Bulletin of Mathematical Biology (1998) **60**, 163–194

An Evolutionary Analytical Model of a Complementary Circular Code Simulating the Protein Coding Genes, the 5′ and 3′ Regions

DIDIER G. ARQUÈS
Equipe de Biologie Théorique,
Université de Marne la Vallée,
Institut Gaspard Monge,
2 rue de la Butte Verte,
93160 Noisy Le Grand, France
Fax: 01 49 32 91 38
E-mail: arques@univ-mlv.fr

JEAN-PAUL FALLOT, CHRISTIAN J. MICHEL*
Equipe de Biologie Théorique,
Institut Polytechnique de Sévenans,
Rue du Château,
Sévenans,
90010 Belfort, France
Fax: 03 84 58 30 30
E-mail: christian.michel@utbm.fr

The self-complementary subset $\mathcal{T}_0 = \mathcal{X}_0 \cup \{\text{AAA, TTT}\}$ with $\mathcal{X}_0 = \{\text{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC}\}$ of 22 trinucleotides has a preferential occurrence in the frame 0 (reading frame established by the ATG start trinucleotide) of protein (coding) genes of both prokaryotes and eukaryotes. The subsets $\mathcal{T}_1 = \mathcal{X}_1 \cup \{\text{CCC}\}$ and $\mathcal{T}_2 = \mathcal{X}_2 \cup \{\text{GGG}\}$ of 21 trinucleotides have a preferential occurrence in the shifted frames 1 and 2 respectively (frame 0 shifted by one and two nucleotides respectively in the 5′–3′ direction). $\mathcal{T}_1$ and $\mathcal{T}_2$ are complementary to each other. The subset $\mathcal{T}_0$ contains the subset $\mathcal{X}_0$ which has the rarity property (6×10^{-8}) to be a complementary maximal circular code with two permutated maximal circular codes $\mathcal{X}_1$ and $\mathcal{X}_2$ in the frames 1 and 2 respectively. $\mathcal{X}_0$ is called a C^3 code.

A quantitative study of these three subsets $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the three frames 0, 1, 2 of protein genes, and the 5′ and 3′ regions of eukaryotes, shows that their occurrence frequencies are constant functions of the trinucleotide positions in the sequences. The frequencies of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the frame 0 of protein genes are 49, 28.5 and 22.5% respectively. In contrast, the frequencies of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the 5′ and 3′ regions of eukaryotes, are independent of the frame. Indeed, the frequency of $\mathcal{T}_0$ in the three frames of 5′ (respectively 3′) regions is equal to 35.5% (respectively 38%) and is greater than the frequencies $\mathcal{T}_1$ and $\mathcal{T}_2$, both equal to 32.25% (respectively 31%) in the three frames.

*Author to whom correspondence should be addressed.

0092-8240/98/010163 + 32 $25.00/0 bu970033

Several frequency asymmetries unexpectedly observed (e.g. the frequency difference between $\mathcal{T}_1$ and $\mathcal{T}_2$ in the frame 0), are related to a new property of the subset $\mathcal{T}_0$ involving substitutions. An evolutionary analytical model at three parameters (p, q, t) based on an independent mixing of the 22 codons (trinucleotides in frame 0) of $\mathcal{T}_0$ with equiprobability (1/22) followed by $t \approx 4$ substitutions per codon according to the proportions $p \approx 0.1$, $q \approx 0.1$ and $r = 1 - p - q \approx 0.8$ in the three codon sites respectively, retrieves the frequencies of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ observed in the three frames of protein genes and explains these asymmetries. Furthermore, the same model $(0.1, 0.1, t)$ after $t \approx 22$ substitutions per codon, retrieves the statistical properties observed in the three frames of the $5'$ and $3'$ regions. The complex behaviour of these analytical curves is totally unexpected and *a priori* difficult to imagine.

1. INTRODUCTION

1.1. ***Historical background.*** The concept of a code without commas, introduced by Crick *et al.* (1957), is a code readable in only one frame and without a start signal. Such a theoretical code 'without commas' is a set $\mathcal{X}$ of codons so that their concatenation (series of codons) leads to genes which have the interesting property of automatically retrieving the concatenation of codons of $\mathcal{X}$, without the use of a start codon, in the case that 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 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) proposed that only 20 among 64 codons would code for the 20 amino acids. However, the determination of a set of 20 codons forming a code $\mathcal{X}$ 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 retrieval of the frame: ...AAA, AAA, AAA, ..., ...A, AAA, AAA, AA... and ...AA, AAA, AAA, A.... Similarly, two codons related to circular permutations, e.g. AAC and ACA (or CAA), cannot belong to such a code at the same time. Indeed, the concatenation of AAC with itself for example, leads to the concatenation of ACA (or CAA) with itself in another frame, making the frame determination impossible. Therefore, by excluding AAA, CCC, GGG and TTT and by gathering the 60 remaining codons in 20 classes of three codons so that, in each class, the three codons are deduced from each other by circular permutations (e.g. AAC, ACA and CAA), a code without commas has

only one codon per class and therefore contains at most 20 codons. This codon number is identical to the amino acid number. This remark has naturally led to the proposition of 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 a G + C composition observed in the actual protein genes (Jukes and Bhushan, 1986).

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

Recently, a code without commas (called circular code in computer terms) on the alphabet {A, C, G, T} has been identified in the protein genes of both prokaryotes and eukaryotes (Arquès and Michel, 1996). This circular code is associated with 20 codons which occur preferentially in reading frame, in comparison with the other codons. The first aim of this paper is to quantify the preferential occurrence of these 20 codons (Section 2). In the actual protein genes, the 64 codons are used for protein synthesis: one start codon for locating the reading frame, 61 codons for coding the 20 amino acids and three stop codons (according to the universal genetic code). However, the circular code observed may be the trace, after evolution, of a primitive structure of protein genes. The second aim of this paper is the development of an evolutionary analytical model in order to propose an explanation for the presence of a circular code in the actual protein genes (Section 3). This model will be based on two processes: a construction process of primitive genes with the circular code and an evolutionary process with different rates of random substitutions in the three codon sites.

In order to understand the circular code identified in protein genes of prokaryotes and eukaryotes on the alphabet {A, C, G, T} (Section 1.3), the concept of circular code is first presented on the alphabet {R, Y} in Section 1.2. Sections 1.2 and 1.3 are the necessary reminders of the results obtained and detailed in Arquès and Michel (1996). Sections 1.4 and 1.5 present the new results. Section 1.4 introduces the quantitative study of the circular code which will identify frequency asymmetries in contradiction with the complementarity property of the circular code. Section 1.5 presents an evolutionary model explaining these asymmetries and leading to the identification of a new property, namely an evolutionary property, of the circular code observed in protein genes.

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

Recall of the DNA complementarity rule (Watson and Crick, 1953). The DNA double helix consists of two nucleotide sequences s_1 and s_2 connected with the nucleotide pairing (hydrogen bonds) according to the complementarity rule $\mathcal{C}$: the nucleotide A (respectively C, G. T) in s_1 pairs with the complementary nucleotide $\mathcal{C}(\mathrm{A}) = \mathrm{T}$ (resp. $\mathcal{C}(\mathrm{C}) = \mathrm{G}, \mathcal{C}(\mathrm{G}) = \mathrm{C}, \mathcal{C}(\mathrm{T}) = \mathrm{A}$) in s_2. The extension of this rule to the alphabet $\mathcal{B}_2$ leads to $\mathcal{C}(\mathrm{R}) = \mathrm{Y}$ and $\mathcal{C}(\mathrm{Y}) = \mathrm{R}$. The two nucleotide sequences s_1 and s_2 run in opposite directions (called antiparallel) in the DNA double helix: the trinucleotide $w = l_1 l_2 l_3, l_1, l_2, l_3 \in \mathcal{B}$, in s_1 pairs with the complementary trinucleotide $\mathcal{C}(w) = \mathcal{C}(l_3)\mathcal{C}(l_2)\mathcal{C}(l_1)$ in s_2 e.g. $\mathcal{C}(\mathrm{AAC}) = \mathrm{GTT}, \mathcal{C}(\mathrm{RRY}) = \mathrm{RRY}$.

Recall of the trinucleotide circular permutation. The circular permutation $\mathcal{P}$ of the trinucleotide $w = l_1 l_2 l_3$, is the permutated trinucleotide $\mathcal{P}(w) = l_2 l_3 l_1$, e.g. $\mathcal{P}(\mathrm{AAC}) = \mathrm{ACA}, \mathcal{P}(\mathrm{RRY}) = \mathrm{RYR}$. If $\mathcal{X}$ is a set of trinucleotides then $\mathcal{P}(\mathcal{X})$ is the set of permutated trinucleotides of $\mathcal{X}$.

Definition of a circular code. A subset $\mathcal{X}$ of $\mathcal{B}^+$ is a circular code if for all $n, m \geq 1$ and $x_1, x_2, \ldots, x_n \in \mathcal{X}, y_1, y_2, \ldots, y_m \in \mathcal{X}$ and $p \in \mathcal{B}^*, s \in \mathcal{B}^+$, 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 $x_i = y_i$, $1 \leq i \leq n$ (Béal, 1993; Berstel and Perrin, 1985). In other terms, every word on $\mathcal{B}$ 'written on a circle' has at most one factorization (decomposition) over $\mathcal{X}$. In the following, $\mathcal{X}$ will be a set of words of length 3 as a protein gene is a concatenation of trinucleotides.

