}$. This very low probability explains the difficulty in identifying such C^3 codes by hand or by the classical methods of automatic construction of circular codes.

Table 3(d) summarizes the circular code statistics on the alphabets {R, Y} and {A, C, G, T}.

Biological consequence: the probability to observe the code X_0 in protein genes is very low and non-random.

3.8. Concatenation properties

The C^3 code X_0 identified in protein genes has some concatenation properties (flexibility) compared to the other C^3 codes:

Property 7: X_0 has the maximum length (13 nucleotides) of the minimal window among the 216 C^3 codes.

The lengths of minimal windows to retrieve the frame in the 216 C^3 codes are 5, 7, 9 and 13 nucleotides (data not shown), and equal to 13 nucleotides for X_0 (Section 3.6).

Property 8: X_0 has a high frequency (27.5%) of misplaced trinucleotides in the shifted frames among the 216 C^3 codes.

The circular permutation property implies that the concatenation of two trinucleotides of X_0 generates with a high probability a trinucleotide of X_1 in frame 1 and a trinucleotide of X_2 in frame 2. For X_0, this property is verified at 72.5%, i.e. there are 27.5% of misplaced trinucleotides in the shifted frames. The concatenation of two identical trinucleotides (process called duplication in biology) of X_0 leads obviously to the expected trinucleotides in the shifted frames (e.g. CAG$\in X_0$ and CAG$\in X_0$ generate AGC$\in X_1$ in frame 1 and GCA$\in X_2$ in frame 2) (first example in Fig. 3(b)). However, the probability of this type of concatenation is too low (1/20) to explain the circularity property. The concatenation of two different trinucleotides of X_0 may lead to the expected trinucleotides in the shifted frames (e.g. CAG$\in X_0$ and GAG$\in X_0$ generate AGG$\in X_1$ in frame 1 and GGA$\in X_2$ in frame 2) or not (GGT$\in X_0$ and ACC$\in X_0$ generate GTA$\notin X_1$ in frame 1 and TAC$\notin X_2$ in frame 2) (second and third examples in Fig. 3(b)). Due to the complementarity property, the frequency of trinucleotides of X_0 (resp. X_2) found in the shifted frame 1 is equal to the frequency of trinucleotides of X_0 (resp. X_1) found in the shifted frame 2. Therefore, the frequencies of misplaced trinucleotides (X_0 and X_2) found in the shifted frame 1 are equal to the frequencies of misplaced trinucleotides (X_0 and X_1) found in the shifted frame 2.

Fig. 4 shows the repartition function of the 216 C^3 codes according to the frequency of misplaced trinucleotides in frame 1 (or 2) generated by the concatenation of trinucleotides of a given C^3 code. This frequency is between 6.5% and 31%, and equal to 27.5% for X_0.

Property 9: X_0 has an occurrence of the four types of nucleotides in the three trinucleotide sites. This structure explains that the code X_0 cannot be generated by the classical methods of automatic construction of circular codes.

Biological consequence (detailed in the discussion): the code X_0 has evolutionary properties.

4. Discussion

4.1. Simulation of the protein coding genes

As the code X_0 leads to 72.5% well-placed trinucleotides in the shifted frames (Fig. 4), the protein genes could be simulated uniquely with the trinucleotides in frame 0 (codons). In order to verify this hypothesis, a population S of 200 simulated genes of 1000 base length is generated by an independent mixing of the 22 codons of T_0 with equiprobability. A sample of 200 000 bases allows precise and stable computations (i.e. there is no random fluctuations in the probability calculus of i-motifs: a sample having, for example, 50 simulated genes of 1000 base length leads to similar

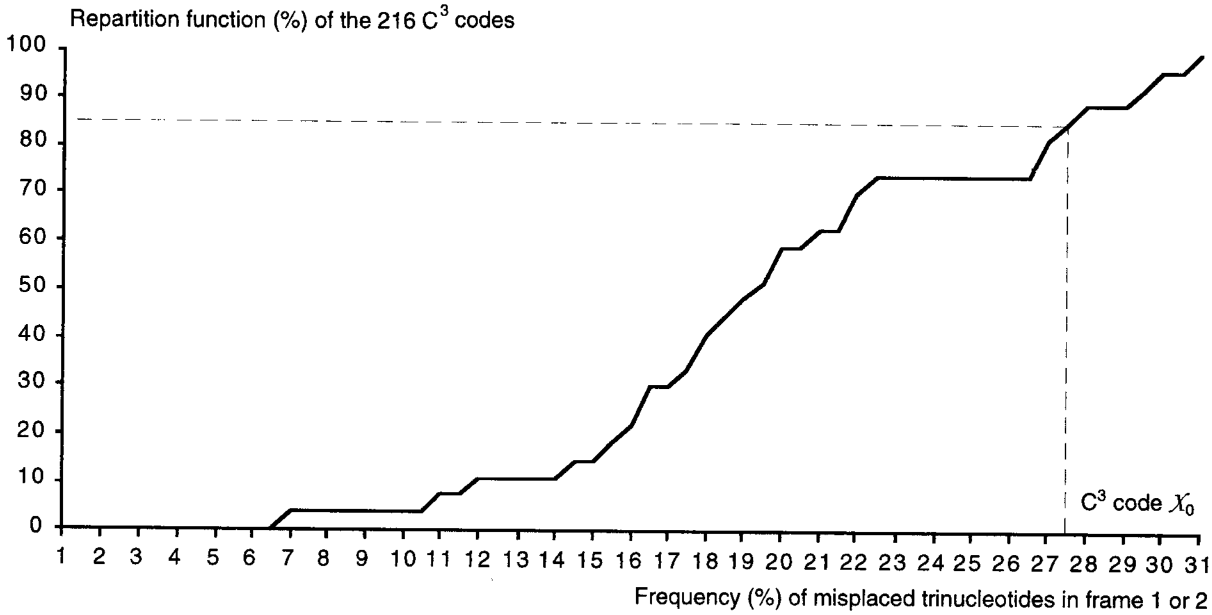


Fig. 4. Repartition function of the 216 C^3 codes according to the frequency of misplaced trinucleotides in frame 1 or 2 generated by the concatenation of trinucleotides of a given C^3 code. The frequency of misplaced trinucleotides for the C^3 code $X_0 = 27.5\%$.

results). The 12 288 autocorrelation functions are computed in this simulated population S as defined in Section 2.1 and classified in the three types of periodicities according to the method given in Section 3.1.

For each trinucleotide w′ and for each frame q in the simulated population S, Table 4, similarly to the Table 2(a),(b), gives the number of classifiable autocorrelation functions $w^p(N)_i w'$ implying the frame q for w′. This simple simulation retrieves the three types of periodicities correctly associated with the 12 288 i-motifs and thus, the two other subsets T_1 and T_2 of trinucleotides in the frames 1 and 2, respectively (except for three trinucleotides). Note: as this simulation uses only 22 trinucleotides, the maximal number of classifiable autocorrelation functions in the simulated genes (110) is obviously less than the one in the real protein genes (189).

