J. theor. Biol. (1997) **189,** 273–290

A Circular Code in the Protein Coding Genes of Mitochondria

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*, *France and* †*Equipe de Biologie Théorique*, *Institut Polytechnique de Sévenans*, *Rue du Château*, *Sévenans*, 90010 *Belfort*, *France*

(*Received on* 23 *May* 1997, *Accepted in revised form on* 9 *July* 1997)

A new maximal circular code $\mathscr{X}_0$(MIT) with two permutated maximal circular codes $\mathscr{X}_1$(MIT) and $\mathscr{X}_2$(MIT) is identified in the protein coding genes of mitochondria. The three subsets of 20 trinucleotides $\mathscr{X}_0$(MIT) = {ACA, ACC, ATA, ATC, CTA, CTC, GAA, GAC, GAT, GCA, GCC, GCT, GGA, GGC, GGT, GTA, GTC, GTT, TTA, TTC}, $\mathscr{X}_1$(MIT) and $\mathscr{X}_2$(MIT) are in frame 0 (reading frame), 1 and 2 respectively. $\mathscr{X}_1$(MIT) and $\mathscr{X}_2$(MIT) are deduced by one and two circular permutations of $\mathscr{X}_0$(MIT) respectively. The code $\mathscr{X}_0$(MIT) has four important properties: a length of the minimal window to automatically retrieve frame 0 which is equal to five nucleotides; an occurrence probability equal to 6.3×10^{-5}; a low frequency (12% in average) of misplaced trinucleotides in the shifted frames; and an occurrence of four types of nucleotides in the first and second trinucleotide sites but no nucleotide G in the third trinucleotide site. Several biological consequences are presented in the Discussion.

1. Introduction

The concept of a code without a 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 $\mathscr{X}$ of codons so that their concatenation (series of codons) leads to genes which have the interesting property of being able to retrieve automatically the concatenation of codons of $\mathscr{X}$ without the use of a start codon in a case where 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 are lost). 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 do we 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) 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 $\mathscr{X}$ 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 the retrieval of 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 to such a code at the same time. 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

‡ Author to whom correspondence should be addressed. E-mail: arques@univ-mlv.fr.

0022–5193/97/230273 + 18 $25.00/0/jt970513

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) proposed a flexible 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 the concept of code without commas on the alphabet $\{A, C, G, T\}$ being disrupted. 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 & Schuster, 1978).

In order to understand the circular code identified here in protein genes of mitochondria on the alphabet $\{A, C, G, T\}$, the concept of a circular code is introduced on the alphabet $\{R, Y\}$ with the RNY codon model (Eigen & 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. when a part of a protein gene without 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. Indeed, the concatenation of two trinucleotides of RNY, ... RRYRRY ..., ... RRYRYY ..., ... RYYRRY ... and ... RYYRYY ... leads to only one decomposition over RNY as the eventual decomposition in frame 1 always has an R in the third position but no trinucleotide of RNY ends with R and as the eventual decomposition in frame 2 always has a 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 frame 0 shifted by one and two nucleotides, respectively, in the $5' \rightarrow 3'$ direction). RNY is self-complementary and, NYR and YRN are complementary to each other. This property allows the two paired reading frames in the DNA double helix to code simultaneously for amino acids according to a purine/pyrimidine genetic code. Furthermore, NYR (resp. YRN) is obtained by one (resp. two) circular permutation of RNY. This property allows NYR and YRN to be deduced from RNY. Finally, the length of the minimal window to retrieve automatically frame 0 in a series RNYRNY ... generated by the circular code RNY 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 us to retrieve automatically the reading frame in any region of the gene (formed by a series of RNY codons), without a start codon.

The computation of the occurrence frequencies of the 64 trinucleotides AAA, ..., TTT in the three frames 0, 1, 2 of mitochondrial protein genes and the assignment of each trinucleotide to the frame associated with its highest frequency, identifies three subsets of trinucleotides per frame which have several important properties: maximal circular code, circular permutation, automatic frame determination, rarity, low frequency of misplaced trinucleotides in the shifted frames and occurrence of the four types of nucleotides in the first and second trinucleotide sites but no nucleotide G in the third trinucleotide site. Several consequences are studied with respect to the three two-letter genetic alphabets (purine/pyrimidine, amino/ceto, strong/weak interaction), the mitochondrial genetic code and the amino acid frequencies in mitochondrial proteins. Finally, several similarities and differences between the two codes identified in protein genes of eukaryotes/prokaryotes (Arquès & Michel, 1996) and mitochondria are presented.

2. Method

The method is obvious. In the genetic alphabet $\{A, C, G, T\}$, there are 64 trinucleotides $w \in \mathscr{T} = \{AAA, \ldots, TTT\}$. In protein genes, the trinucleotide w can be read in three frames $p \in \{0, 1, 2\}$ and then noted w^p with $p = 0$: reading frame established by the start trinucleotide ATG and $p = 1$ (resp. $p = 2$): reading frame shifted by one (resp. two) nucleotide in the $5' \rightarrow 3'$ direction. There are $64 \times 3 = 192$ trinucleotides w^p. The occurrence frequencies $P(w^p)$ are computed in the protein gene population

of mitochondria MIT (1303 sequences, 350963 trinucleotides). This large population, obtained from the release 47 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], allows us to obtain significant statistical results (stable frequencies). Then, each trinucleotide w is classified in the frame associated with its highest frequency.

3. Results

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

Table 1 gives the occurrence frequencies $P(w^p)$ of the 192 trinucleotides w^p in the protein genes of mitochondria MIT. Very unexpectedly, Table 1 shows that the 64 trinucleotides w can easily be classified into three subsets of trinucleotides according to the frame (Table 2). The 21 trinucleotides in frame 0 form the subset $\mathscr{T}_0(\mathrm{MIT}) = \{\mathrm{ACA, ACC, ATA, ATC, CTA, CTC, GAA, GAC, GAT, GCA, GCC, GCT, GGA, GGC, GGG, GGT, GTA, GTC, GTT, TTA, TTC}\}$ and the 21 and 22 trinucleotides in frames 1 and 2, the subsets $\mathscr{T}_1(\mathrm{MIT})$ and $\mathscr{T}_2(\mathrm{MIT})$ respectively, $\mathscr{T} = \mathscr{T}_0(\mathrm{MIT}) \cup \mathscr{T}_1(\mathrm{MIT}) \cup \mathscr{T}_2(\mathrm{MIT})$ with $\mathscr{T}_1(\mathrm{MIT})$ and $\mathscr{T}_2(\mathrm{MIT})$ defined in Table 2. By considering the four trinucleotides with identical nucleotides, three subsets $\mathscr{X}_0(\mathrm{MIT})$, $\mathscr{X}_1(\mathrm{MIT})$ and $\mathscr{X}_2(\mathrm{MIT})$ of 20 trinucleotides can be defined from $\mathscr{T}_0(\mathrm{MIT})$, $\mathscr{T}_1(\mathrm{MIT})$ and $\mathscr{T}_2(\mathrm{MIT})$: $\mathscr{X}_0(\mathrm{MIT}) = \mathscr{T}_0(\mathrm{MIT}) - \{\mathrm{GGG}\}$, $\mathscr{X}_1(\mathrm{MIT}) = \mathscr{T}_1(\mathrm{MIT}) - \{\mathrm{CCC}\}$ and $\mathscr{X}_2(\mathrm{MIT}) = \mathscr{T}_2(\mathrm{MIT}) - \{\mathrm{AAA, TTT}\}$. Among the 192 trinucleotides w^p, a very few ones classified into two frames (3) or misclassified (1), have been assigned to the frame according to the properties identified with the other trinucleotides (Table 1). Note: the frequencies of the two mitochondrial stop trinucleotides TAA and TAG in frame 0 are obviously equal to 0 (the universal stop trinucleotide TGA is coding Trp in mitochondria).

3.2. CIRCULARITY PROPERTY

Definition of the trinucleotide circular permutation: the circular permutation $\mathscr{P}$ of the trinucleotide $w = l_1l_2l_3$, $l_1, l_2, l_3 \in \{\mathrm{A, C, G, T}\}$, is the permutated trinucleotide $\mathscr{P}(w) = l_2l_3l_1$, e.g. $\mathscr{P}(\mathrm{AAC}) = \mathrm{ACA}$ (e.g. $\mathscr{P}(\mathrm{RRY}) = \mathrm{RYR}$ on $\{\mathrm{R, Y}\}$).

Property 1: $\mathscr{P}(\mathscr{X}_0(\mathrm{MIT})) = \mathscr{X}_1(\mathrm{MIT})$ and $\mathscr{P}(\mathscr{X}_1(\mathrm{MIT})) = \mathscr{X}_2(\mathrm{MIT})$ (Table 3). $\mathscr{X}_0(\mathrm{MIT})$ generates $\mathscr{X}_1(\mathrm{MIT})$ by one circular permutation and $\mathscr{X}_2(\mathrm{MIT})$ by another circular permutation (one and two circular permutations with each trinucleotide of $\mathscr{X}_0(\mathrm{MIT})$ lead to the trinucleotides of $\mathscr{X}_1(\mathrm{MIT})$ and $\mathscr{X}_2(\mathrm{MIT})$ respectively).

3.3. CIRCULAR CODE PROPERTY

3.3.1. *Definition of a circular code*

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

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

In the following, $\mathscr{X}$ will be a set of words of length 3 as a protein gene is a concatenation of trinucleotides.