The main consequence of the circular code property is the frame determination property (admitted). If a word is constructed by concatenating words of $\mathcal{X}$ and if the frame of construction is lost, then the code property assures that the frame of construction can be retrieved in a unique way.

On the alphabet $\mathcal{B}_2 = \{\mathrm{R}, \mathrm{Y}\}$, there are nine potential maximal (sets of two trinucleotides) circular codes; (Arquès and Michel, 1996). Two of these nine sets, $\mathcal{X}_a = \{\mathrm{RRY}, \mathrm{RYY}\} = \mathrm{RNY}$ and $\{\mathrm{YRR}, \mathrm{YYR}\} = \mathrm{YNR}$, are complementary maximal circular codes with two permutated maximal circular codes (called C^3 codes; Arquès and Michel, 1996). This concept of circular code is presented with the set $\mathcal{X}_a$ which is associated with the biological model of RNY codons (Eigen and Schuster, 1978). The RNY codon model leads to a protein gene model formed by a series RNYRNY... of nucleotides so that there is one type of trinucleotide RNY ($\mathcal{X}_a$) in frame 0 (reading frame), one type of trinucleotide NYR ($\mathcal{X}_b$) in frame 1 and one type of trinucleotide YRN ($\mathcal{X}_c$) in frame 2 (frames 1 and 2 being the frame 0 shifted by one and two nucleotides respectively in the 5′–3′ direction). NYR (resp. YRN) is obtained by one (resp. two) circular permutation of RNY.

.	.	.	,	R	N	Y	,	R	N	Y	,	R	N	Y	,	. . . :RNY $\in \chi_a$
.	.	.		R	,	N	Y	R	,	N	Y	R	,	N	Y	. , . . :NYR $\notin \chi_a$
.	.	.		R	N	,	Y	R	N	,	Y	R	N	,	Y	. . , . :YRN $\notin \chi_a$

Figure 1. The set $\mathcal{X}_a$ = RNY is a circular code as there is a unique decomposition over $\mathcal{X}_a$.

The set $\mathcal{X}_a = \{\text{RRY}, \text{RYY}\} = \text{RNY}$ is a maximal circular code. Indeed, the concatenation of two trinucleotides of $\mathcal{X}_a$, . . .RRYRRY. . ., . . .RRYRYY. . ., . . .RYYRRY. . . and . . .RYYRYY. . ., leads to only one factorization over $\mathcal{X}_a$ as the eventual decomposition in frame 1 has always an R in the third position but no trinucleotide of $\mathcal{X}_a$ ends with R and as the eventual decomposition in frame 2 has always a Y in the first position but no trinucleotide of $\mathcal{X}_a$ begins with Y (Fig. 1). Furthermore, $\mathcal{X}_a$ is self complementary, i.e. $\mathcal{C}(\mathcal{X}_a) = \mathcal{X}_a$ as RRY and RYY are complementary. Any subset of $\mathcal{X}_a$ is also a circular code but not maximal. Therefore, the RRY model (Crick *et al.*, 1976) is a non-maximal circular code. The two sets $\mathcal{X}_b = \mathcal{P}(\mathcal{X}_a) = \{\text{RYR}, \text{YYR}\} = \text{NYR}$ and $\mathcal{X}_c = \mathcal{P}^2(\mathcal{X}_a) = \{\text{YRR}, \text{YRY}\} = \text{YRN}$ obtained by circular permutations of $\mathcal{X}_a$ are also maximal circular codes (identical proof). Furthermore, $\mathcal{X}_b$ and $\mathcal{X}_c$ are complementary to each other, i.e. $\mathcal{C}(\mathcal{X}_b) = \mathcal{X}_c$ and $\mathcal{C}(\mathcal{X}_c) = \mathcal{X}_b$, as RYR (resp. YYR) and YRY (resp. YRR) are complementary. In summary, the set $\mathcal{X}_a$ = RNY is a complementary maximal circular code with two permutated maximal circular codes $\mathcal{X}_b = \mathcal{P}(\mathcal{X}_a)$ = NYR and $\mathcal{X}_c = \mathcal{P}^2(\mathcal{X}_a)$ =YRN (C^3 code).

1.3. *A C^3 code identified in the protein coding genes on the alphabet {A, C, G, T}.*

In contrast to the alphabet $\mathcal{B}_2$ ={R, Y} where the circular codes can be completely studied by hand, the identification of a circular code on the alphabet $\mathcal{B}_4$ ={A, C, G, T} is obviously more complex and difficult as there are $\approx$ 3.5 milliard potential maximal (sets of 20 trinucleotides) circular codes (Arquès and Michel, 1996, Table 2d). Unexpectedly, a simple method computing the occurrence frequencies of the 64 trinucleotides AAA, . . ., TTT in the three frames 0, 1, 2 of protein (coding) genes and assigning each trinucleotide to the frame associated with its highest frequency, has recently identified three subsets of trinucleotides per frame: $\mathcal{T}_0 = \mathcal{X}_0 \cup \{\text{AAA, TTT}\}$ with $\mathcal{X}_0 = \{\text{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC}\}$ in frame 0, $\mathcal{T}_1 = \mathcal{X}_1 \cup \{\text{CCC}\}$ and $\mathcal{T}_2 = \mathcal{X}_2 \cup \{\text{GGG}\}$ in the shifted frames 1 and 2 respectively, with $\mathcal{X}_1$ and $\mathcal{X}_2$ defined in Table 1. The subsets $\mathcal{T}_0$, $\mathcal{T}_1$ and $\mathcal{T}_2$ have 22, 21 and 21 trinucleotides respectively. Furthermore, the same three subsets $\mathcal{T}_0$, $\mathcal{T}_1$, $\mathcal{T}_2$ are retrieved with very few exceptions for the two large protein gene populations of prokaryotes (13,686 sequences, 4,708,758 trinucleotides) and eukaryotes (26,757 sequences, 11,397,678 trinucleotides).

Table 1. Identification of three subsets of trinucleotides in the protein coding genes of both prokaryotes and eukaryotes (Arquès and Michel, 1996): $\mathcal{T}_0 = \mathcal{X}_0 \cup \{AAA, TTT\}$ in frame 0, $\mathcal{T}_1 = \mathcal{X}_1 \cup \{CCC\}$ in frame 1 and $\mathcal{T}_2 = \mathcal{X}_2 \cup \{GGG\}$ in frame 2.

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

The three subsets $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ of trinucleotides have five important properties (detailed in Arquès and Michel, 1996).

(i) The property of a maximal (20 trinucleotides) circular code for $\mathcal{X}_0$ allowing automatic retrieval of frame 0 in any region of a protein gene model formed by a series of trinucleotides of $\mathcal{X}_0$. A biological consequence of this property is the uselessness of motifs for locating the reading frame (frame 0). In the actual protein genes, the most important motif initiating the reading frame is the start codon ATG. Furthermore, $\mathcal{X}_1$ and $\mathcal{X}_2$ are also maximal circular codes.

(ii) The DNA complementarity property $\mathcal{C} : \mathcal{C}(\mathcal{T}_0) = \mathcal{T}_0$ ($\mathcal{T}_0$ is self-complementary: 11 trinucleotides of $\mathcal{T}_0$ are complementary to the 11 other trinucleotides of $\mathcal{T}_0$), $\mathcal{C}(\mathcal{T}_1) = \mathcal{T}_2$ and $\mathcal{C}(\mathcal{T}_2) = \mathcal{T}_1$ ($\mathcal{T}_1$ and $\mathcal{T}_2$ are complementary to each other: the 21 trinucleotides of $\mathcal{T}_1$ are complementary to the 21 trinucleotides of $\mathcal{T}_2$). This property allows the two paired reading frames of a DNA double helix simultaneously to code for amino acids, in agreement with biological results (Zull and Smith, 1990; Konecny *et al.*, 1993; Béland and Allen, 1994; Konecny *et al.*, 1995).

(iii) The circular permutation property $\mathcal{P} : \mathcal{P}(\mathcal{X}_0) = \mathcal{X}_1$ and $\mathcal{P}(\mathcal{X}_1) = \mathcal{X}_2$ ($\mathcal{X}_0$ generates $\mathcal{X}_1$ by one circular permutation and $\mathcal{X}_2$ by another circular permutation: one and two circular permutations with each trinucleotide of $\mathcal{X}_0$ lead to the trinucleotides of $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively) implying that the two subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ can be deduced from $\mathcal{X}_0$.

(iv) The rarity property: there are 216 codes with the three properties (i)–(iii) among 3^{20} potential maximal circular codes, i.e. the occurrence probability of $\mathcal{X}_0$ is equal to $216/3^{20} = 6 \times 10^{-8}$ (Arquès and Michel, 1996, Table 2d). This probability is very low and therefore, non-random in protein genes. In addition, this code $\mathcal{X}_0$ is observed in two independent and large protein gene populations (prokaryotes: 13,686 sequences and eukaryotes: 26,757 sequences).

(v) Three concatenation properties (Arquès and Michel, 1996, Section 3.7) implying that the code $\mathcal{X}_0$ has flexibility properties.