4.2. Consequences on the purine/pyrimidine alphabet

The three subsets T_0, T_1 and T_2 classify the 64 A/C/G/T trinucleotides according to their preferential occurrence frame. Therefore, a preferential occurrence frame for the eight R/Y trinucleotides can be deduced from the frames of the 64 A/C/G/T trinucleotides by considering for each R/Y trinucleotide, the average frame of the eight frames associated with the eight A/C/G/T specified trinucleotides. Table 5 shows that the subset $\mathcal{Y}_0 = \{RRY, RYY\} = RNY$ occurs preferentially in frame 0, the subset $\mathcal{Y}_1 = \{RYR, YYR\}$, in frame 1, and the subset $\mathcal{Y}_2 = \{YRR, YRY\}$, in frame 2. RRY and RYY have the same number (six) of A/C/G/T trinucleotides in frame 0. The subset $\mathcal{Y}_0$ contains a few A/C/G/T trinucleotides in frames 1 (ACT^1, AGC^1) and 2 (AGT^2, GCT^2) and $\mathcal{Y}_1$ and $\mathcal{Y}_2$, a few A/C/G/T trinucleotides in frame 0 (CAG^0, CTG^0, GTA^0, TAC^0). The subset $\mathcal{Y}_0$ is a C^3 code on the alphabet {R, Y} (Section 3.5.2) and corresponds to the RNY codon model (Eigen and Schuster, 1978). There is no preferential frame for RRR and YYY.

These results also explain different previous works analyzing simultaneously the three frames (average frame) in protein genes with autocorrelation functions on the alphabet {R, Y}, in particular: (i) the periodicity 0 modulo 3 with the

Table 4
For each trinucleotide w′ and for each frame q in the simulated genes by an independent mixing of the 22 trinucleotides of T_0, the number of classifiable autocorrelation functions $w^p(N)_i w'$ implying the frame q for w′ is given

w′ in q = 0	Number	w′ in q = 1	Number	w′ in q = 2	Number
AAA	110	AAA	0	AAA	0
AAC	110	AAC	0	AAC	0
AAG	0	**AAG**	110	AAG	0
AAT	110	AAT	0	AAT	0
ACA	0	**ACA**	110	ACA	0
ACC	110	ACC	0	ACC	0
ACG	0	**ACG**	110	ACG	0
ACT	0	**ACT**	92	ACT	0
AGA	0	AGA	2	*AGA*	0
AGC	0	**AGC**	97	AGC	0
AGG	0	**AGG**	110	AGG	0
AGT	0	AGT	0	**AGT**	82
ATA	0	**ATA**	110	ATA	0
ATC	110	ATC	0	ATC	0
ATG	0	**ATG**	110	ATG	0
ATT	110	ATT	0	ATT	0
CAA	0	CAA	0	**CAA**	110
CAC	0	CAC	0	**CAC**	110
CAG	110	CAG	0	CAG	0
CAT	0	CAT	0	**CAT**	110
CCA	0	**CCA**	25	CCA	0
CCC	0	**CCC**	110	CCC	0
CCG	0	**CCG**	110	CCG	0
CCT	0	CCT	0	**CCT**	110
CGA	0	CGA	0	**CGA**	110
CGC	0	CGC	0	**CGC**	110
CGG	0	CGG	0	**CGG**	110
CGT	0	CGT	0	**CGT**	110
CTA	0	CTA	0	**CTA**	110
CTC	110	CTC	0	CTC	0
CTG	110	CTG	0	CTG	0
CTT	0	CTT	0	**CTT**	110
GAA	110	GAA	0	GAA	0
GAC	110	GAC	0	GAC	0
GAG	110	GAG	0	GAG	0
GAT	110	GAT	0	GAT	0
GCA	0	GCA	86	*GCA*	0
GCC	110	GCC	0	GCC	0
GCG	0	**GCG**	110	GCG	0
GCT	0	GCT	0	**GCT**	90
GGA	0	GGA	0	**GGA**	110
GGC	110	GGC	0	GGC	0
GGG	0	GGG	0	**GGG**	110
GGT	110	GGT	0	GGT	0
GTA	110	GTA	0	GTA	0
GTC	110	GTC	0	GTC	0
GTG	0	**GTG**	110	GTG	0
GTT	110	GTT	0	GTT	0
TAA	0	TAA	0	**TAA**	110
TAC	110	TAC	0	TAC	0
TAG	0	**TAG**	110	TAG	0
TAT	0	TAT	0	**TAT**	110
TCA	0	**TCA**	110	TCA	0
TCC	0	**TCC**	110	TCC	0
TCG	0	**TCG**	110	TCG	0
TCT	0	**TCT**	18	TCT	0
TGA	0	TGA	0	**TGA**	110
TGC	0	*TGC*	0	TGC	86
TGG	0	TGG	0	**TGG**	43
TGT	0	TGT	0	**TGT**	110
TTA	0	**TTA**	110	TTA	0
TTC	110	TTC	0	TTC	0
TTG	0	**TTG**	110	TTG	0
TTT	110	TTT	0	TTT	0

The trinucleotides in bold have a preferential occurrence frame. The three trinucleotides in italics (AGA, GCA, TGC) are reclassified according to the properties of the C^3 code X_0.

autocorrelation function $YRY(N)_iYRY$ (Fig. 5(a), Arquès and Michel, 1987, 1994) as the trinucleotides YRY occurring preferentially in frame 2 generate a number multiple of three of bases between them; and (ii) the absence of the periodicity 0 modulo 3 with the autocorrelation function $RRR(N)_iRRR$ (Fig. 5(b), Arquès and Michel, 1993) as RRR does not occur in a preferential frame.

4.3. Consequences on the genetic code

The codon subset T_0 codes for 13 amino acids (AA): Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Lys, Phe, Thr, Tyr and Val (Table 6). As the subset T_0 has 22 codons, several codons of T_0 code for the same AA. Almost all classes of AA are coded by T_0: the simplest (Gly), aliphatic (Ala, Ile, Leu, Val), hydroxyl (Thr), acidic (Asp, Glu), amide (Asn, Gln), basic (Lys) and aromatic (Phe, Tyr) (two classes are not coded by T_0: the sulfur-containing AA: Cys and Met, and the cyclic AA: Pro). The 12 AA coded by X_0 are equally represented in the two classes of aminoacyl-tRNA synthetases with a class 1 associated with Gln, Glu, Ile, Leu, Tyr and Val, and a class 2, with Ala, Asn, Asp, Gly, Phe and Thr (rewieved in Schimmel et al., 1993; Hartman, 1995; Saks and Sampson, 1995).

4.4. Consequences on the amino acid frequencies in proteins

There are seven amino acids (AA) which are not coded by the codon subset T_0: Arg, Cys, His, Met, Pro, Ser and Trp (Table 6). Therefore, these seven AA should have the lowest frequencies in proteins. In order to verify this consequence of a code in protein genes, the frequencies of the 20 AA are computed in 9510 prokaryotic proteins (3 044 028 AA) and 20 673 eukaryotic proteins (7 521 044 AA). They are obtained from the protein data base SWISS-PROT (released 29, June 1994). Then, these observed AA frequencies are compared with their expected AA frequencies (number of codons coding an AA divided by the total number of non-stop codons, i.e. 61). Table 7 shows that except for Met, the six other AA not coded by T_0, Arg, Cys, His, Pro, Ser and Trp, have the lowest observed/expected frequency ratios (<0.8) in the proteins of both prokaryotes and eukaryotes. For Arg, a difference between the observed frequency and the frequency expected from the universal genetic code has already been mentioned (Jukes et al., 1975). Met is a particular case as the codon coding for Met is also the start codon for establishing the frame 0 (reading frame) in actual genes.