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

A complete study of circular codes with trinucleotides on the alphabet $\mathscr{B}_2$ has not been presented so far. Furthermore, such a study also allows us to introduce the different concepts and properties of circular codes which are complex on the alphabet $\mathscr{B}_4 = \{\mathrm{A, C, G, T}\}$ and cannot be analysed by hand. Indeed, there are nine potential maximal circular codes on $\mathscr{B}_2$ but 3.5 milliards on $\mathscr{B}_4$.

If b is the cardinal of the alphabet $\mathscr{B}$, then $\mathscr{X}$ contains at most b^3 trinucleotides (Table 4). Therefore, on $\mathscr{B}_2$, $\mathscr{X}$ is a subset of $\{\mathrm{RRR, RRY, RYR, RYY, YRR, YRY, YYR, YYY}\}$ of cardinal 8. There are two obvious constraints so that $\mathscr{X}$ is a circular code:

(i) $\mathscr{X}$ cannot have the trinucleotides $w = lll$, $l \in \mathscr{B}$. For example, if $\mathscr{X}$ contains RRR then the word ... RRRRRR ... has three factorizations over $\mathscr{X}$: ... RRR, RRR ..., ... R, RRR, RR... and ...RR, RRR, R ... Therefore, $\mathscr{X}$ will be a subset of $\mathscr{B}'_2 = \{\mathrm{RRY, RYR, RYY, YRR, YRY, YYR}\}$. The cardinal of $\mathscr{B}'$ is $b^3 - b$, i.e. 6 for $\mathscr{B}'_2$.

(ii) $\mathscr{X}$ cannot have at the same time two trinucleotides deduced from each other by circular permutation. For example, if $\mathscr{X}$ contains RRY and RYR (RYR is one circular permutation of RRY) then the word ... RRYRRYRRY ... has two

TABLE 1

Occurrence frequencies $P(w^p)$ of the 64 *trinucleotides w* (*AAA*, . . ., *TTT*) *in each frame p* (0, 1, 2) *in the protein coding genes of mitochondria MIT* (1303 *sequences*, 350963 *trinucleotides*)

w in frame $p = 0$	Frequency (%)	w in frame $p = 1$	Frequency (%)	w in frame $p = 2$	Frequency (%)
AAA	3.07	AAA	2.63	**AAA**	3.30
AAC	1.79	AAC	1.41	**AAC**	2.39
AAG	0.89	**AAG**	1.85	AAG	1.14
AAT	2.86	AAT	1.79	**AAT**	4.38
ACA	2.50	ACA	1.49	ACA	1.61
ACC	1.58	ACC	1.42	ACC	1.44
ACG	0.34	**ACG**	1.09	ACG	0.69
ACT	1.81	ACT	1.36	**ACT**	2.22
AGA	0.93	AGA	0.66	**AGA**	2.49
AGC	0.70	AGC	0.43	**AGC**	2.56
AGG	0.29	AGG	0.63	**AGG**	2.58
AGT	1.10	AGT	0.45	**AGT**	2.10
ATA	3.46	ATA	2.91	ATA	1.36
ATC	2.59	ATC	1.43	ATC	1.75
ATG	1.88	**ATG**	1.98	ATG	1.11
ATT	4.44	ATT	2.19	**ATT**	4.54
CAA	2.12	**CAA**	2.58	CAA	1.47
CAC	1.10	CAC	1.44	**CAC**	1.47
CAG	0.47	**CAG**	2.33	CAG	0.45
CAT	1.33	CAT	2.22	**CAT**	2.29
CCA	1.90	*CCA*	1.57	CCA	1.32
CCC	1.14	**CCC**	1.77	CCC	1.53
CCG	0.31	**CCG**	1.02	CCG	0.52
CCT	1.47	CCT	1.44	**CCT**	3.01
CGA	0.91	CGA	0.40	*CGA*	0.88
CGC	0.38	CGC	0.26	**CGC**	1.04
CGG	0.21	CGG	0.49	**CGG**	0.97
CGT	0.61	CGT	0.26	**CGT**	0.96
CTA	3.35	CTA	2.33	CTA	0.96
CTC	1.69	CTC	1.23	CTC	1.51
CTG	0.78	**CTG**	1.59	CTG	0.74
CTT	2.13	CTT	1.99	**CTT**	2.61
GAA	2.48	GAA	1.63	GAA	0.90
GAC	1.09	GAC	0.92	GAC	0.54
GAG	0.81	**GAG**	1.63	GAG	0.37
GAT	1.87	GAT	1.34	GAT	1.02
GCA	1.94	GCA	0.58	GCA	0.50
GCC	1.87	GCC	0.77	GCC	0.49
GCG	0.39	**GCG**	0.46	GCG	0.23
GCT	2.20	GCT	0.86	GCT	0.68
GGA	2.31	GGA	0.48	GGA	0.89
GGC	1.15	GGC	0.33	GGC	0.85
GGG	0.87	GGG	0.71	GGG	0.77
GGT	1.89	GGT	0.49	GGT	0.78
GTA	2.20	GTA	1.17	GTA	0.38
GTC	0.98	GTC	0.68	GTC	0.39
GTG	0.76	**GTG**	1.13	GTG	0.28
GTT	1.84	GTT	1.24	GTT	1.14
TAA	0.00	**TAA**	4.37	TAA	2.95
TAC	1.38	**TAC**	2.19	TAC	1.83
TAG	0.00	**TAG**	3.91	TAG	1.05
TAT	2.45	*TAT*	3.40	TAT	4.31
TCA	2.23	*TCA*	2.04	TCA	1.60
TCC	1.20	**TCC**	2.41	TCC	1.37
TCG	0.37	**TCG**	1.29	TCG	0.67
TCT	1.66	**TCT**	2.17	TCT	2.04
TGA	1.39	TGA	1.28	**TGA**	1.98
TGC	0.43	TGC	0.88	**TGC**	1.95
TGG	0.64	TGG	1.45	**TGG**	1.91
TGT	0.62	TGT	1.00	**TGT**	1.94
TTA	4.85	TTA	3.74	TTA	1.47
TTC	2.67	TTC	2.33	TTC	1.81
TTG	1.21	**TTG**	3.07	TTG	0.98
TTT	4.10	TTT	3.41	**TTT**	4.54

The trinucleotides in bold have a preferential occurrence frame. The trinucleotides in italics, classified into two frames p and p' ($|P(w^p) - P(w^{p'})| \leqslant 0.4\%$: CCA, CGA, TCA) or misclassified ($|P(w^p) - P(w^{p'})| > 0.4\%$: TAT) have been assigned to the frame according to the properties identified with the other trinucleotides.

TABLE 2

List per frame and in lexicographical order of the trinucleotides deduced from the Table 1

$\mathscr{T}_0$(MIT):	ACA	ACC	ATA	ATC	CTA	CTC	GAA	GAC	GAT	GCA	GCC	GCT	GGA	GGC	GGG	GGT	GTA	GTC	GTT	TTA	TTC
$\mathscr{T}_1$(MIT):	AAG	ACG	ATG	CAA	CAG	CCA	CCC	CCG	CTG	GAG	GCG	GTG	TAA	TAC	TAG	TAT	TCA	TCC	TCG	TCT	TTG
$\mathscr{T}_2$MIT):	AAA	AAC	AAT	ACT	AGA	AGC	AGG	AGT	ATT	CAC	CAT	CCT	CGA	CGC	CGG	CGT	CTT	TGA	TGC	TGG	TGT	TTT

Three subsets of trinucleotides can be identified: $\mathscr{T}_0(\text{MIT}) = \mathscr{X}_0(\text{MIT}) \cup \{\text{GGG}\}$ in frame 0, $\mathscr{T}_1(\text{MIT}) = \mathscr{X}_1(\text{MIT}) \cup \{\text{CCC}\}$ in frame 1 and $\mathscr{T}_2(\text{MIT}) = \mathscr{X}_2(\text{MIT}) \cup \{\text{AAA, TTT}\}$ in frame 2.

TABLE 3

Circularity property with the three subsets $\mathscr{X}_0(MIT)$, $\mathscr{X}_1(MIT)$ and $\mathscr{X}_2(MIT)$ of trinucleotides identified in Table 2

$\mathscr{X}_0$(MIT):	ACA	ACC	ATA	ATC	CTA	CTC	GAA	GAC	GAT	GCA	GCC	GCT	GGA	GGC	GGT	GTA	GTC	GTT	TTA	TTC
$\mathscr{X}_1$(MIT):	CAA	CCA	TAA	TAC	TCC	AAG	ACG	ATG	CAG	CCG	CTG	GAG	GCG	GTG	TAG	TCG	TTG	TAT	TCT	
$\mathscr{X}_2$(MIT):	AAC	CAC	AAT	CAT	ACT	CCT	AGA	CGA	TGA	AGC	CGC	TGC	AGG	CGG	TGG	AGT	CGT	TGT	ATT	CTT

factorizations over $\mathscr{X}$: . . . RRY, RRY, RRY . . . and . . . R, RYR, RYR, RY . . . Therefore, by gathering the six trinucleotides of $\mathscr{B}_2'$ into two classes of three codons so that the three codons are deduced from each other by circular permutations $\{\text{RRY, RYR, YRR}\}$ and $\{\text{RYY, YYR, YRY}\}$, $\mathscr{X}$ has at most one trinucleotide in each class. Therefore, $\mathscr{X}$ contains at most two trinucleotides and $3^2 = 9$ sets $\mathscr{X}$ are potential maximal circular codes. The number of classes invariant by circular permutation is $\text{Card}(\mathscr{B}')/3 = (b^3 - b)/3$, i.e. 2 on $\mathscr{B}_2$ (Table 4). The number of potential maximal circular codes is $3^{(b^3-b)/3}$, i.e. $3^2 = 9$ on $\mathscr{B}_2$ (Table 4).