In summary, the self-complementary subset $\mathcal{T}_0$ of 22 trinucleotides identified in protein genes of prokaryotes and eukaryotes contains the subset $\mathcal{X}_0$ of 20 trinucleotides which is a complementary maximal circular code with two permutated

maximal circular codes (C^3 code) and with concatenation properties allowing retrieval 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 and Schuster, 1978) and of the flexible code (Dounce, 1952). Several consequences of the subset $\mathcal{T}_0$ have been studied with respect to the three two-letter genetic alphabets (purine/pyrimidine, amino/ceto, strong/weak interaction), the genetic code, the amino acid frequencies in proteins and the complementary paired DNA sequence (Arquès and Michel, 1996). A new property, precisely an evolutionary property, of the C^3 code $\mathcal{X}_0$ is identified in this paper.

1.4. ***Identification of unexpected frequency asymmetries.*** As the trinucleotides in $\mathcal{T}_0$ (resp. $\mathcal{T}_1, \mathcal{T}_2$) have a preferential occurrence in frame 0 (resp. 1, 2) (Table 1), the global mean frequency of $\mathcal{T}_0$ (resp. $\mathcal{T}_1, \mathcal{T}_2$) in frame 0 (resp. 1, 2) will be expected to be greater than the global mean frequencies of $\mathcal{T}_1$ and $\mathcal{T}_2$ (resp. $\mathcal{T}_0$ and $\mathcal{T}_2$, and $\mathcal{T}_0$ and $\mathcal{T}_1$) in frame 0 (resp. 1, 2). Furthermore, the complementarity property of the C^3 code $\mathcal{X}_0$ would imply several symmetries with the frequencies of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the three frames, in particular the same frequency of $\mathcal{T}_1$ and $\mathcal{T}_2$ in frame 0 (detailed in Section 3). In Section 2, in order to verify these quantitative consequences of the subset $\mathcal{T}_0$, the occurrence frequencies of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ are computed for each codon position (after the start trinucleotide ATG and before the stop trinucleotide TAA, TAG or TGA) in the three frames of protein genes of higher eukaryotes (large gene populations of primates, rodents, (other) mammals and (other) vertebrates). The strong statistical properties associated with the subset $\mathcal{T}_0$, e.g. the preferential occurrence of $\mathcal{T}_0$ (resp. $\mathcal{T}_1, \mathcal{T}_2$) in the frame 0 (resp. 1, 2), are indeed observed in eukaryotic protein genes. However, several frequency asymmetries are identified in protein genes which are in contradiction with the complementarity property of the C^3 code $\mathcal{X}_0$.

In contrast to the protein genes, the 5′ and 3′ regions have no protein coding function. The previous method applied to the 5′ and 3′ regions of higher eukaryotes, i.e. the computation of the occurrence frequencies of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ for each codon position in the three extended frames before ATG (5′ regions) and after TAA, TAG or TGA (3′ regions), shows indeed other statistical properties, in particular a frequency of $\mathcal{T}_0$ greater than the frequencies of $\mathcal{T}_1$ and $\mathcal{T}_2$ whatever the frame.

1.5. ***An evolutionary C^3 code.*** An evolutionary process by substitutions of nucleotides will allow an explanation of the frequency asymmetries observed in eukaryotic protein genes. The evolutionary model which will be tested in Section 3, is based on two processes.

(i) A construction process generating simulated primitive genes according to an independent mixing of the 22 trinucleotides of the subset $\mathcal{T}_0$ with equiprobability (1/22). These primitive genes have 22 among 64 trinu-

cleotides and frequency symmetries (due to the complementarity property of $\mathcal{T}_0$) which are not observed in the real actual protein genes.

(ii) An evolutionary process based on substitutions in the three trinucleotide sites transforming the simulated primitive genes into simulated actual genes. Substitutions with different rates in the sites of trinucleotides of $\mathcal{T}_0$ allow generation of the trinucleotides of $\mathcal{T}_1$ and $\mathcal{T}_2$ according to a non-balanced way and retrieval of the frequency asymmetries of the real actual protein genes. Therefore, these asymmetries are related to a new property, namely an evolutionary property, of the C^3 code $\mathcal{X}_0$. The measure of these asymmetries in the model quantifies this evolutionary property, i.e. the number of substitutions. According to the degeneracy of the genetic code, the highest substitution rate is expected to occur in the third codon site, which is observed in the model. Indeed, after ≈ 4 substitutions per codon in the three codon sites in the proportions $\approx 0.1, \approx 0.1$ and ≈ 0.8 respectively, the simulated actual genes are correlated with the real actual genes.

The 5′ and 3′ regions surround the protein genes and can be subjected to a high rate of substitutions as they have no protein coding function (no reading frame). These 5′ and 3′ regions appear as a limit case (high substitution rate) of the evolutionary process of protein genes. The same model, i.e. with the same substitution rates $\approx 0.1, \approx 0.1$ and ≈ 0.8 in the three codon sites, will, after ≈ 22 substitutions per codon, indeed retrieve the statistical properties observed in the three frames of the 5′ and 3′ regions.

2. A Quantitative Study of the Subset $\mathcal{T}_0$ in the Protein Coding Genes, 5′ and 3′ Regions of Eukaryotes

2.1. ***Method.*** Let w be a trinucleotide in $\mathcal{T}$ ={AAA, ..., TTT} (64 trinucleotides). Let $f \in \{0, 1, 2\}$ be a frame determined by a series of trinucleotides in a gene s of a population F. The frame $f = 0$ is the reading frame established by the start trinucleotide ATG up to a stop trinucleotide TAA, TAG or TGA and the frames $f = 1$ and $f = 2$ are the frame 0 shifted by one and two nucleotides respectively in the 5′–3′ direction. The concept of frame is extended to the 5′ and 3′ regions by continuing a series of trinucleotides before ATG (5′ region) and after TAA, TAG or TGA (3′ region). By choosing the stop trinucleotide TAA as an example, $f = 0$ is the following frame ..., NNN, ATG, NNN, ..., NNN, TAA, NNN, ... and $f = 1$, ..., NNA, TGN, ..., NNT, AAN, ... and $f = 2$, ..., NAT, GNN, ..., NTA, ANN, ... (N being any nucleotide). Therefore, the population F containing the genes s read in the frame f is noted F_f. By representing the 5′–3′ DNA direction by an axis whose origin is either ATG or a stop trinucleotide, the algebraic position d in a given frame f is defined as being the number of trinucleotides before ATG (5′ region, $d < 0$), after ATG (5′ part of a protein gene, $d > 0$), before a stop trinucleotide (3′ part of a protein gene, $d < 0$) and

after a stop trinucleotide (3′ region, $d > 0$). A positive (resp. negative) position is then related to the 5′–3′ (resp. 3′–5′) direction. For example, $d = 10$ in $f = 0$ (resp. $f = 1$, $f = 2$) is the tenth codon after ATG (resp. TGN, GNN) or after the chosen stop trinucleotide TAA (resp. NNT, NTA) and $d = -10$ in $f = 0$ (resp. $f = 1$, $f = 2$) is the tenth codon before ATG (resp. TGN, GNN) or before the chosen stop trinucleotide TAA (resp. NNT, NTA). For a given frame, a trinucleotide w at the algebraic position d is noted w_d. Let $\mathcal{T}_g$ be the subset of trinucleotides having a preferential occurrence in the frame $g \in \{0, 1, 2\}$ (Table 1). In a given frame f of a gene s, the function

$$\delta_d(\mathcal{T}_g) = \begin{cases} 1 & \text{if } w_d \in \mathcal{T}_g \\ 0 & \text{if } w_d \notin \mathcal{T}_g \end{cases}$$

determines if the trinucleotide w at the position d belongs or not to $\mathcal{T}_g$ with $g = 0, 1, 2$. Then, the occurrence probability $P_d(\mathcal{T}_g, F_f)$ of a subset $\mathcal{T}_g$ at the trinucleotide position d in a gene population F_f, is

$$P_d(\mathcal{T}_g, F_f) = \sum_{s \in F_f} \delta_d(\mathcal{T}_g) \Big/ n(w_d)$$

where $n(w_d)$ is the total number of trinucleotides w at the position d in the gene population F_f.

This probability function is represented as a curve as follows: the abscissa shows the position d in trinucleotides, by varying d in a given range, e.g. [2, 200] (5′ parts of protein genes or 3′ regions) and $[-200, -2]$ (3′ parts of protein genes or 5′ regions), and the ordinate gives the occurrence probability of $P_d(\mathcal{T}_0, F_f)$, $P_d(\mathcal{T}_1, F_f)$ and $P_d(\mathcal{T}_2, F_f)$ in a protein gene population F_f.

REMARKS:

(i) For readability reasons, the ATG, the stop and the first ($d = 1$ and $d = -1$) conserved trinucleotides are not represented in the curves.

(ii) $P_d(\mathcal{T}_f, F_f) > P_d(\mathcal{T}_g, F_f)$, $f, g \in \{0, 1, 2\}$ and $g \neq f$, for any position d in the three frames $f \in \{0, 1, 2\}$ of a protein gene population F_f as $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ have a preferential occurrence in the frames 0, 1 and 2 respectively (Table 1).