In summary, both in prokaryotes and eukaryotes, there is a strong correlation between the usage of the codons of T_0 in protein genes and the amino acid frequencies in proteins.

4.5. Biological consequences which could be associated with the properties of the circular code

The identification of the same subset X_0 of 20 trinucleotides occurring preferentially (i.e. in comparison with X_1 and X_2) in frame 0 (reading frame) of two different taxonomic protein gene populations (prokaryotes and eukaryotes), suggests an evolution of protein genes according to two processes: a construction process of primitive protein genes followed by an evolutionnary process transforming the primitive protein genes into actual protein genes. In the model currently under investigation, the primitive protein genes are constructed of trinucleotides of X_0 (independent mixing with equiprobability 1/20), i.e. X_1 and X_2 do not occur in frame 0. Two simple types of evolutionnary processes may generate X_1 and X_2 in frame 0: (i) substitutions in the three sites of trinucleotides of X_0 or (ii) insertions and deletions of nucleotides (process named RNA editing) in the trinucleotides of X_0 (X_1 and X_2 are obtained in this case by circular permutation). There are several constraints associated with these random evolutionary processes, in particular the frequencies of X_1 and X_2 in frame 0 must remain less than the frequency of X_0 in frame 0. Such a computational model based on an evolution of the number of trinucleotides (from 20 to 64), i.e. on an evolution of the number of amino acids coded by these trinucleotides (from 12 to 20 by using the hypothesis of the actual genetic code), may be related to several evolutionary biological models, in particular to the co-evolution of the aminoacyl-tRNA synthetases and the genetic code (Wetzel, 1995).

Almost all actual protein genes (with X_0, X_1 and X_2 in frame 0) begin with the ATG start codon which does not belong to X_0 ($ATG \in X_1$). However, the scanning mechanism (Kozak, 1978) for initiation of translation in eukaryotes (40S ribosomal subunit carrying Met-tRNA$_i^{met}$ and various initiation factors) is based on the consensus sequence GCCGCCRCCATG (R = A or G; Kozak, 1989). Surprisingly, the three trinucleotides preceding the ATG start codon belong to X_0 ($ACC, GCC \in X_0$). The occurrence probability of such a series is equal to $1/(20)^3 \approx 10^{-4}$. Therefore, this motif of nine base length could have been the translation initiation signal in primitive protein genes (with only X_0 in frame 0). The importance of this motif has been demonstrated by site-directed mutagenesis experiments (e.g. Kozak, 1986a) and confirmed by the discovery of a type of thalassemia in which a mutation in the sequence (the purine base R three bases before ATG changed in C) drastically impairs translation initiation of α-globin (Morlé et al., 1985). Non-ATG start codons exist in actual protein genes but they are weakly recognized like the ATG start codons in a mutated consensus sequence. Surprisingly, some non-ATG start codons identified, belong to X_0, e.g. CTG in the first exon of c-*myc*

Table 5
The eight R/Y trinucleotides (R = purine = A or G, Y = pyrimidine = C or T) are associated with the 64 A/C/G/T trinucleotides by considering their frame (T_0, T_1, T_2)

RRR	RRY	RYR	RYY	YRR	YRY	YYR	YYY
AAA^0	AAC^0	ACA^1	ACC^0	CAA^2	CAC^2	CCA^1	CCC^1
AAG^1	AAT^0	ACG^1	ACT^1	CAG^0	CAT^2	CCG^1	CCT^2
AGA^2	AGC^1	ATA^1	ATC^0	CGA^2	CGC^2	CTA^2	CTC^0
AGG^1	AGT^2	ATG^1	ATT^0	CGG^2	CGT^2	CTG^0	CTT^2
GAA^0	GAC^0	GCA^2	GCC^0	TAA^2	TAC^0	TCA^1	TCC^1
GAG^0	GAT^0	GCG^1	GCT^2	TAG^1	TAT^2	TCG^1	TCT^1
GGA^2	GGC^0	GTA^0	GTC^0	TGA^2	TGC^1	TTA^1	TTC^0
GGG^2	GGT^0	GTG^1	GTT^0	TGG^2	TGT^2	TTG^1	TTT^0
0, 1, 2	0	1	0	2	2	1	0, 1, 2

The last row gives the preferential occurrence frame for the R/Y trinucleotides, e.g. RRR is in the three frames (three A/C/G/T trinucleotides in frame 0, two in frame 1, three in frame 2), YRY is in frame 2 (one A/C/G/T trinucleotide in frame 0, one in frame 1, six in frame 2).

(Geballe and Morris, 1994), ATT in several mitochondrial protein genes of the gastropod mollusc *Cepea nemoralis* (Terrett et al., 1996) and in the mitochondrial protein genes NADH2 of *Homo* and chimpanzees (Arnason et al., 1996). Furthermore, the first ATG codon is not always used for translation initiation which can begin at downstream ATG codons in the following cases: (i) the length between the cap and the first ATG codon is less than nine bases (e.g. Strubin et al., 1986); (ii) the first ATG codon occurs in a mutated consensus sequence (Kozak, 1986b); and (iii) the first ATG codon is associated with alternative promoters and/or splice sites for regulating downstream ATG codons, e.g. in protooncogenes, growth factor genes and homeobox genes (Kozak, 1989; Geballe and Morris, 1994). Finally, the lengths of the consensus motif GCCGCCRCC and of the motif between the cap and the first ATG codon could be considered according to the window of nucleotides automatically retrieving the frame 0 with the identified C^3 code X_0 (see also below).

A protein gene formed by trinucleotides of X_0 has the property to automatically retrieve the frame 0 in any region of the gene without a start codon (X_0 is a circular code by definition). It would be interesting to know if a trace of this property may exist in the actual protein genes formed by a preferential occurrence of trinucleotides of X_0 (i.e. a protein gene formed by trinucleotides of X_0, X_1 and X_2 with lower frequencies for X_1 and X_2). Surprisingly, such a property is observed with a biological process, called translational frameshifting, which is involved in: (i) producing a translational fusion for the morphogenesis of the viral particle, e.g. the retroviral *gag* and *pol* genes (Farabaugh et al., 1993); e.g. the retrotransposon *TYA* and *TYB* genes (Belcourt and Farabaugh, 1990); (ii) regulating gene expression, e.g. the *Escherichia coli prfB* gene (e.g. Craigen and Caskey, 1986) and the rat ODCase antizyme gene (e.g. Matsufuji et al., 1995); and (iii) coding two proteins with a common N-terminal region and a different C-terminal region, e.g. the *Escherichia coli dnaX* gene (e.g. Tsuchihashi and Brown, 1992), and the prokaryotic insertion sequences, e.g. IS*1* (Machida et al., 1984), IS*150*: (Vögele et al., 1991). This frameshifting process allows translation to continue through a stop codon in frame 0, to change the frame and to use two frames. It is found in genes associated with different functions and from a variety of organisms: bacteria, lower eukaryotes (yeasts, plants), higher eukaryotes (animal) and viruses (e.g. retroviruses, bacteriophages, plant viruses, etc.). There are mainly two translational frameshifting processes: hopping and slipping.