Recall of the DNA complementarity rule (Watson & Crick, 1953): (i) the DNA double helix is formed from 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 $\mathscr{C}(\text{A}) = \text{T}$ (resp. $\mathscr{C}(\text{C}) = \text{G}$, $\mathscr{C}(\text{G}) = \text{C}$, $\mathscr{C}(\text{T}) = \text{A}$) 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 $w = l_1 l_2 l_3$, $l_1, l_1, l_3 \in \{\text{A, C, G, T}\}$, 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.

$\mathscr{X}_a = \{\text{RRY, RYY}\} = \text{RNY}$ is a circular code (demonstrated in Introduction). $\mathscr{X}_a$ is a maximal (two trinucleotides) circular code and corresponds to the RNY model (Eigen & Schuster, 1978). $\mathscr{X}_a$ is self-complementary (complementary maximal circular code), i.e. $\mathscr{C}(\mathscr{X}_a) = \mathscr{X}_a$, as RRY and RYY are complementary. Any subset of $\mathscr{X}_a$ 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 $\mathscr{X}_b = \mathscr{P}(\mathscr{X}_a) = \{\text{RYR, YYR}\}$ and $\mathscr{X}_c = \mathscr{P}(\mathscr{X}_b) = \{\text{YRR, YRY}\}$ obtained by circular permutations of $\mathscr{X}_a$ are also maximal circular codes (identical proof). $\mathscr{X}_b$ and $\mathscr{X}_c$ are complementary to each other, i.e. $\mathscr{C}(\mathscr{X}_b) = \mathscr{X}_c$ and $\mathscr{C}(\mathscr{X}_c) = \mathscr{X}_b$, as RYR (resp. YYR) and YRY (resp. YRR) are complementary. The previous results remain unchanged by substituting R by Y and reciprocally. Therefore, $\mathscr{X}_d = \{\text{YRR, YYR}\}$ is a maximal complementary circular code ($\mathscr{C}(\mathscr{X}_d) = \mathscr{X}_d$) whose two subsets $\mathscr{X}_e = \mathscr{P}(\mathscr{X}_d) = \{\text{RRY, YRY}\}$ and $\mathscr{X}_f = \mathscr{P}(\mathscr{X}_e) = \{\text{RYR, RYY}\}$ obtained by circular permutations of $\mathscr{X}_d$ are also maximal circular codes and complementary to each other ($\mathscr{C}(\mathscr{X}_e) = \mathscr{X}_f$ and $\mathscr{C}(\mathscr{X}_f) = \mathscr{X}_e$). The three remaining sets $\mathscr{X}$ are $\mathscr{X}_g = \{\text{RYY, YRR}\}$ and $\mathscr{X}_h = \mathscr{P}(\mathscr{X}_g) = \mathscr{C}(\mathscr{X}_g) = \{\text{RRY, YYR}\}$ which

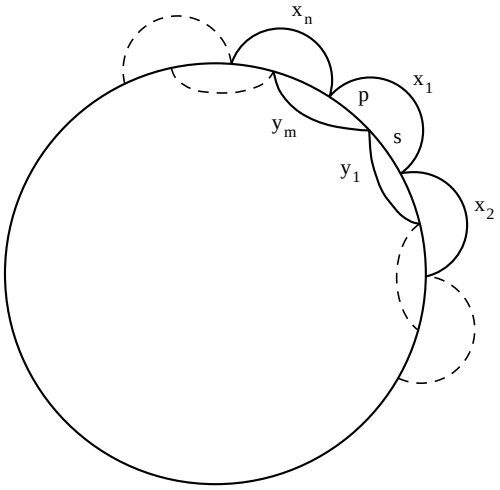

FIG. 1. A representation of the definition of a circular code.

are maximal circular codes and $\mathscr{X}_i = \mathscr{P}(\mathscr{X}_h) = \{RYR, YRY\}$ which is not a circular code as the word ... RYRYRYRYR ... has two factorizations over $\mathscr{X}_i$: ... RYR, YRY, RYR, ... and ... R, YRY, RYR, YR ...

In summary, on $\mathscr{B}_2$, eight of nine sets $\mathscr{X}$ are maximal circular codes and two sets $\mathscr{X}_a = \{RRY, RYY\} = RNY$ and $\mathscr{X}_d = \{YRR, YYR\} = YNR$ are complementary maximal circular codes with two permutated maximal circular codes and called complementary C^3 codes (Table 4).

3.3.3. *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 for 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. 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 circular codes determined by the classical methods of automatic construction, have constraints with 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 2: The subset $\mathscr{X}_0$(MIT) is a maximal (20 trinucleotides) circular code. The subsets $\mathscr{X}_1$(MIT) and $\mathscr{X}_2$(MIT) are also maximal circular codes.

Proof:

(i) Proof of the maximum cardinal of a circular code:
The 60 words of $\mathscr{B}_4' = \{AAA, \ldots, TTT\} - \{AAA, CCC, GGG, TTT\}$ are gathered in $\text{Card}(\mathscr{B}_4')/3 = 20$ classes invariant by circular permutation (Table 4). A circular code with words of length 3 on $\mathscr{B}_4$ has at most one word in each class and then contains at most 20 words.

(ii) Proof that $\mathscr{X}_0$(MIT) is a circular code:
As on $\mathscr{B}_4$ there are $3^{(b^3-b)/3} = 3^{20} = 3486784401$ potential maximal circular codes (Table 4), 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 $\mathscr{A}$ 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).

TABLE 4
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 circular codes with two permutated maximal circular codes (C^3 codes)	6	221544
Probability of C^3 codes	0.67	6.3×10^{-5}
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 maximal circular codes (complementary C^3 codes)	2	216
Probability of complementary C^3 codes	0.22	6.2×10^{-8}

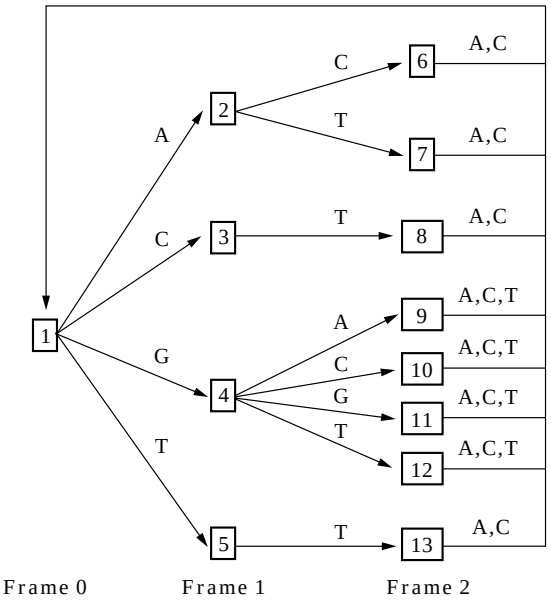

FIG. 2. Flower automaton $\mathscr{F}(\mathscr{X}_0(\text{MIT}))$ associated with the circular code $\mathscr{X}_0(\text{MIT})$.

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

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

Figure 2 gives the flower automaton $\mathscr{F}(\mathscr{X}_0(\text{MIT}))$ associated with $\mathscr{X}_0(\text{MIT})$. To prove that "$\mathscr{X}_0(\text{MIT})$ is a circular code" is equivalent to prove that "$\mathscr{F}(\mathscr{X}_0(\text{MIT}))$ is local", i.e. $\mathscr{F}(\mathscr{X}_0(\text{MIT}))$ 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 $\mathscr{X}$ of $\mathscr{B}_4^+$ verifying the definition of circular code and allows statistics with circular codes on $\mathscr{B}_4$ (Section 3.5).

(iii) Proof that $\mathscr{X}_1(\text{MIT})$ and $\mathscr{X}_2(\text{MIT})$ are circular codes:

Similar to (ii) by constructing the flower automata $\mathscr{F}(\mathscr{X}_1(\text{MIT}))$ and $\mathscr{F}(\mathscr{X}_2(\text{MIT}))$ associated with $\mathscr{X}_1$ (MIT) and $\mathscr{X}_2(\text{MIT})$ respectively.

Remark: the property that $\mathscr{X}_0(\text{MIT})$ is a circular code, does not necessarily imply that $\mathscr{X}_1(\text{MIT})$ and $\mathscr{X}_2(\text{MIT})$ are also circular codes (see e.g. $\mathscr{X}_g$, $\mathscr{X}_h$ and $\mathscr{X}_i$ in Section 3.3.2 and also the statistics in Table 4).

In summary, properties 1 and 2 imply that $\mathscr{X}_0(\text{MIT})$ is a maximal circular code with two permutated maximal circular codes (called C^3 code).

Remark: in contrast to the complementary C^3 code $\mathscr{X}_0(\text{EUK, PRO})$ identified in the protein genes of eukaryotes/prokaryotes (detailed in Section 4.4.1), $\mathscr{X}_0(\text{MIT})$ is not self-complementary. Indeed, for example, $\text{ACA} \in \mathscr{X}_0(\text{MIT})$ but $\mathscr{C}(\text{ACA}) = \text{TGT} \in \mathscr{X}_2(\text{MIT})$ (Table 2).