(iii) $P_d(\mathcal{T}_0, H_f) = 22/64 = 0.34375$ and $P_d(\mathcal{T}_1, H_f) = P_d(\mathcal{T}_2, H_f) = 21/64 = 0.328125$ for any position d in the three frames $f \in \{0, 1, 2\}$ of a random gene population H_f generated with an independent mixing of the four nucleotides A, C, G and T with equiprobability (1/4).

The large gene populations F of higher eukaryotes analyzed here are:

		Protein (coding) genes (P)	
	5′ regions (R5)	5′ parts (P5) and 3′ parts (P3)	3′ regions (R3)
Humans + Rodents + Mammals + Vertebrates (HRMV)	7745 sequences $F = $ R5_HRMV	20,345 sequences $F = $ P5_HRMV and $F = $ P3_HRMV	15,756 sequences $F = $ R3_HRMV
Humans (HUM)	2933 sequences $F = $ R5_HUM	7507 sequences $F = $ P5_HUM and $F = $ P3_HUM	5915 sequences $F = $ R3_HUM
Rodents (ROD)	3134 sequences $F = $ R5_ROD	7594 sequences $F = $ P5_ROD and $F = $ P3_ROD	5919 sequences $F = $ R3_ROD
Mammals (MAM)	721 sequences $F = $ R5_MAM	2586 sequences $F = $ P5_MAM and $F = $ P3_MAM	1869 sequences $F = $ R3_MAM
Vertebrates (VRT)	957 sequences $F = $ R5_VRT	2658 sequences $F = $ P5_VRT and $F = $ P3_VRT	2053 sequences $F = $ R3_VRT

These large populations, obtained from EMBL Nucleotide Sequence Data Library Release 47 (June 1996) in the same way as described in previous studies (see, e.g., Arquès and Michel (1987, 1990) for a description of data acquisitions), allow stable frequencies (consequence of the law of large numbers, Arquès and Michel (1990), Section 2.3.3).

2.2. *Results.*

2.2.1. *Protein coding genes of eukaryotes.* Figures 2a, b show that the probability curve $\mathcal{T}_0$ is, as expected, greater than the two curves $\mathcal{T}_1$ and $\mathcal{T}_2$, for any trinucleotide position d in the frame 0 of the 5′ parts (Fig. 2a: $P_d(\mathcal{T}_g, \text{P5\_HRMV}_0)$) and the 3′ parts (Fig. 2b: $P_d(\mathcal{T}_g, \text{P3\_HRMV}_0)$) of protein genes of humans, rodents, mammals and vertebrates. The curve $\mathcal{T}_0$ is globally horizontal with an average frequency around 49% in P_HRMV_0 (P5_HRMV_0 and P3_HRMV_0) (Table 2a). The two curves $\mathcal{T}_1$ and $\mathcal{T}_2$ are also globally horizontal but unexpectedly distinct (Fig. 2a, b). Indeed, the average frequency of $\mathcal{T}_1$ around 28.5% is greater than the frequency of $\mathcal{T}_2$ around 22.5% in P_HRMV_0 (Table 2a). The probability difference $P_d(\mathcal{T}_1, \text{P\_HRMV}_0) - P_d(\mathcal{T}_2, \text{P\_HRMV}_0) \approx 0.06$ in frame 0 cannot be explained by the difference $1/64 \approx 0.016$ consequent on the fact that $\mathcal{T}_1$ has one stop trinucleotide less than $\mathcal{T}_2$ (TAG $\in \mathcal{T}_1$ and TAA, TGA $\in \mathcal{T}_2$, Table 1) and is a first contradiction with the expected equality resulting from the complementarity property of the C^3 code $\mathcal{X}_0$ (self-complementarity of $\mathcal{T}_0$ and complementarity of $\mathcal{T}_1$ and $\mathcal{T}_2$). The probabilities in frame 0 can be represented by the following set $\mathcal{Q}(\text{P}_0)$ of inequalities: $P_d(\mathcal{T}_0, \text{P\_HRMV}_0) > P_d(\mathcal{T}_1, \text{P\_HMRV}_0) > P_d(\mathcal{T}_2, \text{P\_HMRV}_0)$.

The horizontally as well as the frequency of these three curves are retrieved by increasing the trinucleotide position d, e.g. [2, 500] and $[-500, -2]$, in the frame 0 of protein genes of humans, rodents, mammals and vertebrates (data not shown).

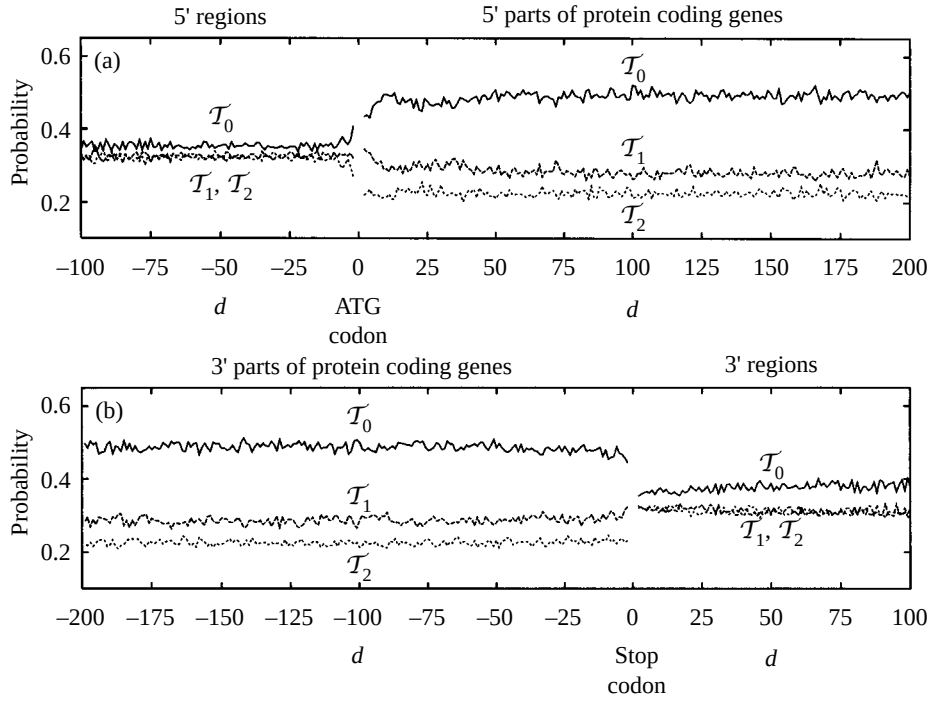

Figure 2. (a) Probability $P_d(\mathcal{T}_g, F_0)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ at the trinucleotide position d in frame 0 of the 5′ regions (F_0 = R5_HRMV$_0$) and the 5′ parts of protein coding genes (F_0 = P5_HRMV$_0$) of humans, rodents, mammals and vertebrates (HRMV$_0$). Three distinct horizontal curves $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in decreasing probabilities occur in the protein coding genes. The horizontal curve $\mathcal{T}_0$ is slightly greater than the two mixed curves $\mathcal{T}_1$ and $\mathcal{T}_2$ in the 5′ regions, similar to the two other frames 1 (Fig. 3a) and 2 (Fig. 4a). (b) Probability $P_d(\mathcal{T}_g, F_0)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ at the trinucleotide position d in frame 0 of the 3′ parts of protein coding genes (F_0 = P3_HRMV$_0$) and the 3′ regions (F_0 = R3_HRMV$_0$) of humans, rodents, mammals and vertebrates (HRMV$_0$). Three distinct horizontal curves $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in decreasing probabilities occur in the protein coding genes. The horizontal curve $\mathcal{T}_0$ is greater than the two mixed curves $\mathcal{T}_1$ and $\mathcal{T}_2$ in the 3′ regions, similar to the two other frames 1 (Fig. 3b) and 2 (Fig. 4b).

These results in frame 0 are also observed with each protein gene subpopulation (data not shown): humans ($P_d(\mathcal{T}_g$, P5_HUM$_0$) and $P_d(\mathcal{T}_g$, P3_HUM$_0$)), rodents ($P_d(\mathcal{T}_g$, P5_ROD$_0$) and $P_d(\mathcal{T}_g$, P3_ROD$_0$)), mammals ($P_d(\mathcal{T}_g$, P5_MAM$_0$) and $P_d(\mathcal{T}_g$, P3_MAM$_0$)) and vertebrates ($P_d(\mathcal{T}_g$, P5_VRT$_0$) and $P_d(\mathcal{T}_g$, P3_VRT$_0$)).