Hopping can be defined as a translational shift ≥ 2 bases in the $5'-3'$ direction (downstream direction) or in the 3′–5′ direction (upstream direction). It requires a take-off trinucleotide and a landing trinucleotide. O'Connor et al. (1989) reported two examples, both decoded as Val, where the take-off trinucleotide (underlined once) belongs to X_1 and the landing trinucleotide (underlined twice), to X_0: GTGTA (GTG$\in X_1$, TGT$\in X_2$, GTA$\in X_0$) and GTGTAAGTT (GTG$\in X_1$, TGT$\in X_2$, GTA$\in X_0$, TAA$\in X_2$, AAG$\in X_1$, AGT$\in X_2$, GTT$\in X_0$). In the first example, the take-off trinucleotide GTG and TGT are not translated and belong to X_1 and X_2, respectively. The translated trinucleotide GTA belongs to X_0. The second example appears to be an extension of the rule applied previously. If the first landing trinucleotide belonging to X_0 (GTA) is not chosen, then the second landing trinucleotide considered is the next downstream trinucleotide belonging to X_0 (GTT).

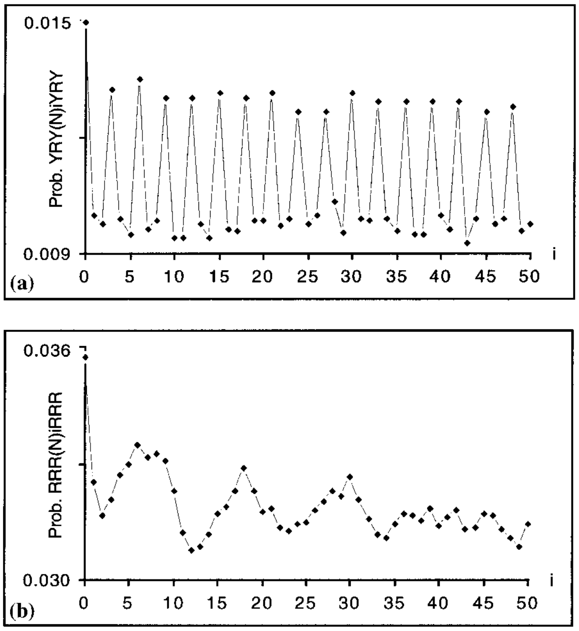


Fig. 5. (a) periodicity 0 modulo 3 in the eukaryotic protein coding genes with the autocorrelation function YRY(N)$_i$YRY without considering the frame. The horizontal axis represents the number i, i$\in$[0, 50], of any bases N between YRY and YRY. The vertical axis represents the autocorrelation function YRY(N)$_i$YRY; (b) absence of the periodicity 0 modulo 3 in the eukaryotic protein coding genes with the autocorrelation function RRR(N)$_i$RRR without considering the frame. The horizontal axis represents the number i, i$\in$[0, 50], of any bases N between RRR and RRR. The vertical axis represents the autocorrelation function RRR(N)$_i$RRR.

Slipping can be defined as a translational shift of one base either in the 3′–5′ direction (-1 frameshift) or in the 5′–3′ direction ($+1$ frameshift). Several trinucleotides which may direct -1 frameshift, have been reported, e.g. AAG (*E. coli dnaX* gene: Tsuchihashi and Brown, 1992; Lindsley and Gallant, 1993; the prokaryotic insertion sequence IS*150*: Vögele et al., 1991), AGC (Bruce et al., 1986), CCG (Dayhuff et al., 1986) and TTA (Jacks et al., 1988; ten Dam et al., 1990). As these trinucleotides belong to X_1 and as X_1 is deduced by one circular permutation of X_0, the theoretical frame 0 is retrieved with -1 frameshift, as expected with the examples mentioned. In contrast, other trinucleotides, e.g. AGT (Farabaugh et al., 1993), CTT (Weiss et al., 1987) and TGA (Craigen and Caskey, 1986; Curran, 1993), may provoke $+1$ frameshift. As these trinucleotides belong to X_2 and as X_2 is deduced by two circular permutations of X_0, the theoretical frame 0 is retrieved with $+1$ frameshift, as expected with the examples given. The trinucleotides AAA (Weiss et al., 1990), GGG (Weiss et al., 1990) and TTT (Fox and Weiss-Brummer, 1980) may induce both -1 and $+1$ frameshifts. Therefore, the trinucleotides belonging to T_0, T_1 and T_2 exclusively (AAA, TTT$\in T_0$, GGG$\in T_2$), may lead to a shift of one base in both directions.

The rules of the take-off and landing trinucleotides in the hopping process and of the frameshift trinucleotides in the slipping process could be analyzed according to the subsets X_0, X_1 and X_2 of trinucleotides in order to identify the frameshift rules encoded in the DNA sequence. As X_0, X_1 and X_2 are associated with the frame 0, 1 and 2, respectively, the type of translational frameshifting at the frameshift site could easily be determined for example, a gene in a shifted frame 1 or 2 retrieving the frame 0, a gene in frame 0 keeping the frame 0 and a gene in frame 0 using a shifted frame 1 or 2. Indeed, a feature classically used for determining the frame 0 after a frameshift site, is the frame among the three possible ones having the highest number of non-stop codons before a stop codon. A set of 20 trinucleotides would be more significant than a stop codon for discriminating the different frames.

Table 6
The subset T_0 (bold) codes for 13 amino acids in the universal genetic code: Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Lys, Phe, Thr, Tyr and Val

