J. theor. Biol. (1996) **182,** 45–58

A Complementary Circular Code in the Protein Coding Genes

DIDIER G. ARQUÈS† AND CHRISTIAN J. MICHEL‡

† *Equipe de Biologie Théorique*, *Université de Marne-la-Vallée*, *Institut Gaspard Monge*, 2 *rue de la butte verte*, 93160 *Noisy le Grand and* ‡ *Equipe de Biologie Théorique*, *Université de Franche-Comté*, *Institut Universitaire de Technologie de Belfort-Montbéliard*, *BP* 527, 90016 *Belfort*, *France*

(*Received on* 16 *June* 1995, *Accepted in revised form on* 8 *May* 1996)

Recently, shifted periodicities 1 modulo 3 and 2 modulo 3 have been identified in protein (coding) genes of both prokaryotes and eukaryotes with autocorrelation functions analysing eight of 64 trinucleotides (Arquès *et al*., 1995). This observation suggests that the trinucleotides are associated with frames in protein genes. In order to verify this hypothesis, a distribution of the 64 trinucleotides AAA, . . . , TTT is studied in both gene populations by using a simple method based on the trinucleotide frequencies per frame. In protein genes, the trinucleotides can be read in three frames: the reading frame 0 established by the ATG start trinucleotide and frame 1 (resp. 2) which is the frame 0 shifted by 1 (resp. 2) nucleotide in the 5′–3′ direction. Then, the occurrence frequencies of the 64 trinucleotides are computed in the three frames. By classifying each of the 64 trinucleotides in its preferential occurrence frame, i.e. the frame associated with its highest frequency, three subsets of trinucleotides can be identified in the three frames. This approach is applied in the two gene populations.

Unexpectedly, the same three subsets of trinucleotides are identified in these two gene populations: $T_0 = X_0 \cup \{\text{AAA,TTT}\}$ with $X_0 = \{\text{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,GAG,GAT,GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC}\}$ in frame 0, $T_1 = X_1 \cup \{\text{CCC}\}$ in frame 1 and $T_2 = X_2 \cup \{\text{GGG}\}$ in frame 2, each subset X_0, X_1 and X_2 having 20 trinucleotides. Surprisingly, these three subsets have five important properties: (i) the property of maximal circular code for X_0 (resp. X_1, X_2) allowing the automatical retrieval of frame 0 (resp. 1, 2) in any region of a protein gene model (formed by a series of trinucleotides of X_0) without using a start codon; (ii) the DNA complementarity property C (e.g. $C(\text{AAC}) = \text{GTT}$): $C(T_0) = T_0$, $C(T_1) = T_2$ and $C(T_2) = T_1$ allowing the two paired reading frames of a DNA double helix simultaneously to code for amino acids; (iii) the circular permutation property P (e.g. $P(\text{AAC}) = \text{ACA}$): $P(X_0) = X_1$ and $P(X_1) = X_2$ 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 equal to 6×10^{-8}; and (v) the concatenation property with: a high frequency (27.5%) of misplaced trinucleotides in the shifted frames, a maximum (13 nucleotides) length of the minimal window to automatically retrieve the frame and an occurrence of the four types of nucleotides in the three trinucleotides sites, in favour of an evolutionary code.

In the Discussion, the identified subsets T_0, T_1 and T_2 replaced in the three two-letter genetic alphabets purine/pyrimidine, amino/ceto and strong/weak interaction, allow us to deduce that the RNY model (R = purine = A or G, Y = pyrimidine = C or T, N = R or Y) (Eigen & Schuster, 1978) is the closest two-letter codon model to the trinucleotides of T_0. 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.

‡Author to whom correspondence should be addressed. Present address: Institut Polytechnique de Sévernans, Rue du Château, Sévernans, 90010 Belfort, France.

1. Introduction

The concept of a code without commas has been introduced by Crick *et al*. (1957) 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 of 64 codons, code for the 20 amino acids. However, the determination of a set of 20 codons forming a code 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 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 interesting property has naturally led to the proposal of a code without commas assigning one codon per amino acid (Crick *et al*., 1957). Unfortunately, the determination of a set of 20 codons forming a code without commas was not solved as there are 3.5 billion potential codes (explanation given in Section 3.4).

In contrast, Dounce (1952) has proposed earlier 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 & Bhushan, 1986).

The two discoveries that the codon TTT, an "excluded" codon in the concept of code without commas, codes for phenylalanine (Nirenberg & 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 on 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 on the alphabet {R,Y} (R = purine = A or G, Y = pyrimidine = C or T) with 2 codon models for the primitive protein genes: RRY (Crick *et al*., 1976) and RNY (N = R or Y) (Eigen & Schuster, 1978). The RNY codon model has interesting properties. Indeed, it 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 code simultaneously for amino acids according to a purine/pyrimidine genetic code. Furthermore, NYR is obtained by one circular permutation of RNY and YRN, by two circular permutations of RNY. This property allows NYR and YRN to be deduced from RNY. Finally, the length of the minimal sub-sequence (called window) to retrieve automatically 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 of automatically retrieving the reading frame in any region of a protein gene model (formed by a series of RNY codons) avoids having to use a start codon. The RNY model has on the reduced alphabet {R, Y} some of the properties of the code identified in protein genes of prokaryotes and eukaryotes on the alphabet {A,C,G,T}.

The recent observation of shifted periodicities (1 modulo 3 and 2 modulo 3) in protein genes of prokaryotes and eukaryotes with autocorrelation functions analysing the eight trinucleotides obtained by specifying YRY on {A,C,G,T} (YRY = {CAC,CAT,. . . ,TGT}), suggests that the trinucleotides have a preferential occurrence frame (Arquès *et al*., 1995). In order to verify this hypothesis, the occurrence frequencies of the 64 trinucleotides AAA, . . . , TTT are computed in the three frames for these two gene populations. By excluding AAA, CCC, GGG and TTT and classifying each of the 60 remaining trinucleotides in the frame associated with its highest frequency, three subsets of 20 trinucleotides are identified in the three frames: X_0

in frame 0 and, X_1 and X_2 in the shifted frames 1 and 2, respectively.

These three subsets X_0, X_1 and X_2 of 20 trinucleotides are identical in these two gene populations and have five important properties (detailed in Section 3): circular code, complementarity, circular permutation, rarity and concatenation.