Figures 3a, b (resp. Fig. 4a, b) show that the probability curve $\mathcal{T}_1$ (resp. $\mathcal{T}_2$) is, as expected, greater than the two curves $\mathcal{T}_2$ and $\mathcal{T}_0$ (resp. $\mathcal{T}_0$ and $\mathcal{T}_1$), for any trinucleotide position d in frame 1 (resp. 2) of the 5′ and 3′ parts of protein genes of humans, rodents, mammals and vertebrates (Fig. 3a: $P_d(\mathcal{T}_g$, P5_HRMV$_1$) and Fig. 3b: $P_d(\mathcal{T}_g$, P3_HRMV$_1$)) (resp. Fig. 4a: $P_d(\mathcal{T}_g$, P5_HRMV$_2$) and Fig. 4b: $P_d(\mathcal{T}_g$, P3_HRMV$_2$)). Note that the curve $\mathcal{T}_1$ (resp. $\mathcal{T}_2$) has the highest probability in the shifted frame 1 (resp. 2) even if $\mathcal{T}_0$ has one trinucleotide more than $\mathcal{T}_1$ or $\mathcal{T}_2$. In frame 1, the three curves $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_0$ are globally horizontal with an

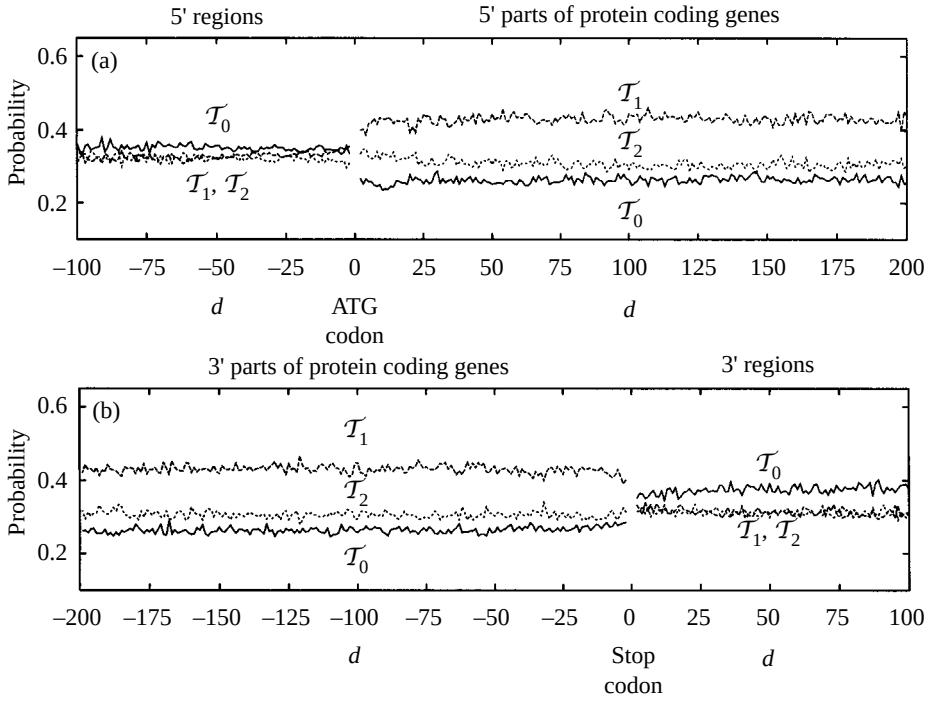

Figure 3. (a) Probability $P_d(\mathcal{T}_g, F_1)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ at the trinucleotide position d in frame 1 of the 5′ regions ($F_1 = \text{R5\_HRMV}_1$) and the 5′ parts of protein coding genes ($F_1 = \text{P5\_HRMV}_1$) of humans, rodents, mammals and vertebrates (HRMV_1). Three distinct horizontal curves $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_0$ in decreasing probabilities occur in the protein coding genes. In contrast to the protein coding genes, the horizontal curve $\mathcal{T}_0$ is slightly greater than the two mixed curves $\mathcal{T}_1$ and $\mathcal{T}_2$ in the 5′ regions, similar to the two other frames 0 (Fig. 2a) and 2 (Fig. 4a). (b) Probability $P_d(\mathcal{T}_g, F_1)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ at the trinucleotide position d in frame 1 of the 3′ parts of protein coding genes ($F_1 = \text{P3\_HRMV}_1$) and the 3′ regions ($F_1 = \text{R3\_HRMV}_1$) of humans, rodents, mammals and vertebrates (HRMV_1). Three distinct horizontal curves $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_0$ in decreasing probabilities occur in the protein coding genes. In contrast to the protein coding genes, the horizontal curve $\mathcal{T}_0$ is greater than the two mixed curves $\mathcal{T}_1$ and $\mathcal{T}_2$ in the 3′ regions, similar to the two other frames 0 (Fig. 2b) and 2 (Fig. 4b).

average frequency around 43, 30.5 and 26.5% respectively (Table 2a). The probabilities in frame 1 can be represented by the following set $\mathcal{Q}(\text{P}_1)$ of inequalities: $P_d(\mathcal{T}_1, \text{P\_HRMV}_1) > P_d(\mathcal{T}_2, \text{P\_HRMV}_1) > P_d(\mathcal{T}_0, \text{P\_HRMV}_1)$. In frame 2, the three curves $\mathcal{T}_2, \mathcal{T}_0, \mathcal{T}_1$ are globally horizontal with an average frequency around 45, 32 and 23% respectively (Table 2a). The probabilities in frame 2 can be represented by the following set $\mathcal{Q}(\text{P}_2)$ of inequalities: $P_d(\mathcal{T}_2, \text{P\_HRMV}_2) > P_d(\mathcal{T}_0, \text{P\_HRMV}_2) > P_d(\mathcal{T}_1, \text{P\_HMRV}_2)$. There is a second contradiction with the complementarity property of the C^3 code $\mathcal{X}_0$ which would have led to the following inequalities in frame 1: $P_d(\mathcal{T}_1, \text{P\_HRMV}_1) > P_d(\mathcal{T}_0, \text{P\_HRMV}_1) > P_d(\mathcal{T}_2, \text{P\_HRMV}_1)$. Indeed, a simulated population S generated, for example, according to an independent mixing of the 22 trinucleotides of $\mathcal{T}_0$ with equiprobabil-

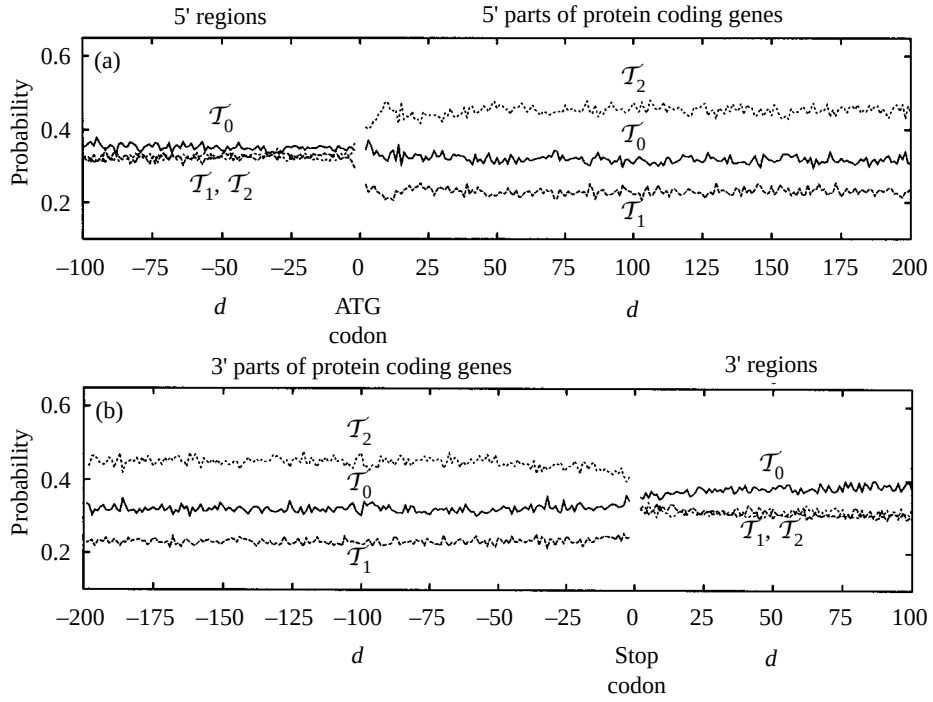

Figure 4. (a) Probability $P_d(\mathcal{T}_g, F_2)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ at the trinucleotide position d in frame 2 of the $5'$ regions ($F_2 = \text{R5\_HRMV}_2$) and the $5'$ parts of protein coding genes ($F_2 = \text{P5\_HRMV}_2$) of humans, rodents, mammals and vertebrates (HRMV_2). Three distinct horizontal curves $\mathcal{T}_2, \mathcal{T}_0, \mathcal{T}_1$ in decreasing probabilities occur in the protein coding genes. In contrast to the protein coding genes, the horizontal curve $\mathcal{T}_0$ is slightly greater than the two mixed curves $\mathcal{T}_1$ and $\mathcal{T}_2$ in the $5'$ regions, similar to the two other frames 0 (Fig. 2a) and 1 (Fig. 3a). (b) Probability $P_d(\mathcal{T}_g, F_2)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ at the trinucleotide position d in frame 2 of the $3'$ parts of protein coding genes ($F_2 = \text{P3\_HRMV}_2$) and the $3'$ regions ($F_2 = \text{R3\_HRMV}_2$) of humans, rodents, mammals and vertebrates (HRMV_2). Three distinct horizontal curves $\mathcal{T}_2, \mathcal{T}_0, \mathcal{T}_1$ in decreasing probabilities occur in the protein coding genes. In contrast to the protein coding genes, the horizontal curve $\mathcal{T}_0$ is greater than the two mixed curves $\mathcal{T}_1$ and $\mathcal{T}_2$ in the $3'$ regions, similar to the two other frames 0 (Fig. 2b) and 1 (Fig. 3b).

ity (1/22) leads to the following inequalities, in frame 0: $P(\mathcal{T}_0, S_0) > P(\mathcal{T}_1, S_0) = P(\mathcal{T}_2, S_0)$, in frame 1: $P(\mathcal{T}_1, S_1) > P(\mathcal{T}_0, S_1) > P(\mathcal{T}_2, S_1)$ and in frame 2: $P(\mathcal{T}_2, S_2) > P(\mathcal{T}_0, S_2) > P(\mathcal{T}_1, S_2)$.