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
TTT	Phe, F Phenylalanine	$\underline{\text{TCT}}$	Ser, S Serine	$\underline{\underline{\text{TAT}}}$	Tyr, Y Tyrosine	$\underline{\underline{\text{TGT}}}$	Cys, C Cysteine
TTC	Phe, F Phenylalanine	$\underline{\text{TCC}}$	Ser, S Serine	**TAC**	Tyr, Y Tyrosine	$\underline{\text{TGC}}$	Cys, C Cysteine
$\underline{\text{TTA}}$	Leu, L Leucine	$\underline{\text{TCA}}$	Ser, S Serine	$\underline{\underline{\text{TAA}}}$	Stop codon ochre	$\underline{\underline{\text{TGA}}}$	Stop codon opal
$\underline{\text{TTG}}$	Leu, L Leucine	$\underline{\text{TCG}}$	Ser, S Serine	$\underline{\text{TAG}}$	Stop codon amber	$\underline{\underline{\text{TGG}}}$	Trp, W Tryptophan
$\underline{\underline{\text{CTT}}}$	Leu, L Leucine	$\underline{\underline{\text{CCT}}}$	Pro, P Proline	$\underline{\underline{\text{CAT}}}$	His, H Histidine	$\underline{\underline{\text{CGT}}}$	Arg, R Arginine
CTC	Leu, L Leucine	$\underline{\text{CCC}}$	Pro, P Proline	$\underline{\underline{\text{CAC}}}$	His, H Histidine	$\underline{\underline{\text{CGC}}}$	Arg, R Arginine
$\underline{\underline{\text{CTA}}}$	Leu, L Leucine	$\underline{\text{CCA}}$	Pro, P Proline	$\underline{\underline{\text{CAA}}}$	Gln, Q Glutamine	$\underline{\underline{\text{CGA}}}$	Arg, R Arginine
CTG	Leu, L Leucine	$\underline{\text{CCG}}$	Pro, P Proline	**CAG**	Gln, Q Glutamine	$\underline{\underline{\text{CGG}}}$	Arg, R Arginine
ATT	Ile, I Isoleucine	$\underline{\text{ACT}}$	Thr, T Threonine	**AAT**	Asn, N Asparagine	$\underline{\underline{\text{AGT}}}$	Ser, S Serine
ATC	Ile, I Isoleucine	**ACC**	Thr, T Threonine	**AAC**	Asn, N Asparagine	$\underline{\text{AGC}}$	Ser, S Serine
$\underline{\text{ATA}}$	Ile, I Isoleucine	$\underline{\text{ACA}}$	Thr, T Threonine	**AAA**	Lys, K Lysine	$\underline{\underline{\text{AGA}}}$	Arg, R Arginine
$\underline{\text{ATG}}$	Met, M Methionine	$\underline{\text{ACG}}$	Thr, T Threonine	$\underline{\text{AAG}}$	Lys, K Lysine	$\underline{\text{AGG}}$	Arg, R Arginine
GTT	Val, V Valine	$\underline{\underline{\text{GCT}}}$	Ala, A Alanine	**GAT**	Asp, D Aspartic acid	**GGT**	Gly, G Glycine
GTC	Val, V Valine	**GCC**	Ala, A Alanine	**GAC**	Asp, D Aspartic acid	**GGC**	Gly, G Glycine
GTA	Val, V Valine	$\underline{\underline{\text{GCA}}}$	Ala, A Alanine	**GAA**	Glu, E Glutamic acid	$\underline{\underline{\text{GGA}}}$	Gly, G Glycine
$\underline{\text{GTG}}$	Val, V Valine	$\underline{\text{GCG}}$	Ala, A Alanine	**GAG**	Glu, E Glutamic acid	$\underline{\underline{\text{GGG}}}$	Gly, G Glycine

Seven amino acids are not coded by T_0: Arg, Cys, His, Met, Pro, Ser and Trp. The subset T_1 (underlined once) is associated with 11 amino acids: Ala, Arg, Cys, Ile, Leu, Lys, Met, Pro, Ser, Thr and Val. The subset T_2 (underlined twice) is associated with 11 amino acids: Ala, Arg, Cys, Gln, Gly, His, Leu, Pro, Ser, Trp and Tyr.

The translational frameshifting in the actual protein genes has been considered in the previous section according to the words of the circular code X_0. In this section, this frameshifting process is analyzed with respect to the window of 13 nucleotides retrieving automatically the frame 0 of the circular code X_0.

In the hopping process, the window of X_0 could be related to the sequence between the take-off and landing trinucleotides.

In the -1 slipping process of eukaryotes and viruses, a secondary structure, usually a pseudoknot, is often associated with frameshifting (ten Dam et al., 1990). Pseudoknots occur on average six nucleotides downstream $(+6)$ of the frameshift trinucleotides (ten Dam et al., 1990). This distance is critical as the insertion or deletion of two nucleotides between frameshift trinucleotides and pseudoknots eliminate frameshifting (Brierley et al., 1992). Furthermore, the four nucleotides upstream (-4) of the frameshift trinucleotides can modify the rate of frameshifting (Brierley et al., 1992). Therefore, these biological observations identify a sequence of 13 $(4+3+6)$

Table 7
Arg, Cys, His, Pro, Ser and Trp have the lowest observed/expected frequency ratios (<0.8) in the proteins of both prokaryotes (9510 sequences, 3 044 028 amino acids) and eukaryotes (20 673 sequences, 7 521 044 amino acids) as expected with the usage of the codons of T_0 in protein genes

Amino acid (number of codons)	Expected frequency (% rounded)	Observed number in proteins		Observed frequency (%) in proteins		Observed frequency/Expected frequency	
		Prokaryotes	Eukaryotes	Prokaryotes	Eukaryotes	Prokaryotes	Eukaryotes
Ala, A, Alanine (4/61)	6.56	287 360	535 229	9.44	7.12	1.44	1.08
Arg, R, Arginine (6/61)	9.84	164 498	377 123	5.40	5.01	0.55	0.51
Asn, N, Asparagine (2/61)	3.28	127 874	337 816	4.20	4.49	1.28	1.37
Asp, D, Aspartic acid (2/61)	3.28	170 363	392 502	5.60	5.22	1.71	1.59
Cys, C, Cysteine (2/61)	3.28	30 213	151 439	0.99	2.01	0.30	0.61
Gln, Q, Glutamine (2/61)	3.28	119 579	312 059	3.93	4.15	1.20	1.26
Glu, E, Glutamic acid (2/61)	3.28	190 544	488 770	6.26	6.50	1.91	1.98
Gly, G, Glycine (4/61)	6.56	234 665	515 987	7.71	6.86	1.18	1.05
His, H, Histidine (2/61)	3.28	63 907	175 381	2.10	2.33	0.64	0.71
Ile, I, Isoleucine (3/61)	4.92	180 442	402 872	5.93	5.36	1.20	1.09
Leu, L, Leucine (6/61)	9.84	288 009	690 767	9.46	9.18	0.96	0.93
Lys, K, Lysine (2/61)	3.28	158 728	466 467	5.21	6.20	1.59	1.89
Met, M, Methionine (1/61)	1.64	74 134	173 337	2.44	2.30	1.48	1.41
Phe, F, Phenylalanine (2/61)	3.28	114 989	311 375	3.78	4.14	1.15	1.26
Pro, P, Proline (4/61)	6.56	132 643	390 457	4.36	5.19	0.66	0.79
Ser, S, Serine (6/61)	9.84	183 680	568 759	6.03	7.56	0.61	0.77
Thr, T, Threonine (4/61)	6.56	173 058	424 350	5.69	5.64	0.87	0.86
Trp, W, Tryptophan (1/61)	1.64	38 564	92 793	1.27	1.23	0.77	0.75
Tyr, Y, Tyrosine (2/61)	3.28	94 076	238 524	3.09	3.17	0.94	0.97
Val, V, Valine (4/61)	6.56	216 702	475 037	7.12	6.32	1.09	0.96

nucleotides involved in frameshifting. The window of X_0 could be connected with this sequence.