3.4. AUTOMATIC FRAME DETERMINATION PROPERTY

Property 3: the length of the minimal window to automatically retrieve frame 0 with $\mathscr{X}_0(\text{MIT})$ is equal to five nucleotides. The lengths of the minimal windows to automatically retrieve frame 1 with $\mathscr{X}_1$ (MIT) and frame 2 with $\mathscr{X}_2(\text{MIT})$ are equal to seven and six nucleotides respectively.

(i) Explanation of the length of the minimal window with $\mathscr{X}_0(\text{MIT})$:

The problem to determine automatically the decomposition of a word into trinucleotides of $\mathscr{X}_0(\text{MIT})$ is introduced with an example. Indeed, the unicity of such a decomposition is not obvious. For example, the word $w' = \text{GATT}$ of length 4 can be decomposed into trinucleotides of $\mathscr{X}_0(\text{MIT})$ in two ways: GAT, Tl_1l_2 [$\text{GAT}, \text{Tl}_1\text{l}_2 \in \mathscr{X}_0(\text{MIT})$; Table 2] or $\text{l}_3\text{GA}, \text{TTl}_4$ [$\text{l}_3\text{GA}, \text{TTl}_4 \in \mathscr{X}_0(\text{MIT})$; Table 2]. In fact, the automatic frame determination property is a consequence of the circular code property. If a word is constructed by concatenating words of $\mathscr{X}_0(\text{MIT})$ and if the frame of construction is lost, then the property of code assures that the frame can be retrieved in an unique way. Such a decomposition is called the reading frame of the word according to the code $\mathscr{X}_0(\text{MIT})$.

The unicity of such a decomposition is proved by using the properties of the associated flower automaton $\mathscr{F}(\mathscr{X}_0(\text{MIT}))$. As all words of $\mathscr{X}_0(\text{MIT})$ have a length of 3 (Fig. 2), the states of the automaton $\mathscr{F}(\mathscr{X}_0(\text{MIT}))$ can be associated with frames according to the code $\mathscr{X}_0(\text{MIT})$, state 1 with frame 0, states 2–5 with frame 1 and states 6–13 with frame 2. If any first letter of a word, obtained by a concatenation of trinucleotides of $\mathscr{X}_0(\text{MIT})$, can be associated with a unique state of the automaton $\mathscr{F}(\mathscr{X}_0(\text{MIT}))$, then a frame can be deduced for this letter because the associated unique state of $\mathscr{F}(\mathscr{X}_0(\text{MIT}))$ is related to a given frame. As a consequence, the word can be decomposed into trinucleotides of $\mathscr{X}_0(\text{MIT})$: its reading frame according to $\mathscr{X}_0(\text{MIT})$ is then retrieved. Then, the problem consists in identifying such a unique state for a letter of the word.

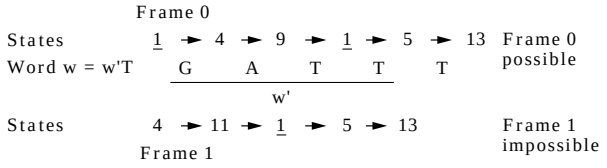

FIG. 3. An example of the automatic frame determination of a word of length $\geqslant 5$ with the flower automaton $\mathscr{F}(\mathscr{X}_0(\mathrm{MIT}))$ (Fig. 2). The states 1 are associated with the position of the commas. The word w' can be decomposed into trinucleotides of $\mathscr{X}_0(\mathrm{MIT})$ in two ways. The word $w = w'\mathrm{T}$ has a unique decomposition into trinucleotides of $\mathscr{X}_0(\mathrm{MIT})$.

Such a unicity is not obvious. In the previous example, the factor $w' = \mathrm{GATT}$ of length 4 can be attributed to two reading frames according to $\mathscr{X}_0(\mathrm{MIT})$: frame 0 (initial state 1 of w' in $\mathscr{F}(\mathscr{X}_0(\mathrm{MIT}))$) or frame 1 (initial state 4 of w' in $\mathscr{F}(\mathscr{X}_0(\mathrm{MIT}))$) (Fig. 3).

In the case of a local automaton $\mathscr{A}$, Definition 1 asserts that n exists 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 $\mathscr{X}_0(\mathrm{MIT})$ can be determined. For $\mathscr{A} = \mathscr{F}(\mathscr{X}_0(\mathrm{MIT}))$, n is equal to five and $\mathscr{F}(\mathscr{X}_0(\mathrm{MIT}))$ is called a five-local automaton. In other words, the length of the minimal window to retrieve always the frame is five letters with $\mathscr{X}_0(\mathrm{MIT})$. In the previous example, the first letter of the word $w = w'\mathrm{T}$ of length 5 is attributed to the unique frame, the frame 0, as there is no edge labelled T leaving the state 13 of $\mathscr{F}(\mathscr{X}_0(\mathrm{MIT}))$ (Figs 2 and 3). Then, the unique decomposition of w according to $\mathscr{X}_0(\mathrm{MIT})$ is GAT, TT (Fig. 3).

The length n of the minimal window to automatically retrieve the reading frame of a word according to $\mathscr{X}_0(\mathrm{MIT})$ 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 3.6).

(ii) Explanation of the lengths of the minimal windows with $\mathscr{X}_1(\mathrm{MIT})$ and $\mathscr{X}_2(\mathrm{MIT})$:

Similar to (i) with the flower automata $\mathscr{F}(\mathscr{X}_1(\mathrm{MIT}))$ and $\mathscr{F}(\mathscr{X}_2(\mathrm{MIT}))$ associated with $\mathscr{X}_1(\mathrm{MIT})$ and $\mathscr{X}_2(\mathrm{MIT})$ respectively.

3.5. RARITY PROPERTY

Property 4: the occurrence probability of $\mathscr{X}_0(\mathrm{MIT})$ is equal to 6.3×10^{-5}. Statistics concerning circular codes with trinucleotides on the four-letter alphabet $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ allow us to determine the occurrence probability of the code $\mathscr{X}_0(\mathrm{MIT})$. There are $3^{20} = 3486784401$ potential maximal circular codes (Section 3.3.3). The number of maximal circular codes computed with an algorithm verifying the definition of circular code among the 3^{20} circular codes, is 12964440. The computed number of maximal circular codes with two permutated maximal circular codes (C^3 codes), is 221544. Therefore, the probability to have a C^3 code, e.g. $\mathscr{X}_0(\mathrm{MIT})$, is $221544/3^{20} = 6.3 \times 10^{-5}$. This 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 4 summarizes and gives different circular code statistics on the alphabets $\{\mathrm{R}, \mathrm{Y}\}$ and $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.

3.6. CONCATENATION PROPERTY

The C^3 code $\mathscr{X}_0(\mathrm{MIT})$ identified in mitochondrial protein genes has some concatenation properties compared to other C^3 codes:

Property 5: $\mathscr{X}_0(\mathrm{MIT})$ has the second lowest length (five nucleotides) of minimal windows among the 221544 C^3 codes. Such a code occurs with a low probability (<0.005) among the 221544 C^3 codes (Table 5).

Table 5 gives the different lengths of minimal windows to retrieve the frame in the 221544 C^3 codes and their associated probabilities.

Property 6: $\mathscr{X}_0(\mathrm{MIT})$ has a low frequency (12% in average) of misplaced trinucleotides in the shifted frames. The circular permutation property implies that the concatenation of two trinucleotides of $\mathscr{X}_0(\mathrm{MIT})$ generates with a high probability a trinucleotide of $\mathscr{X}_1(\mathrm{MIT})$ in frame 1 and a trinucleotide of $\mathscr{X}_2(\mathrm{MIT})$ in frame 2. For $\mathscr{X}_0(\mathrm{MIT})$, this property is verified at 88%, i.e. there are 12% of misplaced trinucleotides in the shifted frames. The concatenation of two identical trinucleotides (process called duplication in biology) of $\mathscr{X}_0(\mathrm{MIT})$

TABLE 5
Lengths of minimal windows in the 221544 C^3 *codes and their associated probabilities*

Lengths of minimal windows	Number	Frequency (%) $= 100 \times$ number/221 544
4	192	0,1
5	744	0,3
6	2400	1,1
7	18 672	8,4
8	22 656	10,2
9	20 952	9,5
10	80 400	36,3
11	38 832	17,5
12	0	0
13	36 696	16,6
Sum	221 544	100

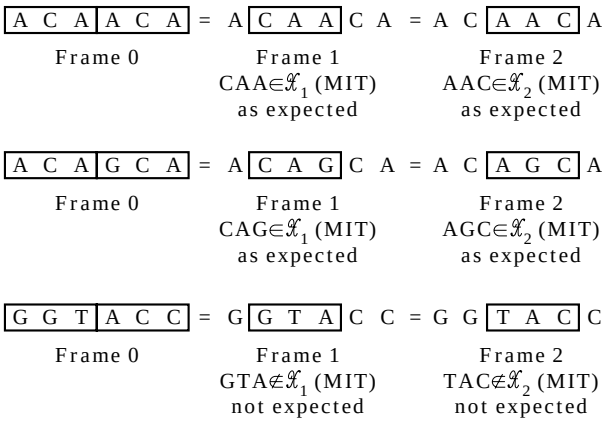

FIG. 4. Examples of three concatenations of two trinucleotides of the circular code $\mathscr{X}_0$(MIT) generating expected and unexpected trinucleotides in the shifted frames 1 and 2.