The horizontality as well as the frequency of the three curves are retrieved by increasing the trinucleotide position d in frames 1 and 2 of protein genes of humans, rodents, mammals and vertebrates (data not shown). These results in frames 1 and 2 are also observed with each protein gene subpopulation (data not shown).

Table 2a. Mean frequencies $P(\mathcal{T}_g, F_f)$ (%) of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ in the three frames $f = 0, 1, 2$ of protein coding genes ($5'$ and $3'$ parts) of humans, rodents, mammals and vertebrates (with rounded averages).

	Protein coding genes of humans, rodents, mammals and vertebrates P_HRMV)		
	Frame 0		
	$5'$ parts P5_HRMV$_0$	$3'$ parts P3_HRMV$_0$	Average P_HRMV$_0$
$\mathcal{T}_0$	48.9	48.6	49
$\mathcal{T}_1$	28.6	28.8	28.5
$\mathcal{T}_2$	22.5	22.6	22.5
	Frame 1		
	$5'$ parts P5_HRMV$_1$	$3'$ parts P3_HRMV$_1$	Average P_HRMV$_1$
$\mathcal{T}_0$	26.2	26.4	26.5
$\mathcal{T}_1$	42.9	43.0	43
$\mathcal{T}_2$	30.9	30.6	30.5
	Frame 2		
	$5'$ parts P5_HRMV$_2$	$3'$ parts P3_HRMV$_2$	Average P_HRMV$_2$
$\mathcal{T}_0$	32.0	32.1	32
$\mathcal{T}_1$	22.9	23.2	23
$\mathcal{T}_2$	45.1	44.7	45

2.2.2. $5'$ *and* $3'$ *regions of eukaryotes.* In contrast to the eukaryotic protein genes, the probability curve $\mathcal{T}_0$ is greater than the probability curves $\mathcal{T}_1$ and $\mathcal{T}_2$, whatever the frame of the $5'$ and $3'$ regions of eukaryotes. Curves $\mathcal{T}_1$ and $\mathcal{T}_2$ are mixed (frame 0 of the $5'$ regions (Fig. 2a): $P_d(\mathcal{T}_g, \text{R5\_HRMV}_0)$, frame 0 of the $3'$ regions (Fig. 2b): $P_d(\mathcal{T}_g, \text{R3\_HRMV}_0)$, frame 1 of the $5'$ regions (Fig. 3a): $P_d(\mathcal{T}_g, \text{R5\_HRMV}_1)$, frame 1 of the $3'$ regions (Fig. 3b): $P_d(\mathcal{T}_g, \text{R3\_HRMV}_1)$, frame 2 of the $5'$ regions (Fig. 4a): $P_d(\mathcal{T}_g, \text{R5\_HRMV}_2)$, frame 2 of the $3'$ regions (Fig. 4b): $P_d(\mathcal{T}_g, \text{R3\_HRMV}_2)$). These probabilities which are common to the $5'$ and $3'$ regions (R) and independent of the frame f, can be represented by the following set $\mathcal{Q}(\text{R}_f)$ of inequalities: $P(\mathcal{T}_0, \text{R}_f) > P(\mathcal{T}_1, \text{R}_f) \approx P(\mathcal{T}_2, \text{R}_f)$. Otherwise, the probability curve $\mathcal{T}_0$ in the $3'$ regions (R3) is greater than the probability curve $\mathcal{T}_0$ in the $5'$ regions (R5) for the three frames (Figs. 2a, b, 3a, b, 4a, b). These results are obvious by smoothing the curves (frame 0 of the $5'$ and $3'$ regions (Fig. 5a): $P_d(\mathcal{T}_g, \text{R5\_HRMV}_0)$ and $P_d(\mathcal{T}_g, \text{R3\_HRMV}_0)$, frame 1 of the $5'$ and $3'$ regions (Fig. 5b): $P_d(\mathcal{T}_g, \text{R5\_HRMV}_1)$ and $P_d(\mathcal{T}_g, \text{R3\_HRMV}_1)$, frame 2 of the $5'$ and $3'$ regions (Fig. 5c): $P_d(\mathcal{T}_g, \text{R5\_HRMV}_2)$ and $P_d(\mathcal{T}_g, \text{R3\_HRMV}_2)$). The average frequency of $\mathcal{T}_0$ is 35.5% in the $5'$ regions (Table 2b) and 38% in the $3'$ regions (Table 2c). The average frequencies of $\mathcal{T}_1$ and $\mathcal{T}_2$ are about 32.25% in the $5'$ regions (Table 2b) and 31% in the $3'$ regions (Table 2c).

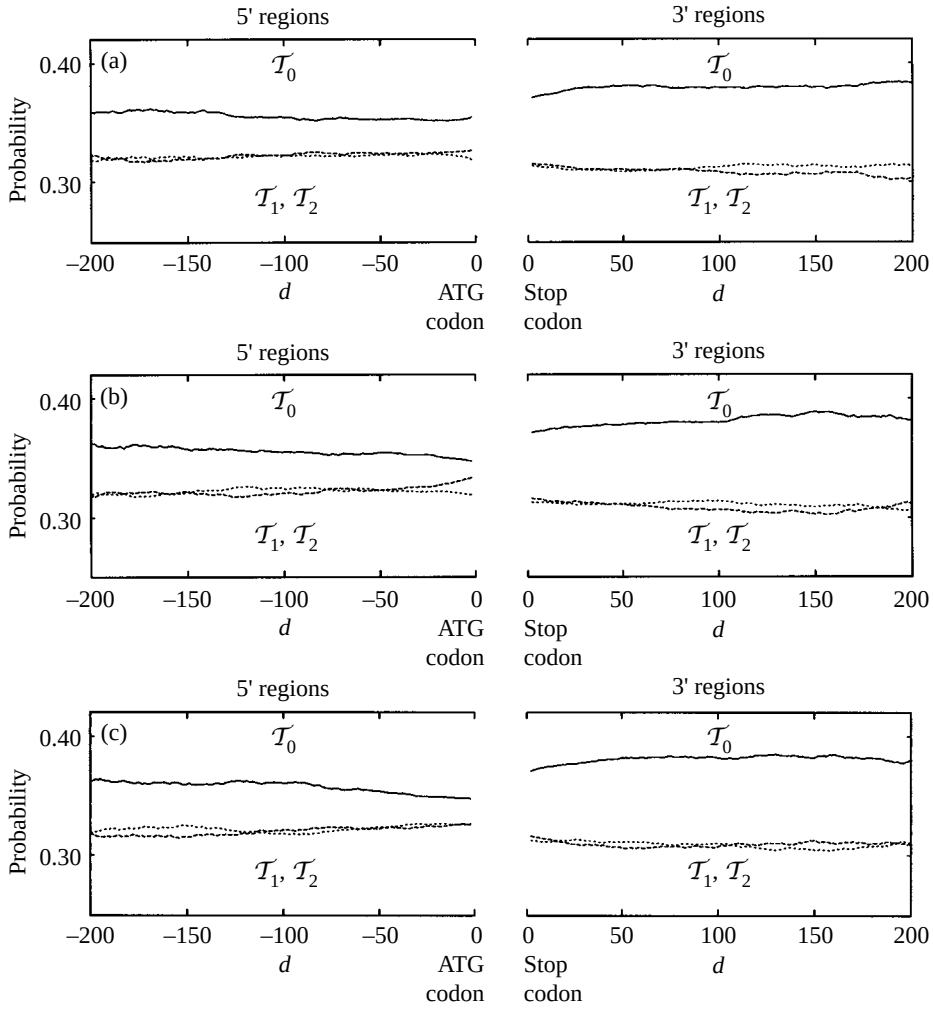

Figure 5. (a) Probability $P_d(\mathcal{T}_g, F_0)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ at the trinucleotide position d in frame 0 of the 5′ regions ($F_0 = \text{R5\_HRMV}_0$) and the 3′ regions ($F_0 = \text{R3\_HRMV}_0$) of humans, rodents, mammals and vertebrates (HRMV_0) (smooth curves). The probability curve $\mathcal{T}_0$ in the 3′ regions is greater than the probability curve $\mathcal{T}_0$ in the 5′ regions, similar to the other frames 1 (Fig. 5b) and 2 (Fig. 5c). (b) Probability $P_d(\mathcal{T}_g, F_1)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ at the trinucleotide position d in frame 1 of the 5′ regions ($F_1 = \text{R5\_HRMV}_1$) and the 3′ regions ($F_1 = \text{R3\_HRMV}_1$) of humans, rodents, mammals and vertebrates (HRMV_1) (smooth curves). The probability curve $\mathcal{T}_0$ in the 3′ regions is greater than the probability curve $\mathcal{T}_0$ in the 5′ regions, similar to the other frames 0 (Fig. 5a) and 2 (Fig. 5c). (c) Probability $P_d(\mathcal{T}_g, F_2)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ at the trinucleotide position d in frame 2 of the 5′ regions ($F2 = \text{R5\_HRMV}_2$) and the 3′ regions ($F_2 = \text{R3\_HRMV}_2$) of humans, rodents, mammals and vertebrates (HRMV_2) (smooth curves). The probability curve $\mathcal{T}_0$ in the 3′ regions is greater than the probability curve $\mathcal{T}_0$ in the 5′ regions, similar to the other frames 0 (Fig. 5a) and 1 (Fig. 5b).