(i) The circular code property: if a sequence is constructed by concatenating trinucleotides of X_0 (frame 0), like a protein gene is composed of a series of codons, and if the frame of construction is lost, e.g. situation observed when a region of a protein gene without start codon is sequenced, then the property of code assures that the constructed sequence is decomposable into a series of trinucleotides of X_0 according to a unique way. This unique decomposition can be retrieved using a window of nucleotides with a minimal length depending on X_0. In the actual protein genes, the frame 0 (reading frame) 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 beginning and without end. We will show that X_0, X_1 and X_2 are maximal (20 trinucleotides) circular codes. Note: the property that X_0 is a circular code does not necessarily imply that X_1 and X_2 are also circular codes.

(ii) The complementarity property: the subset X_0 of trinucleotides is self-complementary (ten trinucleotides of X_0 are complementary to ten other trinucleotides of X_0) and, the subsets X_1 and X_2 of trinucleotides are complementary to each other (20 trinucleotides of X_1 are complementary to 20 trinucleotides of X_2). Therefore, the two paired reading frames may simultaneously code for amino acids, in agreement with biological results (Zull & Smith, 1990; Konecny *et al*., 1993; Béland & Allen, 1994; Konecny *et al*., 1995).

(iii) The circular permutation property: the two subsets X_1 and X_2 of trinucleotides can be deduced from X_0 by circular permutations of one and two nucleotides respectively (one and two circular permutations with each trinucleotide of X_0 leads to the trinucleotides of X_1 and X_2, respectively).

(iv) The rarity property: so far, no statistical analysis has been performed concerning circular codes with trinucleotides on the alphabet $\{A,C,G,T\}$. There are 216 circular codes with the properties mentioned above (complementary circular codes with two permutated circular codes called C^3 codes) among 3.5 billion potential circular codes. Therefore, the occurrence probability of the code X_0 is equal to 6×10^{-8}. In addition, this code X_0 is observed in two independent large gene populations, the protein genes of prokaryotes (13 686 sequences) and eukaryotes (26 757 sequences).

(v) The concatenation property: 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, e.g. the concatenation CAGGAG of CAG and GAG of X_0 generates with a high probability AGG of X_1 in frame 1 and GGA of X_2 in frame 2. For X_0, this property is verified at 72.5% (27.5% of misplaced trinucleotides in the shifted frames), one of the lowest rate among the 216 C^3 codes. Therefore, the length of the minimal window of X_0 to automatically retrieve the frame is maximum (13 nucleotides) among the 216 C^3 codes. Finally, the four types of nucleotides occur in the three trinucleotide sites of X_0. Therefore, the code X_0 cannot be generated by the classical methods of automatic construction of circular codes with impose constraints in the trinucleotide sites, e.g. no nucleotide A in the first trinucleotide site. These three properties (high frequency of misplaced trinucleotides, maximum length of the minimal window and occurrence of the four types of nucleotides) imply that the code X_0 has evolutionary properties (a flexible code for evolution).

In summary, the code X_0 identified in protein genes of prokaryotes and eukaryotes retrieves on the alphabet $\{A,C,G,T\}$ the properties both of the code without commas on the alphabet $\{R,Y\}$ (Crick *et al*., 1976; Eigen & Schuster, 1978) and of the evolutionary code (Dounce, 1952).

In the Discussion, several consequences of the identified subsets X_0, X_1 and X_2 are studied in respect with the two-letter genetic alphabets, the genetic code, the amino acid frequencies in proteins and the complementary paired DNA sequence.

2. Method

The method is obvious. On the genetic alphabet $B = \{A,C,G,T\}$, there are 64 trinucleotides $w \in T = \{AAA,\ldots,TTT\}$. In protein genes, the trinucleotides w^p can be read in three frames $p \in \{0, 1, 2\}$, $p = 0$: reading frame established by the start trinucleotide ATG and $p = 1$ (resp. $p = 2$): reading frame shifted by 1 (resp. 2) nucleotide in the 5′–3′ direction. There are $64 \times 3 = 192$ trinucleotides w^p. The occurrence frequencies $P(w^p)$ are computed in two protein gene populations: prokaryotes (13686 sequences, 4708758 trinucleotides) and (nuclear) eukaryotes (26757 sequences, 11397678 trinucleotides). These large populations, obtained from the release 39 of the EMBL Nucleotide Sequence Data

Library in the same way as described in previous studies (see e.g. Arquès & Michel, 1990a, b for a description of data acquisition), allow to have stable frequencies. Then, each trinucleotide w is classified in its frame associated with its highest frequency.

3. Results

3.1. IDENTIFICATION OF THREE SUBSETS OF TRINUCLEOTIDES IN THE THREE FRAMES

Table 1(a) (resp. Table 1b) gives the occurrence frequencies $P(w^p)$ of the 192 trinucleotides w^p in the protein genes of prokaryotes (resp. eukaryotes). Very unexpectedly, these two tables 1(a, b) show that the 64 nucleotides w can easily be classified in three subsets of trinucleotides according to the frame [Table 2(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 2(a). By considering the four trinucleotides with identical nucleotides, three subsets X_0, X_1 and X_2 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. Among the 192 trinucleotides w^p, the few classified into two frames or misclassified [two for the prokaryotes and four for the eukaryotes, Tables 1(a, b)], have been assigned to the frame according to the properties identified with the other trinucleotides, both in prokaryotes and eukaryotes. Note: the frequencies of the three stop trinucleotides TAA, TAG and TGA in frame 0, are obviously equal to 0.

3.2. COMPLEMENTARITY PROPERTY

Recall of the DNA complementarity rule (*Watson & Crick*, 1953)

(i) The DNA double helix is formed of two nucleotide sequences s_1 and s_2 connected with the nucleotide pairing (hydrogen bonds) according to the complementarity rule C: the nucleotide A (resp. C, G, T) in s_1 pairs with the complementary nucleotide $C(\mathrm{A}) = \mathrm{T}$ (resp. $C(\mathrm{C}) = \mathrm{G}$, $C(\mathrm{G}) = \mathrm{C}$, $C(\mathrm{T}) = \mathrm{A}$) 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 $w = l_1 l_2 l_3$, $l_1, l_2, l_3 \in \{$A,C,G,T$\}$, in s_1 pairs with the complementary trinucleotide $C(w) = C(l_3)C(l_2)C(l_1)$ in s_2.

Property 1

$C(T_0) = T_0$, $C(T_1) = T_2$ and $C(T_2) = T_1$ [Table 2(b)]. T_0 is self-complementary and, T_1 and T_2 are complementary to each other.

Biological consequences

The two paired reading frames may simultaneously code for amino acids.