(e.g. ACA $\in \mathscr{X}_0$(MIT)) leads, by the circularity property, to the expected trinucleotides in the shifted frames [i.e. CAA $\in \mathscr{X}_1$(MIT) in frame 1 and AAC $\in \mathscr{X}_2$(MIT) in frame 2; first example in Fig. 4]. However, the probability of this type of concatenation is too low (1/20) to explain the circularity property. Otherwise, the concatenation of two different trinucleotides of $\mathscr{X}_0$(MIT) may lead to the expected trinucleotides in the shifted frames [e.g. ACA $\in \mathscr{X}_0$(MIT) and GCA $\in \mathscr{X}_0$(MIT) generate CAG $\in \mathscr{X}_1$(MIT) in frame 1 and AGC $\in \mathscr{X}_2$(MIT) in frame 2; second example in Fig. 4] or not [GGT $\in \mathscr{X}_0$(MIT) and ACC $\in \mathscr{X}_0$(MIT) generate GTA $\notin \mathscr{X}_1$(MIT) in frame 1 and TAC $\notin \mathscr{X}_2$(MIT) in frame 2; third example in Fig. 4]. The sum of percentages of trinucleotides of $\mathscr{X}_0$(MIT) (12.3%) and $\mathscr{X}_2$(MIT) (5.6%) found in the shifted frame 1 is equal to 17.9%. The sum of percentages of trinucleotides of $\mathscr{X}_0$(MIT) (0%) and $\mathscr{X}_1$(MIT) (6.1%) found in the shifted frame 2 is equal to 6.1%. The two sums are obviously different as $\mathscr{X}_0$(MIT) has no complementarity property. Therefore, the average percentage of misplaced trinucleotides in the shifted frames is equal to $(17.9 + 6.1)/2 = 12\%$.

Consequence: as the C^3 code $\mathscr{X}_0$(MIT) leads to 88% of well-placed trinucleotides in the shifted frames, the mitochondrial protein genes can be simulated with the trinucleotides associated with the frame 0 uniquely. Indeed, a simulated gene population generated by an independent mixing of the 20 trinucleotides of $\mathscr{X}_0$(MIT) with equiprobability (1/20) retrieves, with a very few exceptions, the two subsets $\mathscr{X}_1$(MIT) and $\mathscr{X}_2$(MIT) of trinucleotides in the frames 1 and 2 respectively (data not shown).

Property 7: $\mathscr{X}_0$(MIT) has an occurrence of the four types of nucleotides in the first and second trinucleotide sites but no nucleotide G in the third trinucleotide site.

4. Discussion

4.1. CONSEQUENCES ON THE TWO-LETTER ALPHABETS

The three subsets $\mathscr{T}_0$(MIT), $\mathscr{T}_1$(MIT) and $\mathscr{T}_2$(MIT) 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 6(a) shows that the trinucleotide RYY occurs preferentially in frame 0, the trinucleotide YYR, in frame 1, and the trinucleotide YRY, in frame 2. Indeed, RYY contains six A/C/G/T trinucleotides in frame 0 and two A/C/G/T trinucleotides in frame 2 (ACT^2, ATT^2). YYR contains six A/C/G/T trinucleotides in frame 1 and two A/C/G/T trinucleotides in frame 0 (CTA^0, TTA^0). YRY contains six A/C/G/T trinucleotides in frame 2 and two A/C/G/T trinucleotides in frame 1 (TAC^1, TAT^1). RYY is a non-maximal circular code with two permutated non-maximal circular codes YYR and YRY. RYY is complementary to the non-maximal circular code RRY corresponding to the RRY codon model (Crick *et al*., 1976; Section 3.3.2). RRY, RYR and YRR occur in two frames, frames 0 and 2 for RRY, frames 0 and 1 for RYR, and frames 1 and 2 for YRR [Table 6(a)]. There is no preferential frame for RRR and YYY [Table 6(a)].

These results also explain the results of different previous works analysing the three frames simultaneously (average frame) in mitochondrial protein genes with autocorrelation functions on the alphabet $\{R, Y\}$, in particular 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 of bases between them.

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 6(b) shows that the trinucleotide KKM occurs preferentially in frame 0, the trinucleotide KMK, in frame 1, and the trinucleotide MKK, in frame 2. On the alphabet $\{S, W\}$, Table 6(c) shows that the trinucleotides SSW and SWW occur preferentially in frame 0, SWS and WWS, in frame 1,

TABLE 6

(a) *Preferential occurrence frame for the trinucleotides on the two-letter alphabets*

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

The eight R/Y trinucleotides (R = purine = A or G, Y = pyrimdine = C or T) are associated with the 64 A/C/G/T trinucleotides by considering their frame ($\mathscr{T}_0$(MIT), $\mathscr{T}_1$(MIT), $\mathscr{T}_2$(MIT)). The last row gives the preferential occurrence frame for the R/Y trinucleotides, e.g. RRR is in the 3 frames (3 A/C/G/T trinucleotides in frame 0, 2 in frame 1, 3 in frame 2), RYY is in frame 0 (6 A/C/G/T trinucleotides in frame 0, 0 in frame 1, 2 in frame 2).

(b)

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

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 ($\mathscr{T}_0$(MIT), $\mathscr{T}_1$(MIT), $\mathscr{T}_2$(MIT)). The last row gives the preferential occurrence frame for the K/M trinucleotides.

(c)

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

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 ($\mathscr{T}_0$(MIT), $\mathscr{T}_1$(MIT), $\mathscr{T}_2$(MIT)). The last row gives the preferential occurrence frame for the S/W trinucleotides.

and WSS and WSW, in frame 2. KKM (resp. $\{$SSW, SWW$\}$) is a non-maximal (resp. maximal) circular code with two permutated non-maximal (resp. maximal) circular codes.

There is no obvious two-letter alphabet close to the alphabet $\{$A, C, G, T$\}$: the A/C/G/T trinucleotides in frame 0 are spread on the eight trinucleotides of each two-letter alphabet, RYY and KKM in frame 0 contain each six A/C/G/T trinucleotides in frame 0, RYR and KMM in frames 0 and 1 contain each four A/C/G/T trinucleotides in frame 0, RRY and MKM in frames 0 and 2 contain each four A/C/G/T trinucleotides in frame 0, and SSW and SWW have a different number of A/C/G/T trinucleotides in frame 0, four and five respectively [Tables 6(a–c)].

4.2. CONSEQUENCES ON THE MITOCHONDRIAL GENETIC CODE

The mitochondrial genetic code has several code changes compared to the universal genetic code. The

TABLE 7

Amino acids coded by the three subsets $\mathscr{X}_0$(MIT), $\mathscr{X}_1$(MIT) and $\mathscr{X}_2$(MIT) according to the mitochondrial genetic code

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

The subset $\mathscr{X}_0$(MIT) (bold) codes for 10 amino acids in the mitochondrial genetic code: Ala, Asp, Glu, Gly, Ile, Leu, Met, Phe, Thr and Val. 10 amino acids are not coded by $\mathscr{X}_0$(MIT): Arg, Asn, Cys, Gln, His, Lys, Pro, Ser, Trp and Tyr. The subset $\mathscr{X}_1$(MIT) (underlined one time) is associated with 11 amino acids: Ala, Gln, Glu, Leu, Lys, Met, Pro, Ser, Thr, Tyr and Val. The subset $\mathscr{X}_2$(MIT) (underlined two times) is associated with 10 amino acids: Arg, Asn, Cys, His, Ile, Leu, Pro, Ser, Thr and Trp. The codons AAA, CCC, GGG and TTT are in addition in italic.

two most common changes are TGA from stop to Trp and ATA from Ile to Met (Osawa *et al*., 1992). The codon subset $\mathscr{X}_0$(MIT) codes for 10 amino acids (AA): Ala, Asp, Glu, Gly, Ile, Leu, Met, Phe, Thr and Val (Table 7). Note that $\mathscr{X}_0$(MIT) and $\mathscr{T}_0$(MIT) code the same number of AA (GGG $\in \mathscr{T}_0$(MIT) codes for Gly). As $\mathscr{X}_0$(MIT) has 20 codons, several codons of $\mathscr{X}_0$(MIT) code for the same AA. The subsets $\mathscr{X}_1$(MIT) and $\mathscr{X}_2$(MIT) are associated with 11 and 10 AA respectively (Table 7). The 10 AA coded by $\mathscr{X}_0$(MIT) are equally represented in the two classes of aminoacyl-tRNA synthetases with a class 1 associated with Glu, Ile, Leu, Met and Val, and a class 2, with Ala, Asp, Gly, Phe and Thr [reviewed in Schimmel *et al*. (1993), Hartman (1995) and Saks & Sampson (1995)].