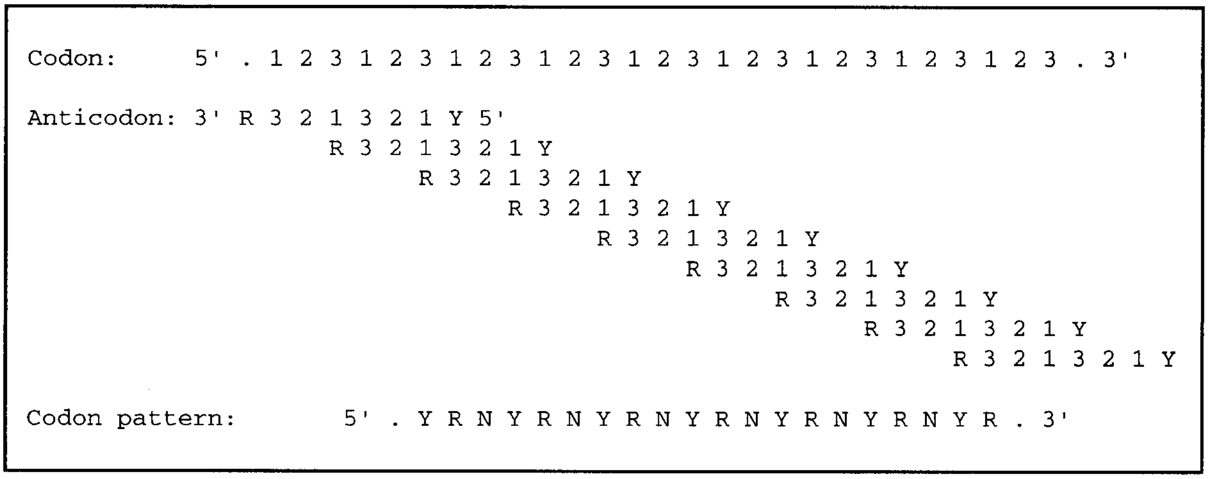
In the −1 slipping process of prokaryotes, a secondary structure, usually a hairpin, and a Shine–Dalgarno site (Shine and Dalgarno, 1974) are often involved in frameshifting (Vögele et al., 1991; Larsen et al., 1994). The Shine–Dalgarno sites occur on average 14 nucleotides upstream (−14) of the frameshift trinucleotides (Larsen et al., 1994), and the hairpin, seven nucleotides downstream (+7) of the frameshift trinucleotides (Vögele et al., 1991). If the spacing between the Shine–Dalgarno site and the frameshift trinucleotide is reduced to seven nucleotides then +1 frameshift may occur (Larsen et al., 1994). Notes: the Shine–Dalgarno sites are observed six nucleotides upstream (−6) of the frameshift trinucleotides in the +1 slipping process of prokaryotes (Weiss et al., 1988); a potential pseudoknot occurs five nucleotides downstream (+5) of the frameshift trinucleotide in the +1 slipping process of the ornithine decarboxylase antizyme gene of eukaryotes (Matsufuji et al., 1995). The window of X_0 could be involved in these different sequences of critical lengths (Larsen et al., 1994).

The window of X_0 can also be considered according to the codon–anticodon pairing process as X_0 is a complementary circular code. The codon–anticodon pairing is classically based on three nucleotides. However, different models of extended codon–anticodon pairing have been proposed.

In the primitive translation RRY model proposed by Crick et al. (1976), the primitive tRNAs have an anticodon involving seven conserved nucleotides (two nucleotides on both sides of the anticodon in actual tRNAs; Barrell and Clark, 1974) in two alternative stacking configurations. The aminoacyl-tRNA adopts the 5′ stacked configuration with the five base anticodon $YYa_1a_2a_3$, a_1 being the wobble base (Woese, 1970). On the other hand, the peptidyl-tRNA adopts the 3′ stacked configuration with the five base anticodon $a_1a_2a_3RN$ (Fuller and Hodgson, 1967). A stacked configuration pairs with a five base codon. Note: a primitive codon–anticodon pairing involving only three bases such as an actual one without ribosome is not stable enough for codon translation (Grosjean et al., 1978). The flip mechanism (detailed below for the RNY model) between these two stacked configurations leads to primitive protein genes consisting of a series of RRY codons (RRY model) (Crick et al., 1976). In the primitive translation RNY model proposed by Eigen and Schuster (1978), these biological concepts are considered with an anticodon involving only five conserved nucleotides (one nucleotide on both sides of the anticodon in actual tRNAs). Therefore, the aminoacyl-tRNA with the 5′ stacked configuration has the four base anticodon $Ya_1a_2a_3$ ($3'a_3a_2a_1Y$) pairing with the four base codon $c_1c_2c_3R$. The peptidyl-tRNA with the 3′ stacked configuration has the four base anticodon $a_1a_2a_3R$ ($3'Ra_3a_2a_1$) pairing with the four base codon $Yc_1c_2c_3$ (Fig. 6). The flip mechanism between these two stacked configurations keeps the three anticodon bases $3'a_3a_2a_1$ paired with the three codon bases $c_1c_2c_3$ for codon translation. The sequence $3'Ra_3a_2a_1a_3a_2a_1Y$ of the peptidyl and aminoacyl tRNAs paired with mRNA during the flip mechanism leads to primitive protein coding genes constituted of a series of RNY codons (Fig. 6). Thus, the window $3'Ra_3a_2a_1a_3a_2a_1Y$ of seven nucleotides with a shift of three bases according to a flip mechanism allows to generate the C^3 code RNY on the 2-letter alphabet {R, Y}. A four base codon–anticodon pairing is also observed with the actual shifty tRNAs which are particular tRNAs involved in the translational frameshift sites (Riddle and Carbon, 1973; Beremand and Blumenthal, 1979; Tuohy et al., 1992). A biological model, such as the flip mechanism, for generating the C^3 code X_0 on the 4-letter alphabet {A, C, G, T} remains to be found.

Finally, the window of X_0 may be related to the acceptor stem nucleotides for the correct recognition of the amino acid that the cognate aminoacyl-tRNA synthetase attaches to the tRNA 3′-terminus (Schimmel et al., 1993; Hartman, 1995; Saks and Sampson, 1995).

As the circular code X_0 is self-complementary, the two paired frames 0 in the two DNA double helix may simultaneously code for amino acids without the start codon (Fig. 7). Furthermore, as X_1 and X_2 are also circular codes and complemen-



```
Codon:      5' . 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 . 3'

Anticodon:  3' R 3 2 1 3 2 1 Y 5'
                     R 3 2 1 3 2 1 Y
                           R 3 2 1 3 2 1 Y
                                 R 3 2 1 3 2 1 Y
                                       R 3 2 1 3 2 1 Y
                                             R 3 2 1 3 2 1 Y
                                                   R 3 2 1 3 2 1 Y
                                                         R 3 2 1 3 2 1 Y
                                                               R 3 2 1 3 2 1 Y

Codon pattern:        5' . Y R N Y R N Y R N Y R N Y R N Y R N Y R . 3'
```

Fig. 6. The primitive translation RNY model of Eigen and Schuster (1978).

tary to each other and as a stop codon is never complementary to another stop codon, several frames among the six frames in the two DNA double helix may simultaneously code for amino acids without start codon, leading to an optimal information storage. For example, the insertion sequence IS*1*, i.e. a transposable element in bacteria, includes six potential reading frames with two genes encoded on the same strand of IS*1* from partially overlapping reading frames (Machida et al., 1984). Overlapping genes on the same strand and genes on both DNA strands are classically found in the genomes of small size for maximizing the coding function, in particular in the mitochondrial genome, e.g. ATPase subunits six and eight, cytochrome oxidase subunit two and an adjacent ORF, etc. (Gray, 1992), and in the viral genome (Ziff, 1980). The concept of proteins coded by complementary strands has also been investigated for properties between amino acids and codons and for rules of the two complementary strands of DNA (Zull and Smith, 1990; Konecny et al., 1993; Béland and Allen, 1994; Konecny et al., 1995).