3.3. CIRCULARITY PROPERTY

Definition of the trinucleotide circular permutation

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

Property 2

$P(X_0) = X_1$ and $P(X_1) = X_2$ [Table (2c)]. X_0 generates X_1 by one circular permutation and X_2 by another circular permutation.

Biological consequences

The two subsets X_1 and X_2 can be deduced from X_0.

3.4. CIRCULAR CODE PROPERTY

Recall of a few notations

Let B be a genetic alphabet, $B_2 = \{$R,Y$\}$ and $B_4 = \{$A,C,G,T$\}$. B^* denotes the words on B of finite length including the empty word of length 0. B^+ denotes the words on B of finite length $\geqslant 1$. Let $w_1 w_2$ be the concatenation of the two words w_1 and w_2.

Definition of circular code

A subset X of B^+ is a circular code if for all $n, m \geqslant 1$ and $x_1, x_2, \ldots, x_n \in X$, $y_1, y_2, \ldots, y_m \in X$ and $p \in B^*$, $s \in B^+$, the equalities $s x_2 x_3 \ldots x_n p = y_1 y_2 \ldots y_m$ and $x_1 = ps$ imply $n = m$, $p = 1$ and $x_i = y_i$, $1 \leqslant i \leqslant n$ [Béal, 1993; Berstel & Perrin, 1985 and Fig. 1(a)]. In other terms, every word on B "written on a circle" has at most one factorization (decomposition) over X.

Note: in the following, X will be a set of words of three letters as a protein gene is a concatenation of trinucleotides.

Basic properties and examples

(i) If b is the cardinal of the alphabet B, then X contains at most b^3 trinucleotides [Table 2(d)]. Therefore, on B_2, X is a subset of $\{$RRR,RRY,RYR,RYY,YRR,YRY,YYR,YYY$\}$ and on B_4, X is a subset of $\{$AAA,. . . ,TTT$\}$.
(ii) There are obvious constraints so that X is a circular code.
$-$ X cannot have the trinucleotides $w = 111$, $l \in B$. For example on B_2, if X contains RRR then the word

TABLE 1(a)

Occurrence frequencies $P(w^p)$ of the 64 trinucleotide w in each frame p in the prokaryotic protein coding genes (13686 sequences, 4708758 trinucleotides)

w in frame $p = 0$	Frequency (%)	w in frame $p = 1$	Frequency (%)	w in frame $p = 2$	Frequency (%)
AAA	3.38	AAA	2.75	AAA	2.44
AAC	2.18	AAC	1.59	AAC	1.38
AAG	1.98	**AAG**	3.21	AAG	0.81
AAT	2.17	AAT	1.37	AAT	1.69
ACA	1.22	**ACA**	1.91	ACA	1.11
ACC	2.09	ACC	1.60	ACC	0.79
ACG	1.30	**ACG**	2.49	ACG	0.68
ACT	1.13	**ACT**	1.17	ACT	1.09
AGA	0.61	AGA	1.59	**AGA**	2.47
AGC	1.42	**AGC**	1.83	AGC	1.71
AGG	0.31	**AGG**	2.21	AGG	1.45
AGT	0.87	AGT	0.97	**AGT**	1.26
ATA	0.83	**ATA**	2.15	ATA	0.66
ATC	2.61	ATC	1.66	ATC	0.82
ATG	2.38	**ATG**	2.82	ATG	0.41
ATT	2.50	ATT	1.38	ATT	1.50
CAA	1.70	CAA	1.47	**CAA**	2.74
CAC	1.01	CAC	0.91	**CAC**	1.66
CAG	2.21	CAG	1.76	CAG	1.16
CAT	1.06	CAT	0.83	**CAT**	2.12
CCA	0.88	**CCA**	1.81	CCA	1.50
CCC	0.80	**CCC**	1.16	CCC	0.98
CCG	1.76	**CCG**	2.74	CCG	1.36
CCT	0.84	CCT	1.12	**CCT**	1.67
CGA	0.41	CGA	1.48	**CGA**	3.50
CGC	1.90	CGC	2.01	**CGC**	2.62
CGG	0.79	CGG	2.35	**CGG**	2.77
CGT	1.32	CGT	0.79	**CGT**	1.93
CTA	0.53	CTA	1.08	**CTA**	1.08
CTC	1.34	CTC	0.85	CTC	1.00
CTG	3.44	CTG	1.95	CTG	0.94
CTT	1.26	CTT	1.00	**CTT**	1.62
GAA	3.83	GAA	0.78	GAA	2.14
GAC	2.54	GAC	0.43	GAC	1.41
GAG	2.40	GAG	0.67	GAG	0.63
GAT	3.04	GAT	0.40	GAT	2.21
GCA	1.93	GCA	1.79	*GCA*	1.88
GCC	2.93	GCC	1.42	GCC	1.41
GCG	2.62	**GCG**	2.65	GCG	1.51
GCT	1.89	GCT	1.17	**GCT**	2.48
GGA	1.26	GGA	0.83	**GGA**	2.57
GGC	3.14	GGC	1.02	GGC	2.63
GGG	1.07	GGG	1.08	**GGG**	1.55
GGT	2.17	GGT	0.50	GGT	2.06
GTA	1.22	GTA	1.21	GTA	0.61
GTC	1.82	GTC	0.92	GTC	0.82
GTG	2.18	*GTG*	1.68	GTG	0.47
GTT	1.83	GTT	0.94	GTT	1.56
TAA	0.00	TAA	1.33	**TAA**	2.39
TAC	1.42	TAC	0.74	TAC	1.29
TAG	0.00	**TAG**	1.26	TAG	0.60
TAT	1.74	TAT	0.79	**TAT**	2.00
TCA	0.94	**TCA**	2.18	TCA	1.51
TCC	0.99	**TCC**	1.33	TCC	1.11
TCG	0.95	**TCG**	2.95	TCG	0.86
TCT	1.04	*TCT*	1.19	TCT	1.32
TGA	0.00	TGA	2.50	**TGA**	3.27
TGC	0.57	**TGC**	2.41	TGC	2.41
TGG	1.25	TGG	3.17	*TGG*	1.87
TGT	0.40	TGT	1.20	**TGT**	1.79
TTA	1.55	**TTA**	1.84	TTA	1.00
TTC	1.87	TTC	1.36	TTC	1.29
TTG	1.25	**TTG**	2.90	TTG	0.50
TTT	1.93	TTT	1.35	TTT	1.93

The trinucleotides in bold have a preferential occurrence frame. The trinucleotides in italics, classified into two frames p and p' ($|P(w^p) - P(w^{p'})| \leqslant 0.2\%$: GCA, TCT) or misclassified ($|P(w^p) - P(w^{p'})| > 0.2\%$: GTG, TGG), have been assigned to the frame according to the properties identified with the other trinucleotides, both in prokaryotes and eukaryotes.