4.3. CONSEQUENCES ON THE AMINO ACID FREQUENCIES IN MITOCHONDRIAL PROTEINS

There are 10 amino acids (AA) which are not coded by the codon subset $\mathscr{X}_0$(MIT): Arg, Asn, Cys, Gln, His, Lys, Pro, Ser, Trp and Tyr (Table 7). Therefore, these 10 AA should have the lowest frequencies in mitochondrial proteins. In order to verify this consequence of a code in mitochondrial protein genes, the frequencies of the 20 AA are computed in 2304 mitochondrial proteins (715660 AA). They are obtained from the protein data base SWISS-PROT (release 34 in October 1996). 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. 62). Table 8 shows that six of 10 AA not coded by $\mathscr{X}_0$(MIT), Arg, Cys, His, Pro, Ser and Trp, have the lowest observed/expected frequency ratios (<0.8) in the mitochondrial proteins. 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). Gln which is not coded by $\mathscr{X}_0$(MIT), has a frequency ratio <1.0 (Table 8). The three exceptions, Asn, Lys and Tyr, not coded

TABLE 8

Correlation between the usage of the subset $\mathscr{X}_0$(MIT) in mitochondrial protein genes and the amino acid frequencies in mitochondrial proteins

Amino acid (number of codons)	Expected frequency (% rounded)	Observed number in mitochondrial proteins	Observed frequency (%) in mitochondrial proteins	Observed frequency/ Expected frequency
Ala, A, Alanine (4/62)	6.45	52 184	7.29	1.13
Arg, R, Arginine (6/62)	9.68	30 547	4.27	0.44
Asn, N, Asparagine (2/62)	3.23	31 260	4.37	1.35
Asp, D, Aspartic acid (2/62)	3.23	28 837	4.03	1.25
Cys, C, Cysteine (2/62)	3.23	9 179	1.28	0.40
Gln, Q, Glutamine (2/62)	3.23	22 601	3.16	0.98
Glu, E, Glutamic acid (2/62)	3.23	34 138	4.77	1.48
Gly, G, Glycine (4/62)	6.45	49 250	6.88	1.07
His, H, Histidine (2/62)	3.23	16 347	2.28	0.71
Ile, I, Isoleucine (2/62)	3.23	49 609	6.93	2.15
Leu, L, Leucine (6/62)	9.68	80 786	11.29	1.17
Lys, K, Lysine (2/62)	3.23	39 937	5.58	1.73
Met, M, Methionine (2/62)	3.23	23 211	3.24	1.01
Phe, F, Phenylalanine (2/62)	3.23	38 725	5.41	1.68
Pro, P, Proline (4/62)	6.45	33 486	4.68	0.73
Ser, S, Serine (6/62)	9.68	51 643	7.22	0.75
Thr, T, Threonine (4/62)	6.45	41 069	5.74	0.89
Trp, W, Tryptophan (2/62)	3.23	11 449	1.60	0.50
Tyr, Y, Tyrosine (2/62)	3.23	25 446	3.56	1.10
Val, V, Valine (4/62)	6.45	45 956	6.42	1.00

Arg, Cys, His, Pro, Ser and Trp have the lowest observed/expected frequency ratios (<0.8) in the proteins of mitochondria (2304 sequences, 715660 amino acids) as expected with the usage of the trinucleotides of $\mathscr{X}_0$(MIT) in the mitochondrial protein genes.

for by $\mathscr{X}_0$(MIT) with frequency ratios >1.0, may be explained by the various code changes existing in the gene subpopulations of mitochondria. Indeed, these three AA have different codon assignments, e.g. AAA may code for Asn (instead of Lys) in mitochondria of echinoderms and platyhelminths, TAA, instead to be a stop codon, may code for Tyr in mitochondria of platyhelminths (Osawa *et al*., 1992). A greater number of sequences in these various subpopulations should improve this statistical analysis in future.

In summary, in mitochondria, there is a correlation (however not perfect) between the usage of the trinucleotides of $\mathscr{X}_0$(MIT) in mitochondrial protein genes and the amino acid frequencies in mitochondrial proteins.

4.4. SIMILARITIES AND DIFFERENCES BETWEEN THE COMPLEMENTARY C^3 CODE OF EUKARYOTES/ PROKARYOTES AND THE C^3 CODE OF MITOCHONDRIA

4.4.1. *Recall concerning the complementary C^3 code of eukaryotes/prokaryotes* (*Arquès & Michel*, 1996)

The same three subsets of trinucleotides per frame have recently been identified in two large protein gene populations of eukaryotes EUK (26757 sequences, 11397678 trinucleotides) and prokaryotes PRO (13686 sequences, 4708758 trinucleotides): $\mathscr{T}_0(\text{EUK, PRO}) = \mathscr{X}_0(\text{EUK, PRO}) \cup \{\text{AAA, TTT}\}$, $\mathscr{T}_1(\text{EUK, PRO}) = \mathscr{X}_1(\text{EUK, PRO}) \cup \{\text{CCC}\}$ and $\mathscr{T}2(\text{EUK, PRO}) = \mathscr{X}_2(\text{EUK, PRO}) \cup \{\text{GGG}\}$ in frames 0, 1 and 2 respectively, with the three subsets $\mathscr{X}_0(\text{EUK, PRO})$, $\mathscr{X}_1(\text{EUK, PRO})$ and $\mathscr{X}_2(\text{EUK, PRO})$ of 20 trinucleotides defined in Table 9(a). The subset $\mathscr{X}_0(\text{EUK, PRO})$ associated with $\mathscr{X}_1(\text{EUK, PRO})$ and $\mathscr{X}_2(\text{EUK, PRO})$ has seven important properties [detailed in Arquès & Michel (1996)]:

(i) A maximal (20 trinucleotides) circular code for $\mathscr{X}_0(\text{EUK, PRO})$ [resp. $\mathscr{X}_1(\text{EUK, PRO})$ and $\mathscr{X}_2(\text{EUK, PRO})$] allowing to retrieve automatically the frame 0 (resp. 1, 2) in any region of a protein gene model [formed by a series of trinucleotides of $\mathscr{X}_0(\text{EUK, PRO})$] without a start codon.

(ii) Circular permutations $\mathscr{P}$ of $\mathscr{X}_0(\text{EUK, PRO})$ allowing us to deduce $\mathscr{X}_1(\text{EUK, PRO})$ and $\mathscr{X}_2(\text{EUK, PRO})$: $\mathscr{P}(\mathscr{X}_0(\text{EUK, PRO})) = \mathscr{X}_1(\text{EUK, PRO})$ and $\mathscr{P}(\mathscr{X}_1(\text{EUK, PRO})) = \mathscr{X}_2(\text{EUK, PRO})$ [Table 9(b)].

(iii) The DNA complementarity property $\mathscr{C}$:

$$\mathscr{C}(\mathscr{X}_0(\text{EUK, PRO})) = \mathscr{X}_0(\text{EUK, PRO})$$

($\mathscr{X}_0(\text{EUK, PRO})$ is self-complementary: 10 trinucleotides of $\mathscr{X}_0(\text{EUK, PRO})$ are complementary to the 10 other trinucleotides of $\mathscr{X}_0(\text{EUK, PRO})$), $\mathscr{C}(\mathscr{X}_1(\text{EUK, PRO})) = \mathscr{X}_2(\text{EUK, PRO})$ and $\mathscr{C}(\mathscr{X}_2(\text{EUK, PRO})) = \mathscr{X}_1(\text{EUK, PRO})$ [$\mathscr{X}_1(\text{EUK, PRO})$ and $\mathscr{X}_2(\text{EUK, PRO})$ are complementary to each other: the 20 trinucleotides of $\mathscr{X}_1(\text{EUK, PRO})$

Table 9

(a) *List per frame and in lexicopraphical order of the trinucleotides of the complementary C^3 code identified in protein coding genes of eukaryotes EUK and prokaryotes PRO (Arquès & Michel, 1996)*

$\mathscr{T}_0$(EUK, PRO):	AAA	AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC	TTT
$\mathscr{T}_1$(EUK, PRO):	AAG	ACA	ACG	ACT	AGC	AGG	ATA	ATG	CCA	CCC	CCG	GCG	GTG	TAG	TCA	TCC	TCG	TCT	TGC	TTA	TTG	
$\mathscr{T}_2$(EUK, PRO):	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: $\mathscr{T}_0$(EUK, PRO) = $\mathscr{X}_0$(EUK, PRO) $\cup$ {AAA, TTT} in frame 0, $\mathscr{T}_1$(EUK, PRO) = $\mathscr{X}_1$(EUK, PRO) $\cup$ {CCC} in frame 1 and $\mathscr{T}_2$(EUK, PRO) = $\mathscr{X}_2$(EUK, PRO) $\cup$ {GGG} in frame 2.

(b) *Circularity property with the three subsets $\mathscr{X}_0$(EUK, PRO), $\mathscr{X}_1$(EUK, PRO) and $\mathscr{X}_2$(EUK, PRO) of trinucleotides of the complementary C^3 code identified in protein coding genes of eukaryotes EUK and prokaryotes PRO [Table 9(a)]*

$\mathscr{X}_0$(EUK, PRO):	AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC
$\mathscr{X}_1$(EUK, PRO):	ACA	ATA	CCA	TCA	TTA	AGC	TCC	TGC	AAG	ACG	AGG	ATG	CCG	GCG	GTG	TAG	TCG	TTG	ACT	TCT
$\mathscr{X}_2$(EUK, PRO):	CAA	TAA	CAC	CAT	TAT	GCA	CCT	GCT	AGA	CGA	GGA	TGA	CGC	CGG	TGG	AGT	CGT	TGT	CTA	CTT

(c) *Complementarity property with the three sets $\mathscr{T}_0$(EUK, PRO), $\mathscr{T}_1$(EUK, PRO) and $\mathscr{T}_2$(EUK, PRO) of trinucleotides of the complementary C^3 code identified in protein coding genes of eukaryotes EUK and prokaryotes PRO [Table 9(a)]*

$\mathscr{T}_0$(EUK, PRO):	AAA	AAC	AAT	ACC	ATC	CAG	CTC	GAA	GAC	GCC	GTA								
$\mathscr{T}_0$(EUK, PRO):	TTT	GTT	ATT	GGT	GAT	CTG	GAG	TTC	GTC	GGC	TAC								
$\mathscr{T}_1$(EUK, PRO):	AAG	ACA	ACG	ACT	AGC	AGG	ATA	ATG	CCA	CCC	CCG	GCG	GTG	TAG	TCA	TCC	TCG	TCT	TGC	TTA	TTG
$\mathscr{T}_2$(EUK, PRO):	CTT	TGT	CGT	AGT	GCT	CCT	TAT	CAT	TGG	GGG	CGG	CGC	CAC	CTA	TGA	GGA	CGA	AGA	GCA	TAA	CAA

are complementary to the 20 trinucleotides of $\mathscr{X}_2$(EUK, PRO); Table 9(c)].