4.6. Conclusion summarizing the results

Based on a complete statistical analysis with 12 288 autocorrelation functions, three subsets of 20 trinucleotides (by excluding AAA, CCC, GGG and TTT) are identified in the three frames of protein genes of both prokaryotes and eukaryotes: X_0 in frame 0 and, X_1 and X_2 in the shifted frames 1 and 2, respectively (X_0, X_1 and X_2 being defined in Table 3(a)). Unexpectedly, X_0, X_1 and X_2 are maximal circular codes on the alphabet {A, C, G, T}, a concept suggested by Crick et al. (1957), then abandoned. These three circular codes have several relations between them. X_0 is self-complementary and, X_1 and X_2 are complementary to each other. X_0 generates X_1 by one circular permutation and X_2 by another circular permutation. Therefore, X_0 is a complementary Circular code with two permutated Circular codes (called C^3 code). The C^3 code X_0 has several properties: automatic retrieval of the frame after 13 nucleotides, rarity (6×10^{-8}) and concatenation properties (high frequency of misplaced trinucleotides in the shifted frames, maximum length of the minimal window to automatically retrieve the frame and the occurrence of the four types of nucleotides in the three trinucleotide sites). The different properties of the C^3 code X_0 as well as the few biological consequences analyzed here, in particular the possibility that several trinucleotides of X_0 may code for the same AA and the strong correlation between the usage of the trinucleotides of X_0 in protein genes and the amino acid frequencies in proteins, suggest that this code could have had a function in gene evolution and

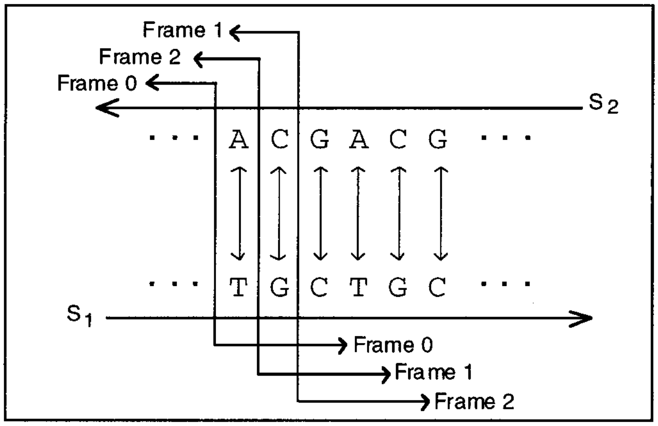


Fig. 7. The self-complementary circular code X_0 allows the two paired frames 0 (reading frames) to code simultaneously for amino acids without a start codon.

that the primitive alphabet could have been {A, C, G, T} rather than {R, Y}. An evolutionary model of the C^3 code X_0 based on substitutions is currently being investigated.

Acknowledgements

We thank Professor D. Perrin for his advice concerning the circular code and the referees for improving the presentation. This work was supported by GIP GREG grant (Groupement d'Intérêt Public, Groupement de Recherches et d'Etudes sur les Génomes), INSERM grant (Contrat de Recherche Externe No. 930101) and Mr Jean-Marc Vassards (Director of the society RVH, Mulhouse).

References

Arnason, U., Xu, X., Gullberg, A., 1996. Comparison between the complete mitochondrial DNA sequences of Homo and the common chimpanzee based on nonchimeric sequences. J. Mol. Evol. 42, 145–152.

Arquès, D.G., Lapayre, J.-C., Michel, C.J., 1995. Identification and simulation of shifted periodicities common to protein coding genes of eukaryotes, prokaryotes and viruses. J. Theor. Biol. 172, 279–291.

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

Arquès, D.G., Michel, C.J., 1990a. Periodicities in coding and noncoding regions of the genes. J. Theor. 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, Part 2: Simulation model, Part 3: Return of the model to the reality. Bull. Math. Biol. 52, 741–772.

Arquès, D.G., Michel, C.J., 1993. Identification and simulation of new non-random statistical properties common to different eukaryotic gene subpopulations. Biochimie 75, 399–407.

Arquès, D.G., Michel, C.J., 1994. Analytical expression of the purine/pyrimidine autocorrelation function after and before random mutations. Math. Biosci. 123, 103–125.

Barrell, B.G., Clark, B.F.C., 1974. Handbook of Nucleic Acid Sequences. Joynson–Bruvvers, Oxford.

Béal, M.-P., 1993. Codage Symbolique. Masson, Paris.

Béland, P., Allen, T.F.H., 1994. The origin and evolution of the genetic code. J. Theor. Biol. 170, 359–365.

Belcourt, M.F., Farabaugh, P.J., 1990. Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a seven nucleotide minimal site. Cell 62, 339–352.

Beremand, M.N., Blumenthal, T., 1979. Overlapping genes in RNA phage: a new protein implicated in lysis. Cell 18, 257–266.

Berstel, J., Perrin, D., 1985. Theory of Codes. Academic Press, New York.

Brierley, I., Jenner, A.J., Inglis, S.C., 1992. Mutational analysis of the 'slippery-sequence' component of a coronavirus ribosomal frameshifting signal. J. Mol. Biol. 227, 463–479.

Bruce, A.G., Atkins, J.F., Gesteland, R.F., 1986. tRNA anticodon replacement experiments show that ribosomal frameshifting can be caused by doublet decoding. Proc. Natl. Acad. Sci. USA 83, 5062–5066.

Craigen, W.J., Caskey, C.T., 1986. Expression of peptide chain release factor 2 requires high-efficiency frameshift. Nature 322, 273–275.

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

Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proc. Natl. Acad. Sci. USA 43, 416–421.

Curran, J.F., 1993. Analysis of effects of tRNA: message stability on frameshift frequency at the *Escherichia coli* RF2 programmed frameshift site. Nucleic Acids Res. 21, 1837–1843.

Dayhuff, T., Atkins, J., Gesteland, R., 1986. Characterization of ribosomal frameshif events by protein sequence analysis. J. Biol. Chem. 261, 7491–7500.

Dounce, A.L., 1952. Duplicating mechanism for peptide chain and nucleic acid synthesis. Enzymologia 15, 251–258.

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

Farabaugh, P.J., Zhao, H., Vimaladithan, A., 1993. A novel programmed frameshift expresses the POL3 gene of retrotransposon Ty3 of yeast: frameshifting without tRNA slippage. Cell 74, 93–103.