TABLE 1(b)

Occurrence frequencies $P(w^p)$ of the 64 trinucleotide w in each frame p in the eukaryotic protein coding genes (26 757 sequences, 1 1 39 7 678 trinucleotides)

w in frame $p = 0$	Frequency (%)	w in frame $p = 1$	Frequency (%)	w in frame $p = 2$	Frequency (%)
AAA	2.64	AAA	2.16	AAA	2.24
AAC	2.40	AAC	1.33	AAC	1.12
AAG	3.56	*AAG*	2.64	AAG	1.05
AAT	2.10	AAT	1.24	AAT	1.39
ACA	1.48	**ACA**	2.79	ACA	1.28
ACC	1.93	ACC	1.90	ACC	1.04
ACG	0.75	**ACG**	1.62	ACG	0.43
ACT	1.51	**ACT**	1.71	ACT	1.03
AGA	1.23	AGA	2.97	*AGA*	2.84
AGC	1.57	**AGC**	2.28	AGC	1.57
AGG	0.95	**AGG**	3.05	AGG	1.46
AGT	1.10	AGT	1.45	*AGT*	1.29
ATA	0.88	**ATA**	1.51	ATA	0.61
ATC	2.32	ATC	1.29	ATC	1.00
ATG	2.31	**ATG**	3.08	ATG	0.57
ATT	1.98	ATT	1.25	ATT	1.33
CAA	1.65	CAA	1.55	**CAA**	3.71
CAC	1.28	CAC	1.14	**CAC**	2.17
CAG	2.66	CAG	2.22	CAG	1.82
CAT	1.01	CAT	1.02	**CAT**	2.75
CCA	1.66	**CCA**	2.91	CCA	2.04
CCC	1.50	**CCC**	1.60	CCC	1.50
CCG	0.73	**CCG**	1.48	CCG	1.03
CCT	1.46	CCT	1.59	**CCT**	2.16
CGA	0.53	CGA	0.62	**CGA**	1.94
CGC	0.96	CGC	0.78	**CGC**	1.16
CGG	0.71	CGG	1.10	**CGG**	1.28
CGT	0.67	CGT	0.48	**CGT**	1.19
CTA	0.73	CTA	1.20	*CTA*	1.17
CTC	1.64	CTC	1.44	CTC	1.64
CTG	2.94	CTG	2.85	CTG	1.32
CTT	1.27	CTT	1.20	**CTT**	2.13
GAA	3.09	GAA	1.06	GAA	2.96
GAC	2.58	GAC	0.63	GAC	1.37
GAG	3.54	GAG	1.27	GAG	1.28
GAT	2.68	GAT	0.57	GAT	1.70
GCA	1.58	GCA	1.93	*GCA*	1.81
GCC	2.51	GCC	1.37	GCC	1.42
GCG	0.83	**GCG**	1.11	GCG	0.80
GCT	2.17	GCT	1.28	*GCT*	1.97
GGA	1.76	GGA	1.27	**GGA**	3.49
GGC	2.08	GGC	0.95	GGC	2.08
GGG	1.20	GGG	1.26	**GGG**	1.55
GGT	1.74	GGT	0.61	GGT	1.67
GTA	0.78	GTA	0.91	GTA	0.73
GTC	1.60	GTC	0.81	GTC	1.05
GTG	2.40	*GTG*	1.93	GTG	0.77
GTT	1.55	GTT	0.75	GTT	1.35
TAA	0.00	TAA	1.02	**TAA**	1.79
TAC	1.76	TAC	0.68	TAC	1.02
TAG	0.00	**TAG**	1.03	TAG	0.70
TAT	1.34	TAT	0.67	**TAT**	1.40
TCA	1.20	**TCA**	2.82	TCA	1.47
TCC	1.63	**TCC**	1.85	TCC	1.40
TCG	0.66	**TCG**	1.37	TCG	0.61
TCT	1.56	**TCT**	1.71	TCT	1.42
TGA	0.00	TGA	2.46	**TGA**	3.61
TGC	1.08	*TGC*	2.00	TGC	2.28
TGG	1.21	TGG	3.39	*TGG*	2.48
TGT	0.89	TGT	1.36	**TGT**	2.17
TTA	1.02	**TTA**	1.29	TTA	0.74
TTC	2.20	TTC	1.36	TTC	1.34
TTG	1.50	**TTG**	2.69	TTG	0.61
TTT	1.75	TTT	1.14	TTT	1.70

The trinucleotides in bold have a preferential occurrence frame. The trinucleotides in italics, classified into two frames p and p' ($|P(w^p) - P(w^{p'})| \leqslant 0.2\%$: AGA, AGT, CTA, GCA, GCT, GTA) or misclassified ($|P(w^p) - (w^{p'})| > 0.2\%$: AAG, GTG, TGC, TGG), have been assigned to the frame according to the properties identified with the other trinucleotides, both in prokaryotes and eukaryotes.

. . .RRRRRRR. . . has three factorizations over X: . . .RRR,RRR,. . . , . . .R,RRR,RR. . . and . . . RR,RRR,R. . . Therefore, X will be a subset of $B_2' = \{\text{RRY,RYR,RYY,YRR,YRY,YYR}\}$. Similarly on B_4, X will be a subset of $B_4' = \{\text{AAA,}. . . , \text{TTT}\} - \{\text{AAA,CCC,GGG,TTT}\}$. The cardinal of B' is $b^3 - b$, i.e. six for B_2' and 60 for B_4'.