(iv) A length of the minimal window to automatically retrieve frame 0 equal to 13 nucleotides.

(v) An occurrence probability equal to 6×10^{-8} (Table 4).

(vi) A high frequency (24.6%) of misplaced trinucleotides in the shifted frames (due to the complementarity property, the sum of percentages of trinucleotides of $\mathscr{X}_0$(EUK, PRO) (11.9%) and $\mathscr{X}_2$(EUK, PRO) (12.7%) found in the shifted frame 1 is equal to the sum of percentages of trinucleotides of $\mathscr{X}_0$(EUK, PRO) (11.9%) and $\mathscr{X}_1$(EUK, PRO) (12.7%) found in the shifted frame 2 and are equal to 24.6%).

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

In summary, the subset $\mathscr{X}_0$(EUK, PRO) of 20 trinucleotides identified in protein genes of eukaryotes/prokaryotes is a complementary maximal circular code with two permutated maximal circular codes (complementary C^3 code).

4.4.2. *Consequences on the amino acid coding*

The subset $\mathscr{X}_0$(EUK, PRO) codes for 12 amino acids (AA): Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Phe, Thr, Tyr and Val (Arquès & Michel, 1996). The subset $\mathscr{X}_0$(MIT) codes only for 10 AA: Ala, Asp, Glu, Gly, Ile, Leu, Met, Phe, Thr and Val (Table 7). The nine AA commonly coded for by the two subsets are Ala, Asp, Glu, Gly, Ile, Leu, Phe, Thr and Val. There are three AA coded for by $\mathscr{X}_0$(EUK, PRO) and not coded for by $\mathscr{X}_0$(MIT): Asn, Gln and Tyr. There is one AA, Met, coded for by $\mathscr{X}_0$(MIT) and not coded for by $\mathscr{X}_0$(EUK, PRO). Six AA not coded for by the two subsets $\mathscr{X}_0$(EUK, PRO) and $\mathscr{X}_0$(MIT), Arg, Cys, His, Pro, Ser and Trp, have the lowest observed/expected frequency ratios (<0.8) in the proteins of both eukaryotes/prokaryotes (Arquès & Michel, 1996) and mitochondria (Table 8) as expected with the usage of these two codes in protein genes. Surprisingly, the only exception observed among the eight AA not coded for by $\mathscr{X}_0$(EUK, PRO) is Met (frequency ratio >1; Arquès & Michel, 1996) which is coded by $\mathscr{X}_0$(MIT) (Table 7).

4.4.3. *Consequences on the DNA double helix coding*

As the code $\mathscr{X}_0$(EUK, PRO) is self-complementary, the two paired frames 0 in the DNA double helix may simultaneously code for amino acids without a start codon. Furthermore, as $\mathscr{X}_1$(EUK, PRO) and $\mathscr{X}_2$(EUK, PRO) are also 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 DNA double helix, may simultaneously code for amino acids without a start codon, leading to an optimal information storage [detailed in Arquès & Michel (1996)]. For example, the insertion sequence IS1, i.e. a transposable element in bacteria, includes six potential reading frames with two genes encoded on the same strand of IS1 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 viral genome for maximizing the coding function (e.g. Ziff, 1980). The concept of proteins coded by complementary strands has also been investigated from properties between amino acids and codons and from rules of the two complementary strands of DNA (Zull & Smith, 1990; Konecny *et al*., 1993, 1995; Béland & Allen, 1994).

In contrast, the code $\mathscr{X}_0$(MIT) with no complementary property, does not allow that the two paired frames 0 in the DNA double helix simultaneously code for amino acids.

4.4.4. *Evolutionary consequences*

The identification of a subset $\mathscr{X}_0$ [$\mathscr{X}_0$(EUK, PRO) or $\mathscr{X}_0$(MIT)] of 20 trinucleotides occurring preferentially (i.e. in comparison with $\mathscr{X}_1$ and $\mathscr{X}_2$) in frame 0 (reading frame) of different taxonomic protein gene populations (eukaryotes, prokaryotes, mitochondria), suggests an evolution of protein genes according to two processes: a construction process of primitive protein genes followed by an evolutionary process transforming the primitive protein genes into actual protein genes. A model at three parameters (p, q, k) based on an independent mixing of the 20 trinucleotides of $\mathscr{X}_0$(EUK, PRO) with equiprobability (1/20) followed by $k \approx 5$ substitutions per codon in the three codon sites in proportions $p \approx 0.1$, $q \approx 0.1$ and $r = 1 - p - q \approx 0.8$ respectively, allows us to simulate an evolution of protein genes of eukaryotes/prokaryotes (Arquès *et al*., 1997). 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). A computational model also based on the construction and evolutionary processes, is currently in investigation for an evolutionary simulation of mitochondrial protein genes.

The complementary code $\mathscr{X}_0$(EUK, PRO) has flexibility properties compared to the code $\mathscr{X}_0$(MIT): a higher frequency of misplaced trinucleotides in the shifted frames (24.6% compared to 12%), a greater length of the minimal window to automatically retrieve the frame 0 (13 nucleotides compared to five nucleotides) and an occurrence of the four types of nucleotides in the three trinucleotide sites while $\mathscr{X}_0$(MIT) has no nucleotide G in the third trinucleotide site.

The two codes $\mathscr{X}_0$(EUK, PRO) and $\mathscr{X}_0$(MIT) have 13 trinucleotides in common (Table 10): ACC, ATC, CTC, GAA, GAC, GAT, GCC, GGC, GGT, GTA, GTC, GTT and TTC. Ten of these 13 trinucleotides are complementary to each other: $\mathscr{C}$(ACC) = GGT, $\mathscr{C}$(ATC) = GAT, $\mathscr{C}$(GAA) = TTC, $\mathscr{C}$(GAC) = GTC, $\mathscr{C}$(GCC) = GGC. Three trinucleotides are not complementary: CTC, GTA, GTT. The seven non-common trinucleotides with the two codes can be deduced by circular permutations. One permutation of three trinucleotides of $\mathscr{X}_0$(EUK, PRO) leads to three trinucleotides of $\mathscr{X}_0$(MIT): $\mathscr{P}$(AAC) = ACA, $\mathscr{P}$(AAT) = ATA, $\mathscr{P}$(ATT) = TTA and two permutations (or one permutation to the left) of four trinucleotides of $\mathscr{X}_0$(EUK, PRO) leads to four trinucleotides of $\mathscr{X}_0$(MIT): $\mathscr{P}(\mathscr{P}$(CAG)) = GCA, $\mathscr{P}(\mathscr{P}$(CTG)) = GCT, $\mathscr{P}(\mathscr{P}$(GAG)) = GGA, $\mathscr{P}(\mathscr{P}$(TAC)) = CTA. Only two of these seven non-common trinucleotides can be deduced by only one substitution, precisely at the third trinucleotide site: ATT $\in \mathscr{X}_0$(EUK, PRO) with ATA $\in \mathscr{X}_0$(MIT) and CTG $\in \mathscr{X}_0$(EUK, PRO) with CTA $\in \mathscr{X}_0$(MIT). Therefore, the link between the two codes seems to be related to permutations rather than substitutions.

4.4.5. *Consequences associated with the property to retrieve automatically the frame*

A protein gene formed by trinucleotides of a code $\mathscr{X}_0$ [$\mathscr{X}_0$(EUK, PRO) or $\mathscr{X}_0$(MIT)] has the property to automatically retrieve frame 0 in any region of the gene without a start codon. It would be interesting to know if a trace of the uselessness of the start codon for locating the reading frame may exist in the actual protein genes which have a preferential occurrence of trinucleotides of $\mathscr{X}_0$ (i.e. a protein gene formed by trinucleotides of $\mathscr{X}_0$, $\mathscr{X}_1$ and $\mathscr{X}_2$ in frame 0 with lower frequencies for $\mathscr{X}_1$ and $\mathscr{X}_2$). Almost all actual protein genes begin with the ATG start codon which does not belong to $\mathscr{X}_0$ [ATG belongs to $\mathscr{X}_1$ in both codes; Tables 2 and 9(a)]. 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 motif GCCGCCRCCATG (Kozak, 1989). Surprisingly,

TABLE 10

Trinucleotides common (bold) to the complementary C^3 code $\mathscr{X}_0$(EUK, PRO) identified in protein coding genes of eukaryotes EUK and prokaryotes PRO [Table 9(a)] and to the C^3 code $\mathscr{X}_0$(MIT) identified in protein coding genes of mitochondria MIT (Table 2)

$\mathscr{X}_0$(EUK, PRO):	AAC	AAT	**ACC**	**ATC**	ATT	CAG	**CTC**	CTG	**GAA**	**GAC**	GAG	**GAT**	**GCC**	**GGC**	**GGT**	**GTA**	**GTC**	**GTT**	TAC	**TTC**
$\mathscr{X}_0$(MIT):	ACA	**ACC**	ATA	**ATC**	CTA	**CTC**	**GAA**	**GAC**	**GAT**	GCA	**GCC**	GCT	GGA	**GGC**	**GGT**	**GTA**	**GTC**	**GTT**	TTA	**TTC**

the three trinucleotides preceding the ATG start codon belong to $\mathscr{X}_0$ in both codes [ACC, GCC $\in \mathscr{X}_0$(EUK, PRO), $\mathscr{X}_0$(MIT); Tables 2 and 9(a)]. The occurrence probability of such a series is equal to $1/3^3 \approx 0.04$. Therefore, this motif of nine base length could have been the translation initiation signal in primitive protein genes (with only $\mathscr{X}_0$ 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 this motif (the base R three bases before ATG changed in C) drastically impairs translation initiation of α-globin (Morlé *et al*., 1985). Note that YCC, i.e. CCC and TCC, belongs to $\mathscr{T}_1$ in both codes [Tables 2 and 9(a)]. Otherwise, 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 motif [reviewed in 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 [reviewed in Kozak (1989) and Geballe & Morris (1994)]. Finally, the length of the consensus motif GCCGCCRCC could be considered according to the windows of nucleotides retrieving automatically the frame 0 with the two codes $\mathscr{X}_0$ (see also below).