Table 2b. Mean frequencies $P(\mathcal{T}_g, F_f)$ (%) of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ in the three frames $f = 0, 1, 2$ and in the rounded average frame of $5'$ regions of humans, rodents, mammals and vertebrates.

	$5'$ regions of humans, rodents, mammals and vertebrates (R5_HRMV)			
	Frame 0 $R5\_HRMV_0$	Frame 1 $R5\_HRMV_1$	Frame 2 $R5\_HRMV_2$	Average frame R5_HRMV
$\mathcal{T}_0$	35.5	35.1	35.2	35.5
$\mathcal{T}_1$	32.5	32.8	32.3	32.5
$\mathcal{T}_2$	32.0	32.1	32.5	32

Table 2c. Mean frequencies $P(\mathcal{T}_g, F_f)$ (%) of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ in the three frames $f = 0, 1, 2$ and in the rounded average frame of $3'$ regions of humans, rodents, mammals and vertebrates.

	$3'$ regions of humans, rodents, mammals and vertebrates (R3_HRMV)			
	Frame 0 $R3\_HRMV_0$	Frame 1 $R3\_HRMV_1$	Frame 2 $R3\_HRMV_2$	Average frame R3_HRMV
$\mathcal{T}_0$	37.7	37.7	37.8	38
$\mathcal{T}_1$	31.1	31.1	31.1	31
$\mathcal{T}_2$	31.2	31.2	31.1	31

These results in the three frames are also observed with each subpopulation (humans, rodents, mammals, vertebrates) of the $5'$ and $3'$ regions (data not shown).

3. An Evolutionary Analytical Model of the Subset $\mathcal{T}_0$

3.1. ***Presentation of the model.*** The three subsets $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ of trinucleotides in the analyzed gene populations present several statistical properties. The probability curves are constant functions of the trinucleotide position (horizontal curves) in the three frames of protein genes, $5'$ and $3'$ regions of eukaryotes. Therefore, these curves can be characterized by their probabilities $P(\mathcal{T}_g, F_f)$ (instead of $P_d(\mathcal{T}_g, F_f)$) (Table 2a–c). These probabilities are highly statistically significant as they are computed in a large population (HRMV) and retrieved in its subpopulations (HUM, ROD, MAM, VRT).

In the protein genes (P), the probabilities $P(\mathcal{T}_g, \mathrm{P}_f)$ of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the three frames $f = 0, 1, 2$ of P can be represented by three sets of inequalities, $\mathcal{Q}(\mathrm{P}_0)$ in frame 0: $P(\mathcal{T}_0, \mathrm{P}_0) > P(\mathcal{T}_1, \mathrm{P}_0) > P(\mathcal{T}_2, \mathrm{P}_0)$, $\mathcal{Q}(\mathrm{P}_1)$ in frame 1: $P(\mathcal{T}_1, \mathrm{P}_1) > P(\mathcal{T}_2, \mathrm{P}_1) > P(\mathcal{T}_0, \mathrm{P}_1)$ and $\mathcal{Q}(\mathrm{P}_2)$ in frame 2: $P(\mathcal{T}_2, \mathrm{P}_2) > P(\mathcal{T}_0, \mathrm{P}_2) > P(\mathcal{T}_1, \mathrm{P}_2)$ (Table 2a). As detailed in Section 2.2, these probability inequalities seem to be in contradiction with the complementarity property of the C^3 code $\mathcal{X}_0$.

In the $5'$ and $3'$ regions (R), the probabilities $P(\mathcal{T}_g, \mathrm{R}_f)$ of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the three frames $f = 0, 1, 2$ of R are independent of the frame f and can be associated with the set of inequalities $\mathcal{Q}(\mathrm{R}_f)$: $P(\mathcal{T}_0, \mathrm{R}_f) > P(\mathcal{T}_1, \mathrm{R}_f) \approx P(\mathcal{T}_2, \mathrm{R}_f)$ (Table 2b, c).

A new property of the subset $\mathcal{T}_0$ related to evolution by substitution is studied in this section. Precisely, the problem investigated here is whether a unique

evolutionary analytical model can explain the properties of the subsets $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ observed in the three frames of actual protein genes, and the 5′ and 3′ regions, in particular the asymmetries observed in protein genes and the statistical properties observed in the three frames of 5′ and 3′ regions. The main evolutionary process of DNA sequences is determined by substitutions of nucleotides. Editing of RNA (Benne *et al.*, 1986) by insertions and deletions of nucleotides, is an evolutionary process only observed in particular protein genes (mainly mitochondrial transcripts of the kinetoplastid protozoa and *Physarum polycephalum*) as it destroys the reading frame (reviews Benne, 1989; Feagin, 1990; Simpson, 1990; Stuart, 1991). However, RNA editing also occurs in the 5′ and 3′ regions, with a reduced extent compared with the protein genes (e.g. Feagin *et al.*, 1988; Shaw *et al.*, 1988). Therefore, as the actual protein genes have a preferential occurrence of the subset $\mathcal{T}_0$ (in frame 0; note also that $P(\mathcal{T}_0, \mathrm{P}_0) > P(\mathcal{T}_2, \mathrm{P}_2) > P(\mathcal{T}_1, \mathrm{P}_1)$ in Table 2a) and as the main process of gene evolution is determined by nucleotide substitutions, the model which will be tested, is based on two processes.

(i) A construction process based on an independent mixing of the 22 trinucleotides of $\mathcal{T}_0$ with equiprobability (1/22).
(ii) An evolutionary process based on substitutions in the three trinucleotide sites.

Such models based on two successive processes, construction and evolution (substitutions, insertions and deletions of nucleotides), have already been developed on the purine/pyrimidine alphabet (Arquès and Michel, 1990, 1992, 1993, 1994).

3.2. ***Method.*** In order to determine the exact probabilities of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ after substitutions, the analytical solutions giving the probabilities of the eight trinucleotides on the alphabet {R, Y} after a unique substitution rate per codon (Arquès and Michel, 1994) are generalized both to the 64 trinucleotides on the alphabet {A, C, G, T} and to the three substitution rates p, q and $r = 1 - p - q$ of the three codon sites respectively.

By convention, in the following, the indexes i or $j \in [1, 64]$ represent the trinucleotides AAA, ..., TTT in the alphabetical order. The occurrence probability $P_i(t + dt)$ of a trinucleotide i at a time $t + dt$ is equal to the occurrence probability $P_i(t)$ of this trinucleotide i at the time t minus the substitution probability of this trinucleotide i during $[t, t+dt]$ and plus the substitution probabilities of the trinucleotides j, $j \neq i$, into the trinucleotide i during $[t, t + dt]$:

$$P_i(t + dt) = P_i(t) - \alpha dt P_i(t) + \alpha dt \sum_{j=1}^{64} P(j \to i) P_j(t) \qquad (1)$$

where α is the probability that a trinucleotide is subjected to one substitution during a unit interval of time and where $P(j \to i)$ is the substitution probability of a trinucleotide j, $j \neq i$, into a trinucleotide i. The probability $P(j \to i)$ is

equal to 0 if the substitution is impossible (j is different from i, and j and i differ by more than one nucleotide as dt is assumed to be small enough so that a codon cannot substitute successively twice during $[t, t + dt]$), otherwise it is given in the function of the three substitution rates p, q and $r = 1 - p - q$ in the three codon sites respectively (matrix A, Table 3). For example with $i = 1$, $P(\text{AAC} \rightarrow \text{AAA}) = P(\text{AAG} \rightarrow \text{AAA}) = P(\text{AAT} \rightarrow \text{AAA}) = r/3$, $P(\text{ACA} \rightarrow \text{AAA}) = P(\text{AGA} \rightarrow \text{AAA}) = P(\text{ATA} \rightarrow \text{AAA}) = q/3$, $P(\text{CAA} \rightarrow \text{AAA}) = P(\text{GAA} \rightarrow \text{AAA}) = P(\text{TAA} \rightarrow \text{AAA}) = p/3$ and $P(j \rightarrow \text{AAA}) = 0$ with $j \notin \{\text{AAC}, \text{AAG}, \text{AAT}, \text{ACA}, \text{AGA}, \text{ATA}, \text{CAA}, \text{GAA}, \text{TAA}\}$. Formula (1) considers only the case of one substitution during dt as the case of several substitutions leads to negligible terms function of dt^2.