Fox, T.D., Weiss-Brummer, B., 1980. Leaky +1 and −1 frameshift mutations at the same site in a yeast mitochondrial gene. Nature 288, 60–63.

Fuller, W., Hodgson, A., 1967. Conformation of the anticodon loop in tRNA. Nature 215, 817–821.

Geballe, A.P., Morris, D.R., 1994. Initiation codons within 5′-leaders of mRNAs as regulators of translation. TIBS 19, 159–164.

Gray, M.W., 1992. The endosymbiont hypothesis revisited. Int. Rev. Cytol. 141, 233–357.

Grosjean, H.J., de Henau, S., Crothers, D.M., 1978. On the physical basis for ambiguity in genetic coding interactions. Proc. Natl. Acad. Sci. USA 75, 610–614.

Hartman, H., 1995. Speculations on the origin of the genetic code. J. Mol. Evol. 40, 541–544.

Jacks, T., Madhani, H.D., Marsiarz, F.R., Varmus, H.E., 1988. Signals for ribosomal frameshifting in the rous sarcoma virus *gag-pol* region. Cell 55, 447–458.

Jukes, T.H., Bhushan, V., 1986. Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. J. Mol. Evol. 24, 39–44.

Jukes, T.H., Holmquist, R., Moise, H., 1975. Amino acid composition of proteins: selection against the genetic code. Science 189, 50–51.

Konecny, J., Eckert, M., Schöniger, M., Hofacker, G.L., 1993. Neutral adaptation of the genetic code to double-strand coding. J. Mol. Evol. 36, 407–416.

Konecny, J., Schöniger, M., Hofacker, G.L., 1995. Complementary coding conforms to the primeval comma-less code. J. Theor. Biol. 173, 263–270.

Kozak, M., 1978. How do eucaryotic ribosomes select initiation regions in messenger RNA?. Cell 15, 1109–1123.

Kozak, M., 1986a. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J. Mol. Biol. 196, 947–950.

Kozak, M., 1986b. Bifunctional messenger RNAs in eukaryotes. Cell 47, 481–483.

Kozak, M., 1989. The scanning model for translation: an update. J. Cell Biol. 108, 229–241.

Larsen, B., Wills, N.M., Gesteland, R.F., Atkins, J.F., 1994. rRNA-mRNA base pairing stimulates a programmed −1 ribosomal frameshift. J. Bacteriol. 176, 6842–6851.

Lindsley, D., Gallant, J.A., 1993. On the directional specificity of ribosome frameshifting at a hungry codon. Proc. Natl. Acad. Sci. USA 90, 5469–5473.

Machida, Y., Machida, C., Ohtsubo, E., 1984. Insertion element IS*1* encodes two structural genes required for its transposition. J. Mol. Biol. 177, 229–245.

Matsufuji, S., Matsufuji, T., Miyazaki, Y., Murakami, Y., Atkins, J.F., Gesteland, R.F., Hayashi, S., 1995. Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. Cell 80, 51–60.

Morlé, F., Lopez, B., Henni, T., Godet, J., 1985. α-thalassemia associated with the deletion of two nucleotides at position −2 and −3 preceding the AUG codon. EMBO J. 4, 1245–1250.

Nirenberg, M.W., Matthaei, J.H., 1961. The dependance of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc. Natl. Acad. Sci. USA 47, 1588–1602.

O'Connor, M., Gesteland, R.F., Atkins, J.F., 1989. tRNA hopping: enhancement by an expanded anticodon. EMBO J. 8, 4315–4323.

Riddle, D.L., Carbon, J., 1973. Frameshift suppression: a nucleotide addition in the anticodon of a glycine tRNA. Nat. New Biol. 242, 230–234.

Saks, M.E., Sampson, J.R., 1995. Evolution of tRNA recognition systems and tRNA gene sequences. J. Mol. Evol. 40, 509–518.

Schimmel, P., Giegé, R., Moras, D., Yokoyama, S., 1993. An operational RNA code for amino acids and possible relationship to genetic code. Proc. Natl. Acad. Sci. USA 90, 8763–8768.

Shine, J., Dalgarno, L., 1974. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc. Natl. Acad. Sci. USA 71, 1342–1346.

Strubin, M., Long, E.O., Mach, B., 1986. Two forms of the Ia antigen-associated invariant chain result from alternative initiations at two in-phase AUGs. Cell 47, 619–625.

ten Dam, E.B., Pleij, C.W.A., Bosch, L., 1990. RNA pseudoknots: translational frameshifting and readthrough of viral RNAs. Virus Genes 4, 121–136.

Terrett, J.A., Miles, S., Thomas, R.H., 1996. Complete DNA sequence of the mitochondrial genome of *Cepea nemoralis* (gastropoda: pulmonata). J. Mol. Evol. 42, 160–168.

Tsuchihashi, Z., Brown, P.O., 1992. Sequence requirements for efficient translational frameshifting in the *E. coli dnaX* gene and the role of an unstable interaction between tRNA$^{\mathrm{Lys}}$ and an AAG lysine codon. Genes Dev. 6, 511–519.

Tuohy, T.M.F., Thompson, S., Gesteland, R.F., Atkins, J.F., 1992. Seven, eight and nine-membered anticodon loop mutants of $tRNA_2^{Arg}$ which cause +1 frameshifting. J. Mol. Biol. 228, 1042–1054.

Vögele, K., Schwartz, E., Welz, C., Schiltz, E., Rak, B., 1991. High-level ribosomal frameshifting directs the synthesis of IS*150* gene products. Nucleic Acids Res. 19, 4377–4389.

Watson, J.D., Crick, F.H.C., 1953. A structure for deoxyribose nucleic acid. Nature 171, 737–738.

Weiss, R.B., Dunn, D.M., Atkins, J.F., Gesteland, R.F., 1987. Slippery runs, shifty stops, backward steps and forward hops: −2, −1, +1, +2, +5, and +6 ribosomal frameshifting. Cold Spring Harbor Symp. Quant. Biol. 52, 687–693.

Weiss, R.B., Dunn, D.M., Dahlenberg, A.E., Atkins, J.F., Gesteland, R.F., 1988. Reading frame switch caused by base-pair formation between the 3′ end of 16S rRNA and the mRNA during elongation of protein synthesis in *Escherichia coli*. EMBO J. 7, 1503–1507.

Weiss, R.B., Dunn, D.M., Atkins, J.F., Gesteland, R.F., 1990. Ribosomal frameshifting from − 2 to +50 nucleotides. Prog. Nucleic Acids Res. Mol. Biol. 39, 159–183.

Wetzel, R., 1995. Evolution of the aminoacyl-tRNA synthetases and the origin of the genetic code. J. Theor. Biol. 40, 545–550.

Woese, C.R., 1970. Molecular mechanics of translation: a reciprocating ratchet mechanism. Nature 226, 817–820.

Ziff, E.B., 1980. Transcription and RNA processing by the DNA tumour viruses. Nature 287, 491–499.

Zull, J.E., Smith, S.K., 1990. Is genetic code redundancy related to retention of structural information in both DNA strands?. Trends Biochem. Sci. 15, 257–261.