$-X$ cannot have two trinucleotides at the same time deduced from each other by circular permutation. For example on B_2, 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 B_2' in two classes of three codons so that the three codons are deduced from each other by circular permutations $\{\text{RRY,RYR,YRR}\}$ and $\{\text{RYY,YYR,YRY}\}$, X has at most one trinucleotide in each class. The number of classes invariant by circular permutation is $\text{Card}(B')/3 = (b^3 - b)/3$, i.e. two on B_2 and 20 on B_4 [Table 2(d)]. Therefore, X contains at most two trinucleotides and $3^2 = 9$ sets X are potential (maximal) circular codes. The number of potential (maximal) circular codes is $3^{(b^3 - b)/3}$, i.e. $3^2 = 9$ on B_2 and $3^{20} = 3\ 486\ 784\ 401$ on B_4 [Table 2(d)]. For example, $X_a = \{\text{RRY,RYY}\}$ is a circular code. Indeed, the concatenation of two trinucleotides of X_a, . . .RRYRRY. . . , . . . RRYRYY . . . , . . . RYYRRY . . . and . . . RYYRYY . . . leads to only one factorization over X_a as the eventual decomposition in frame 1 always has a R in the third position but no trinucleotide of X_a ends with R and as the eventual decomposition in frame 2 always has a Y in the first position but no trinucleotide of X_a begins with Y. X_a is a maximal (two trinucleotides) circular code and corresponds to the RNY model (Eigen & Schuster, 1978). X_a is self-complementary (complementary circular code), i.e. $C(X_a) = X_a$, 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 circular code and corresponds to the RRY model (Crick *et al*., 1976). The two subsets

TABLE 2(a)

List of the trinucleotides per frame in lexicographical order deduced from the Tables 1(*a*,*b*)

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	

Three subsets of trinucleotides can be identified: $T_0 = X_0 \cup \{\text{AAA,TTT}\}$ in frame $p = 0$, $T_1 = X_1 \cup \{\text{CCC}\}$ in frame $p = 1$ and $T_2 = X_2 \cup \{\text{GGG}\}$ in frame $p = 2$.

TABLE 2(b)

Complementarity property with the three subsets T_0, T_1 *and* T_2 *of trinucleotides identified in Table* 2(*a*)

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	

TABLE 2(c)

Circularity property with the three subsets X_0, X_1 *and* X_2 *of trinucleotides identified in Table* 2(*a*)

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	ACT		
X_2	CAA	TAA	CAC	CAT	TAT	GCA	CCT	GCT	AGA	CGA	GGA	TGA	CGC	CGG	TGG	AGT	CGT	TGT	CTA	CTT		

TABLE 2(d)

Circular code statistics on the alphabets $\{R,Y\}$ *and* $\{A,C,G,T\}$

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	$3^2 = 9$	$3^{20} = 3486784401$
Number of (maximal) circular codes	8	12964440
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 2 permutated circular codes (C^3 codes)	2	216
Probability of C^3 codes	0.22	6.2×10^{-8}

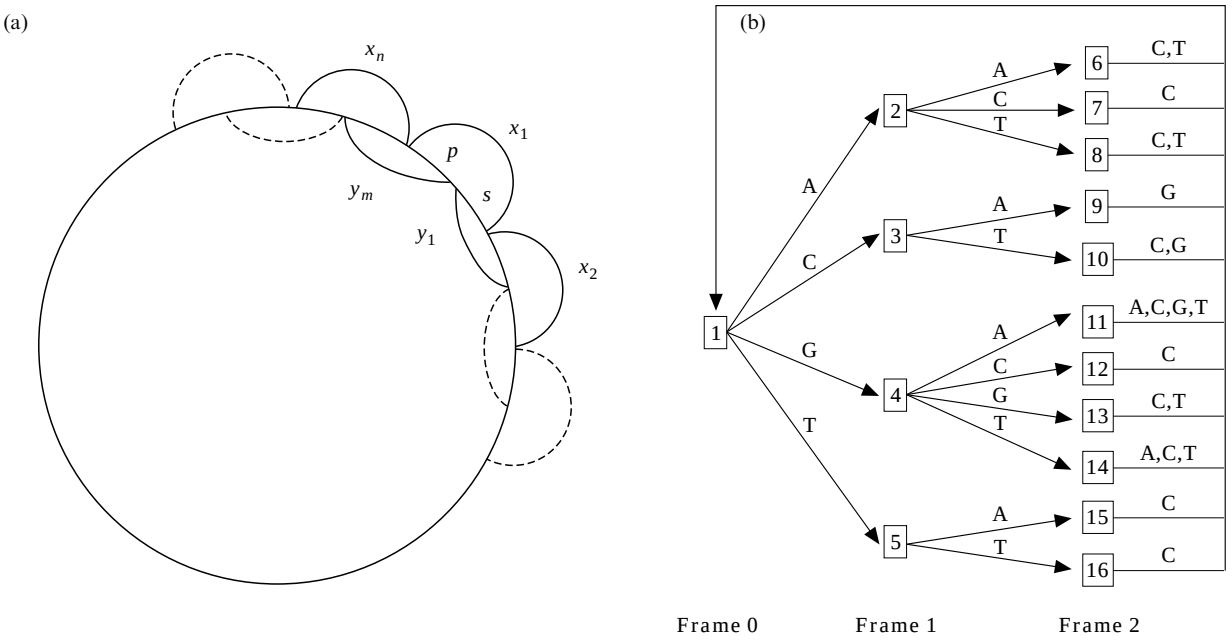

FIG. 1. (a) A representation of the definition of a circular code. (b) Flower automaton $F(X_0)$ associated with the circular code X_0.

$X_b = P(X_a) = \{\text{RYR,YYR}\}$ and $X_c = P(X_b) = \{\text{YRR,YRY}\}$ 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. $C(X_b) = X_c$ and $C(X_c) = X_b$, as RYR (resp. YYR) and YRY (resp. YRR) are complementary. The previous results remain unchanged by substituting R by Y and reciprocally. Therefore, $X_d = \{\text{YRR,YYR}\}$ is a maximal complementary circular code ($C(X_d) = X_d$) whose two subsets $X_e = P(X_d) = \{\text{RRY,YRY}\}$ and $X_f = P(X_e) = \{\text{RYR,RYY}\}$ obtained by circular permutations of X_d are also maximal circular codes and complementary to each other ($C(X_e) = X_f$ and $C(X_f) = X_e$). The three remaining sets X are $X_g = \{\text{RYY,YRR}\}$ and $X_h = P(X_g) = C(X_g) = \{\text{RRY,YYR}\}$ which are circular codes and $X_i = P(X_h) = \{\text{RYR,YRY}\}$ which is not a circular code as the word ...RYRYRYRYR ... has two factorizations over X_i: ... RYR,YRY,RYR,... and ...R,YRY,RYR,YR... On B_2 eight among nine sets X are circular codes and two sets $X_a = \{\text{RRY,RYY}\} = \text{RNY}$ and $X_d = \{\text{YRR,YYR}\} = \text{YNR}$ are complementary circular codes with two permutated circular codes (called C^3 codes) [Table 2(d)].

The study of circular codes on 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 & Perrin, 1985). The results obtained on {A,C,G,T} are new and very significant, from a biological as well as a computational point of view.