Similarly, it would be interesting to know if a trace of the property to retrieve automatically the frame 0 in any region of the gene may exist in the actual protein genes. 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 [reviewed in Farabaugh *et al*. (1993)], the retrotransposon *TYA* and *TYB* genes (Belcourt & Farabaugh, 1990); (ii) regulating gene expression, e.g. the *E. coli prfB* gene (e.g. Craigen & 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 *E. coli dnaX* gene (e.g. Tsuchihashi & Brown, 1992) and the prokaryotic insertion sequences, e.g. IS1 (Machida *et al*., 1984), IS150 (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 $\geqslant$two bases in the $5' \rightarrow 3'$ direction (downstream direction) or in the $3' \rightarrow 5'$ 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 one time) belongs to $\mathscr{X}_1$ in both codes and the landing trinucleotide (underlined two times), to $\mathscr{X}_0$ in both codes: GTGTA [GTG $\in \mathscr{X}_1$(EUK, PRO), $\mathscr{X}_1$(MIT); TGT $\in \mathscr{X}_2$(EUK, PRO), $\mathscr{X}_2$(MIT); GTA $\in \mathscr{X}_0$(EUK, PRO), $\mathscr{X}_0$(MIT); Tables 2 and 9(a)] and GTGTAAGTT [GTG $\in \mathscr{X}_1$(EUK, PRO), $\mathscr{X}_1$(MIT); TGT $\in \mathscr{X}_2$(EUK, PRO), $\mathscr{X}_2$(MIT); GTA $\in \mathscr{X}_0$(EUK, PRO), $\mathscr{X}_0$(MIT); TAA $\in \mathscr{X}_2$(EUK, PRO), $\mathscr{X}_1$(MIT); AAG $\in \mathscr{X}_1$(EUK, PRO), $\mathscr{X}_1$(MIT); AGT $\in \mathscr{X}_2$(EUK, PRO), $\mathscr{X}_2$(MIT); GTT $\in \mathscr{X}_0$(EUK, PRO), $\mathscr{X}_0$(MIT); Tables 2 and 9(a)]. In the first example, the take-off trinucleotide GTG and TGT are not translated and belong to $\mathscr{X}_1$ and $\mathscr{X}_2$ respectively in both codes. The translated trinucleotide GTA belongs to $\mathscr{X}_0$ in both codes. The second example appears to be an extension of the rule applied previously. If the first landing trinucleotide belonging to $\mathscr{X}_0$ in both codes (GTA) is not chosen, then the second landing trinucleotide considered is the next downstream trinucleotide belonging to $\mathscr{X}_0$ in both codes (GTT). All the other non-translated trinucleotides belong to $\mathscr{X}_1$ and $\mathscr{X}_2$, but not to $\mathscr{X}_0$, even for a given code. Note that TAA belongs to $\mathscr{X}_2$ in the code of eukaryotes/prokaryotes but to $\mathscr{X}_1$ in the code of mitochondria.

Slipping can be defined as a translational shift of one base either in the $3' \rightarrow 5'$ direction (-1 frameshift) or in the $5' \rightarrow 3'$ direction ($+1$ frameshift). Several trinucleotides which may direct -1 frameshift, have been reported, e.g. AAG (Tsuchihashi & Brown, 1992; Lindsley & Gallant, 1993; Vögele *et al*., 1991) and CCG (Dayhuff *et al*., 1986). As these trinucleotides belong to $\mathscr{X}_1$ in both codes [Tables 2 and 9(a)] and as $\mathscr{X}_1$ is deduced by one circular permutation of $\mathscr{X}_0$, 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 & Caskey, 1986; Curran, 1993), may provoke $+1$ frameshift. As these trinucleotides belong to $\mathscr{X}_2$ in both codes [Tables 2 and 9(a)] and as $\mathscr{X}_2$ is deduced by two circular permutations of $\mathscr{X}_0$, 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 & Weiss-Brummer, 1980) may induce both -1 and $+1$ frameshifts. Therefore, the trinucleotides belonging to $\mathscr{T}_0$, $\mathscr{T}_1$ and $\mathscr{T}_2$ exclusively 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 analysed according to the subsets $\mathscr{X}_0$, $\mathscr{X}_1$ and $\mathscr{X}_2$ of trinucleotides in order to identify the frameshift rules encoded in the DNA sequence. As $\mathscr{X}_0$, $\mathscr{X}_1$ and $\mathscr{X}_2$ 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 or 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.

The translational frameshifting in the actual protein genes has been considered in the previous section according to the words of the two circular codes $\mathscr{X}_0$. In this section, this frameshifting process is also analysed with respect to the windows of nucleotides automatically retrieving frame 0 of the two codes $\mathscr{X}_0$. The windows are equal to 13 nucleotides in the code $\mathscr{X}_0$(EUK, PRO) and five nucleotides in the code $\mathscr{X}_0$(MIT). These two lengths automatically retrieve frame 0 in the worst case, i.e. smaller lengths, depending on the type of words, can also retrieve automatically the frame 0. In the hopping process, the windows of the two codes $\mathscr{X}_0$ could be related to the sequences 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). Pseudknots occur in an average six nucleotides downstream ($+6$) of 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 frameshift trinucleotides can modify the rate of frameshifting (Brierley *et al*., 1992). The windows of

the two codes $\mathscr{X}_0$ could be connected with these sequences.

In the -1 slipping process of prokaryotes, a secondary structure, usually a hairpin, and a Shine–Dalgarno site (Shine & Dalgarno, 1974) are often involved in frameshifting (Vögele *et al*., 1991; Larsen *et al*., 1994). The Shine–Dalgarno sites occur in an average 14 nucleotides upstream (-14) of frameshift trinucleotides (Larsen *et al*., 1994), and the hairpin, seven nucleotides downstream ($+7$) of frameshift trinucleotides (Vögele *et al*., 1991). If the spacing between the Shine–Dalgarno site and the frameshift trinucleotides is reduced to seven nucleotides then $+1$ frameshift may occur (Larsen *et al*., 1994). Notes: the Shine–Dalgarno site is 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 trinucleotides in the $+1$ slipping process of the ornithine decarboxylase antizyme gene of eukaryotes (Matsufuji *et al*., 1995). Similarly, the windows of the two codes $\mathscr{X}_0$ could be involved in these different sequences of critical lengths (Larsen *et al*., 1994).

5. Conclusion

After the identification of a complementary C^3 code $\mathscr{X}_0$(EUK, PRO) in protein genes of eukaryotes and prokaryotes, a new C^3 code $\mathscr{X}_0$(MIT) is discovered in protein genes of mitochondria. The exceptional properties of these two codes as well as their different biological consequences, suggest that these codes could have had a function in gene evolution and that the primitive alphabet could have been $\{A, C, G, T\}$ rather than $\{R, Y\}$.

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. (1994). Analytical expression of the purine/pyrimidine autocorrelation function after and before random mutations. *Math. Biosc.* **123,** 103–125.

Arquès, D. G. & Michel, C. J. (1996). A complementary circular code in the protein coding genes. *J. theor. Biol.* **182,** 45–58.

Arquès, D. G., Fallot, J.-P. & Michel, C. J. (1997). An evolutionary model of a complementary circular code. *J. theor. Biol.* **185,** 241–253.

Béal, M.-P. (1993). *Codage symbolique*. New York: Masson.

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 7 nucleotide minimal site. *Cell* **62,** 339–352.

Berstel, J. & Perrin, D. (1985). *Theory of Codes*. London: Academic Press.

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.

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., Griffith, J. S. & Orgel, L. E. (1957). Codes without commas. *Proc. Natl. Acad. Sci. U.S.A.* **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.

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

Dayhuff, T., Atkins, J. & Gesteland, R. (1986). Characterization of ribosomal frameshift 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 *POL*3 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.

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

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

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. U.S.A.* **90,** 5469–5473.

Machida, Y., Machida, C. & Ohtsubo, E. (1984). Insertion element IS1 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. U.S.A.* **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.

OSAWA, S., JUKES, T. H., WATANABE, K. & MUTO, A. (1992). Recent evidence for evolution of the genetic code. *Microbiol. Rev*. **56,** 229–264.

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. U.S.A.* **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. U.S.A.* **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.

TSUCHIHASHI, Z. & BROWN, P. O. (1992). Sequence requirements for efficient translational frameshifting in the *Escherichia coli dna*X gene and the role of an unstable interaction between $tRNA^{Lys}$ and an AAG lysine codon. *Genes Dev*. **6,** 511–519.

VÖGELE, K., SCHWARTZ, E., WELZ, C., SCHILTZ, E. & RAK, B. (1991). High-level ribosomal frameshifting directs the synthesis of IS150 gene products. *Nucl. 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. Nucl. 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.

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.