Property 3

The subset X_0 is a maximal (20 trinucleotides) circular code. The subsets X_1 and X_2 are also maximal circular codes.

Proof: (i) Proof of the maximum cardinal of a circular code:
The 60 words of $B_4' = \{\text{AAA},\ldots,\text{TTT}\} - \{\text{AAA,CCC,GGG,TTT}\}$ are gathered in 20 classes invariant by circular permutation. A circular code with words of three letters on 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 B_4 there are $3^{20} = 3\ 486\ 784\ 401$ potential circular codes, the use of some theorems in coding theory are necessary to the development of algorithms for determining automatically the circular codes. We give these basic theorems, the algorithms written in Pascal will be described elsewhere.

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

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

Definition 2: The flower automaton $F(X)$ associated with a subset X of B^+ has a particular state [labelled

1 in Fig. 1(b)] and cycles issued from this state 1 and labelled by words of X.

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

Fig. 1(b) gives the flower automaton $F(X_0)$ associated with X_0. To prove that "X_0 is a circular code" is equivalent to prove that "$F(X_0)$ is local", i.e. $F(X_0)$ does not contain two cycles labelling 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 B^+ verifying the proof of circular code and allows statistics with circular codes (Section 3.6).

Properties 1, 2 and 3 imply that X_0 is a complementary circular code with two permutated circular codes (called C^3 code).

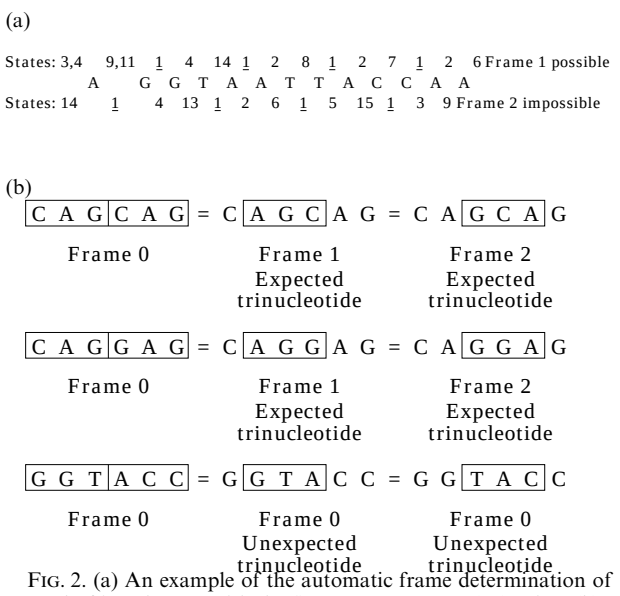

FIG. 2. (a) An example of the automatic frame determination of a word of length $\geqslant 13$ with the flower automaton $F(X_0)$ [Fig. 1(b)]. (b) Examples of three concatenations of two trinucleotides of the C^3 code X_0 generating expected and unexpected trinucleotides in the shifted frames 1 and 2.

3.5. AUTOMATIC FRAME DETERMINATION PROPERTY

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 the code assures that it can be retrieved in a unique way. Such a decomposition is called the reading frame of the word according to the code X_0.

The associated flower automaton $F(X_0)$ has several properties as all words of X_0 have a length of three letters [Fig. 1(b)]. Therefore, the states 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 words of X_0, can be associated with a unique state of the automaton $F(X_0)$, then a frame can be deduced for this letter because the associated unique state of $F(X_0)$ is related to a given frame. As a consequence, the word can be decomposed in words of X_0: its reading frame according to X_0 is then retrieved. The problem consists then in identifying such a unique state for a letter of the word. Such a unicity is not obvious. For example, the factor $w' = \text{AGGTAATTACCA}$ of length 12 can be attributed to two reading frames according to X_0: frame 1 (initial state 3 or 4 of w' in $F(X_0)$) or frame 2 (initial state 14 of w' in $F(X_0)$) [Fig. 2(a)].

In the case of a local automaton A, Definition 1 asserts that there exists n so that for a word of length $\geqslant n$, 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 $A = F(X_0)$, n is equal to 13 and $F(X_0)$ is called a 13-local automaton. In other words, the size of the minimal window to retrieve always the frame is 13 letters with X_0. For 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 $F(X_0)$. Then, the unique decomposition of w according to X_0 is AG,GTA,ATT,ACC,AA.

The length n of the minimal window to retrieve automatically the reading frame of a word according to X_0 can be determined by algorithm testing all possible paths in the automaton (Section 3.7).

Biological consequences

The code X_0 can retrieve automatically the frame 0 in any region of a protein gene model (formed by a series of trinucleotides of X_0) without using a start codon.

3.6. RARITY PROPERTY

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 3.4 and Table 2(d)].

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 determining such C^3 codes by hand or by the classical construction methods since the proposition of a code without commas by Crick *et al*. in 1957.

Table 2(d) summarizes the circular code statistics on the alphabets {R,Y} and {A,C,G,T}.

Biological consequences

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

3.7. CONCATENATION PROPERTY

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

(i) The largest window length. The lengths of minimal windows to retrieve the frame in the 216 C^3 codes are 5, 7, 9 and 13 (13 for X_0, Section 3.5).

(ii) A circularity property with a high frequency of misplaced trinucleotides in the shifted frames. The concatenation of two identical trinucleotides (process called duplication in biology) of X_0 (e.g. CAG $\in X_0$) leads obviously to the expected trinucleotides in the shifted frames [first example in Fig. 2(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 $\in X_1$ and GGA $\in X_2$) or not (GGT $\in X_0$ and ACC $\in X_0$ generate GTA $\notin X_1$ and TAC $\notin X_2$) [second and third examples in Fig. 2(b)]. Figure 3 shows the repartition function of the 216 C^3 codes according to the frequency of misplaced trinucleotides in frame 1 or 2 (the two shifted frames having the same frequency of misplaced trinucleotides by the complementarity property) 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.

(iii) An occurrence of the four types of nucleotides in the three trinucleotide sites.

Biological consequences

The code X_0 has evolutionary properties.

As the C^3 code X_0 leads to 72.5% (100–27.5) well-placed trinucleotides in the shifted frames (Fig. 3), the protein genes can be simulated with an independent mixing of the 20 trinucleotides of X_0 with equiprobability. Indeed, such a simulation retrieves the two other subsets X_1 and X_2 of trinucleotides in the frames 1 and 2 respectively (data not shown).

4. Discussion

4.1. CONSEQUENCES ON THE TWO-LETTER ALPHABETS

The three subsets T_0, T_1 and T_2 classify the 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 3(a) shows that the subset $Y_0 = \{\text{RRY,RYY}\} = \text{RNY}$ occurs preferentially in frame 0, the subset $Y_1 = \{\text{RYR,YYR}\}$, in frame 1, and the subset $Y_2 = \{\text{YRR,YRY}\}$, in frame 2. RRY and RYY have the same number (6) of A/C/G/T trinucleotides in frame 0. Y_0 contains a few A/C/G/T

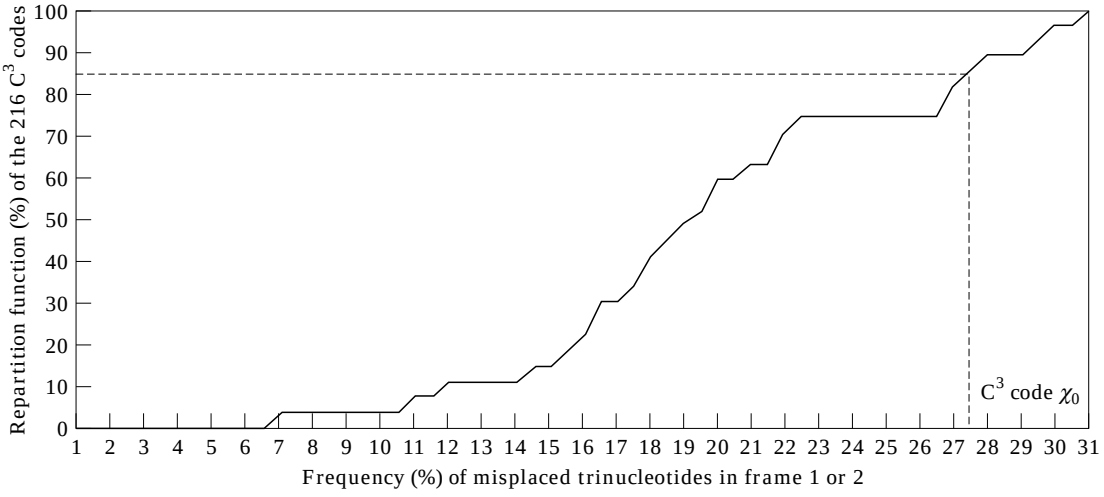

FIG. 3. Repartition function of the 216 C^3 codes according to the frequency of misplaced trinucleotides in frame 1 or 2 generated by the concatenation of trinucleotides of a given C^3 code. The frequency of misplaced trinucleotides for the C^3 code X_0 is equal to 27.5%.

trinucleotides in frames 1 (ACT^1, AGC^1) and 2 (AGT^2, GCT^2) and Y_1 and Y_2, a few A/C/G/T trinucleotides in frame 0 (CAG^0, CTG^0, GTA^0, TAC^0). The subset Y_0 is a C^3 code on {R,Y} (Section 3.4) and corresponds to the RNY codon model (Eigen & Schuster, 1978). There is no preferential frame for RRR and YYY.

These results also explain different previous works analysing the three frames simultaneously (average frame) in protein genes with autocorrelation functions on the alphabet {R,Y}, in particular: (i) the periodicity 0 modulo 3 with the autocorrelation function $YRY(N)_iYRY$ (Arquès & Michel, 1987, 1994) as the trinucleotides YRY occurring preferentially in frame 2 generate a number multiple of three bases between them; and (ii) the absence of the periodicity 0 modulo 3 with the autocorrelation function $RRR(N)_iRRR$ (Arquès & Michel,

TABLE 3(a)

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).

TABLE 3(b)

The eight K/M trinucleotides (K = ceto = G or T, M = amino = A or C) are associated with the 64 A/C/G/T trinucleotides by considering their frame (T_0, T_1, T_2)

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

The last row gives the preferential occurrence frame for the K/M trinucleotides.

TABLE 3(c)

The eight S/W trinucleotides (S = strong interaction = C or G, W = weak interaction = A or T) are associated with the 64 A/C/G/T trinucleotides by considering their frame (T_0, T_1, T_2)

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

The last row gives the preferential occurrence frame for the S/W trinucleotides.

1993) as RRR does not occur in a preferential frame.

This approach can be applied to the alphabets {K, M} (K = ceto = G or T, M = amino = A or C) and {S, W} (S = strong interaction = C or G, W = weak interaction = A or T). On the alphabet {K, M}, Table 3(b) shows that the subset $Z_0 = \{KKM,KMM\}$ occurs preferentially in frame 0, the subset $Z_1 = \{KMK,MMK\}$, in frame 1, and the subset $Z_2 = \{MKK,MKM\}$, in frame 2. On the alphabet {S, W}, Table 3(c) shows that SWS occurs preferentially in frame 0, WSS, in frame 1, and SSW, in frame 2. The subset $Z_0 = \{KKM,KMM\}$ (resp. SWS) is a maximal (resp. not maximal) circular code with two permutated circular codes (but no complementary).

The subset Y_0 (resp. Z_0, SWS) in frame 0 contains 12 (resp. 8, 6) A/C/G/T trinucleotides in frame 0. Therefore, the alphabet {R,Y} is the closest two-letter alphabet to the alphabet {A,C,G,T} and the subset $Y_0 = RNY$, the closest two-letter codon model to the trinucleotides of T_0.

4.2. 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 4(a)]. As T_0 has 22 codons, several codons of T_0 code for the same AA.

TABLE 4(a)

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

TABLE 4(b)

Number of trinucleotides of each subset T_0, T_1 and T_2 coding an amino acid

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
T_0	1	0	2	2	0	1	2	2	0	2	2	1	0	2	0	0	1	0	1	3
T_1	1	1	0	0	1	0	0	0	0	1	2	1	1	0	3	5	3	0	0	1
T_2	2	5	0	0	1	1	0	2	2	0	2	0	0	0	1	1	0	1	1	0

Six essential AA are coded by T_0: Ile, Leu, Lys, Phe, Thr and Val (two essential AA are not coded by T_0: Met and Trp). 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). There are some symmetrical features unexplained. The subsets T_1 and T_2 code a same number (11) of AA [Table 4(a)]. Table 4(b) gives the number of trinucleotides of each subset T_0, T_1 and T_2 coding an AA. Ser (resp. Arg) is coded by five trinucleotides of T_1 (resp. T_2) and by one trinucleotide of T_2 (resp. T_1). Val (resp. Thr) is coded by three trinucleotides of T_0 (resp. T_1) and by one trinucleotide of T_1 (resp. T_0). Otherwise, Asn, Asp, Glu and Phe are only coded by T_0, Met, by T_1 and, His and Trp, by T_2.

4.3. CONSEQUENCES ON THE AMINO ACID FREQUENCIES IN PROTEINS

Seven amino acids (AA) are not coded by the codon subset T_0: Arg, Cys, His, Met, Pro, Ser and Trp [Table 4(a)]. 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 (3044028 AA) and 20673 eukaryotic proteins (7521044 AA). They are obtained from the protein data base SWISS-PROT (release 29 from 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 5 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.

4.4. CONSEQUENCES ON THE COMPLEMENTARY PAIRED DNA SEQUENCE

As the set X_0 of trinucleotides is a circular code (property to retrieve automatically the frame 0) and self-complementary, the two paired frames 0 (reading frames) in the two DNA double helix may simultaneously code for amino acids without using a start codon (Fig. 4), in agreement with biological arguments (Zull & Smith, 1990; Konecny *et al*., 1993;

TABLE 5

Observed/expected frequency ratios in the proteins of prokaryotes and eukaryotes

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	287360	535229	9.44	7.12	1.44	1.08
Arg, R, Arginine (6/61)	9.84	164498	377123	5.40	5.01	0.55	0.51
Asn, N, Asparagine (2/61)	3.28	127874	337816	4.20	4.49	1.28	1.37
Asp, D, Aspartic acid (2/61)	3.28	170363	392502	5.60	5.22	1.71	1.59
Cys, C, Cysteine (2/61)	3.28	30213	151439	0.99	2.01	0.30	0.61
Gln, Q, Glutamine (2/61)	3.28	119579	312059	3.93	4.15	1.20	1.26
Glu, E, Glutamic acid (2/61)	3.28	190544	488770	6.26	6.50	1.91	1.98
Gly, G, Glycine (4/61)	6.56	234665	515987	7.71	6.86	1.18	1.05
His, H, Histidine (2/61)	3.28	63907	175381	2.10	2.33	0.64	0.71
Ile, I, Isoleucine (3/61)	4.92	180442	402872	5.93	5.36	1.20	1.09
Leu, L, Leucine (6/61)	9.84	288009	690767	9.46	9.18	0.96	0.93
Lys, K, Lysine (2/61)	3.28	158728	466467	5.21	6.20	1.59	1.89
Met, M, Methionine (1/61)	1.64	74134	173337	2.44	2.30	1.48	1.41
Phe, F, Phenylalanine (2/61)	3.28	114989	311375	3.78	4.14	1.15	1.26
Pro, P, Proline (4/61)	6.56	132643	390457	4.36	5.19	0.66	0.79
Ser, S, Serine (6/61)	9.84	183680	568759	6.03	7.56	0.61	0.77
Thr, T, Threonine (4/61)	6.56	173058	424350	5.69	5.64	0.87	0.86
Trp, W, Tryptophan (1/61)	1.64	38564	92793	1.27	1.23	0.77	0.75
Tyr, Y, Tyrosine (2/61)	3.28	94076	238524	3.09	3.17	0.94	0.97
Val, V, Valine (4/61)	6.56	216702	475037	7.12	6.32	1.09	0.96

Arg, Cys, His, Pro, Ser and Trp have the lowest observed/expected frequency ratios (<0.8) in the proteins of both prokaryotes (9510 sequences, 3044028 amino acids) and eukaryotes (20673 sequences, 7521044 amino acids) as expected with the usage of the codons of T_0 in protein genes.

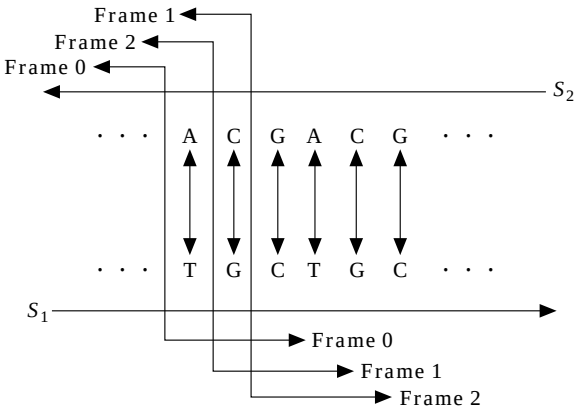

FIG. 4. The self-complementary circular code X_0 allows the two paired frames 0 (reading frames) simultaneously to code for amino acids without using a start codon.

Béland & Allen, 1994; Konecny *et al*., 1995). Furthermore, as the sets X_1 and X_2 of trinucleotides are circular codes and complementary 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 using a start codon, leading to an optimal information storage.

5. Conclusion

The identification in the protein genes of both prokaryotes and eukaryotes, of a circular code on the alphabet {A,C,G,T} which automatically retrieves the frame after 13 nucleotides and which has exceptional properties of complementarity, circular permutation, rarity (6×10^{-8}), concatenation (high frequency of misplaced trinucleotides in the shifted frames, maximum length of the minimal window to automatically retrieve the frame and occurrence of the four types of nucleotides in the three trinucleotide sites) and coding amino acids by several codons, suggests that this code could have had a function in gene evolution and that the primitive alphabet could have been {A,C,G,T} rather than {R,Y}.

We thank Prof. D. Perrin for his advice concerning the circular code. 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

ARQUÈS, D. G. & MICHEL, C. J. (1987). A purine-pyrimidine motif verifying an identical presence in almost all gene taxonomic groups. *J. theor. Biol.* **128,** 457–461.

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. Biosc.* **123,** 103–125.

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.

BÉAL, M.-P. (1993). *Codage Symbolique*. Paris: Masson.

BÉLAND, P. & ALLEN, T. F. H. (1994). The origin and evolution of the genetic code. *J. theor. Biol.* **170,** 359–365.

BERSTEL, J. & PERRIN, D. (1985). Theory of Codes. London: Academic Press.

CRICK, F. H. C., GRIFFITH, J. S. & ORGEL, L. E. (1957). Codes without commas. *Proc. Natl. Acad. Sci.* **43,** 416–421.

CRICK, F. H. C., BRENNER, S., KLUG, A. & PIECZENIK, G. (1976). A speculation on the origin of protein synthesis. *Origins of Life* **7,** 389–397.

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.

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.

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.* **47,** 1588–1602.

WATSON, J. D. & CRICK, F. H. C. (1953). A structure for deoxyribose nucleic acid. *Nature* **171,** 737–738.

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.