Table 3. Substitution matrix A (64, 64) of the 4096 trinucleotide substitutions given in the function of the three substitution rates p, q and $r = 1 - p - q$ of the three codon sites respectively. The lines and the columns correspond to the trinucleotides given in alphabetical order. The matrix A (64, 64) is symmetrical, i.e. the lines or the columns can correspond to the trinucleotides before or after substitutions. The square matrix A can be represented by a square block matrix B (4, 4) whose four diagonal elements are formed by four identical square submatrices M (16, 16) (Table 4) and whose 12 non-diagonal elements are formed by 12 identical square submatrices $(p/3)\text{Id}$ (16, 16).

	1:AAA	...	16:ATT	17:CAA	...	32:CTT	33:GAA	...	48:GTT	49:TAA	...	64:TTT
1:AAA												
⋮	M			$(p/3)\text{Id}$			$(p/3)\text{Id}$			$(p/3)\text{Id}$		
16:ATT												
17:CAA												
⋮	$(p/3)\text{Id}$			M			$(p/3)\text{Id}$			$(p/3)\text{Id}$		
32:CTT												
33:GAA												
⋮	$(p/3)\text{Id}$			$(p/3)\text{Id}$			M			$(p/3)\text{Id}$		
48:GTT												
49:TAA												
⋮	$(p/3)\text{Id}$			$(p/3)\text{Id}$			$(p/3)\text{Id}$			M		
64:TTT												

With an appropriate unit of time, $\alpha = 1$ (i.e. one substitution per codon per unit of time in average and, therefore, the time t is then equivalent to the mean

number of substitutions per codon) and formula (1) becomes

$$\frac{P_i(t+dt)-P_i(t)}{dt} \approx P_i'(t) = -P_i(t) + \sum_{j=1}^{64} P(j \to i) P_j(t). \qquad (2)$$

For example, a sequence of 100 codons after $t = 0.1$ substitutions per codon according to the site proportions $p = 0.2, q = 0.3$ and $r = 1 - 0.2 - 0.3 = 0.5$ means that $100 \times 0.1 = 10$ codons, randomly chosen in the sequence, have mutated, $10 \times 0.2 = 2$ codons in the first site, $10 \times 0.3 = 3$ codons in the second site and $10 \times 0.5 = 5$ codons in the third site.

By considering the column vector $P(t) = [P_i(t)]_{1 \le i \le 64}$ made of the $64 P_i(t)$ and the substitution matrix A $(64, 64)$ of the 4096 trinucleotide substitutions $P(j \to i)$ (Table 3), the differential equation (2) can be represented by the following matrix equation

$$P'(t) = -P(t) + A \cdot P(t) = (A - \mathrm{Id}) \cdot P(t)$$

where the symbol $\cdot$ represents the matrix product.

The real matrix A is symmetrical (see below). Therefore, the real matrix $A - \mathrm{Id}$ is symmetrical. Then, it exists as an eigenvector matrix Q and a diagonal matrix D of characteristic roots $(\lambda_k - 1)$ (where λ_k are the characteristic roots of A ordered in the same way as the eigenvector columns in Q) so that $A - \mathrm{Id} = Q \cdot D \cdot Q^{-1}$. Then,

$$P'(t) = Q \cdot D \cdot Q^{-1} \cdot P(t)$$

$$Q^{-1} \cdot P'(t) = (Q^{-1} \cdot P(t))' = D \cdot Q^{-1} \cdot P(t).$$

This equation has the classical solution

$$Q^{-1} \cdot P(t) = e^{Dt} \cdot Q^{-1} \cdot P(0)$$

where e^{Dt} is the diagonal matrix of exponential characteristic roots $e^{(\lambda_k - 1)t}$. Finally,

$$P(t) = Q \cdot e^{Dt} \cdot Q^{-1} \cdot P(0).$$

The characteristic roots $(\lambda_k - 1)$ of the matrix $A - \mathrm{Id}$ are deduced from the characteristic roots λ_k of the matrix A which can be obtained by determining the roots of the characteristic equation $\det(A - \lambda \mathrm{Id}) = 0$ of A.

Table 4. Square submatrix M (16, 16) forming the four diagonal elements of the square block matrix B (4, 4) of the substitution square matrix A (64, 64) (Table 3). The square matrix M (16, 16) is symmetrical and can be represented by a square block matrix C (4, 4) whose four diagonal elements are formed by four identical square submatrices N (4, 4) (Table 6) and whose 12 non-diagonal elements are formed by 12 identical square submatrices $(q/3)\mathrm{Id}$ (4, 4) (Table 7).

N	$(q/3)\mathrm{Id}$	$(q/3)\mathrm{Id}$	$(q/3)\mathrm{Id}$
$(q/3)\mathrm{Id}$	N	$(q/3)\mathrm{Id}$	$(q/3)\mathrm{Id}$
$(q/3)\mathrm{Id}$	$(q/3)\mathrm{Id}$	N	$(q/3)\mathrm{Id}$
$(q/3)\mathrm{Id}$	$(q/3)\mathrm{Id}$	$(q/3)\mathrm{Id}$	N

Table 5. Square block matrix B (4, 4) (Table 3) after linear combinations.

$M-(p/3)\mathrm{Id}$	0	0	0
0	$M+p\mathrm{Id}$	0	0
0	$(2p/3)\mathrm{Id}$	$M-(p/3)\mathrm{Id}$	0
0	$(p/3)\mathrm{Id}$	0	$M-(p/3)\mathrm{Id}$

The substitution square matrix A (64, 64) (Table 3) can be represented by the square block matrix B (4, 4) whose four diagonal elements are formed by four identical square submatrices M (16, 16) (Table 4) and whose 12 non-diagonal elements are formed by 12 identical square submatrices $(p/3)\mathrm{Id}$ (16, 16) (the product of the identity matrix (16, 16) and the parameter $p/3$). The square block matrix B (4, 4) (Table 3) after linear combinations (Table 5) leads to the following determinant for the matrix $A-\lambda\mathrm{Id}$

$$\begin{aligned}\det(A-\lambda\mathrm{Id}) &= \det{}^3(M-\lambda\mathrm{Id}-(p/3)\mathrm{Id})\det(M-\lambda\mathrm{Id}+p\mathrm{Id})\\ &= \det{}^3(M-(\lambda+p/3)\mathrm{Id})\det(M-(\lambda-p)\mathrm{Id}).\end{aligned}\tag{3}$$

The submatrix M (Table 4) can be represented by the square block matrix C (4, 4) whose four diagonal elements are formed by four identical square submatrices N (4, 4) (Table 6) and whose 12 non-diagonal elements are formed by 12 identical square submatrices $(q/3)\mathrm{Id}$ (4, 4) (Table 7). Therefore, the determinant for the matrix $M-\mu\mathrm{Id}$ is as before

$$\det(M-\mu\mathrm{Id}) = \det{}^3(N-\mu\mathrm{Id}-(q/3)\mathrm{Id})\det(N-\mu\mathrm{Id}+q\mathrm{Id}).$$

By substituting in (3) with $\mu=\lambda+p/3$ or $\mu=\lambda-p$,

$$\begin{aligned}\det(A-\lambda\mathrm{Id}) &= [\det{}^3(N-(\lambda+p/3)\mathrm{Id}-(q/3)\mathrm{Id})\det(N-(\lambda+p/3)\mathrm{Id}+q\mathrm{Id}]^3\\ &\quad\times\det{}^3(N-(\lambda-p)\mathrm{Id}-(q/3)\mathrm{Id})\det(N-(\lambda-p)\mathrm{Id}+q\mathrm{Id})\\ &= \det{}^9(N-(\lambda+p/3+q/3)\mathrm{Id})\det{}^3(N-(\lambda+p/3-q)\mathrm{Id})\\ &\quad\times\det{}^3(N-(\lambda-p+q/3)\mathrm{Id})\det(N-(\lambda-p-q)\mathrm{Id}).\end{aligned}\tag{4}$$

Table 6. Square submatrix N (4, 4) forming the four diagonal elements of the square block matrix C (4, 4) (Table 4).

0	$r/3$	$r/3$	$r/3$
$r/3$	0	$r/3$	$r/3$
$r/3$	$r/3$	0	$r/3$
$r/3$	$r/3$	$r/3$	0

Table 7. Square submatrix $(q/3)\mathrm{Id}$ (4, 4) forming the 12 non-diagonal elements of the square block matrix C (4, 4) (Table 4).

$q/3$	0	0	0
0	$q/3$	0	0
0	0	$q/3$	0
0	0	0	$q/3$

The submatrix N (Table 6) after linear combinations similar to the block matrix B (Table 5) leads to the following determinant for the matrix $N - \mu\mathrm{Id}$:

$$\det(N - \mu\mathrm{Id}) = (-\mu - r/3)^3(-\mu + r).$$

By substituting in (4) with $\mu = \lambda + p/3 + q/3$, $\mu = \lambda + p/3 - q$, $\mu = \lambda - p + q/3$ or $\mu = \lambda - p - q$,

$$\begin{aligned}
\det(A - \lambda\mathrm{Id}) = {} & [(-(\lambda + p/3 + q/3) - r/3)^3(-(\lambda + p/3 + q/3) + r)]^9 \\
& \times [(-(\lambda + p/3 - q) - r/3)^3(-(\lambda + p/3 - q) + r)]^3 \\
& \times [(-(\lambda - p + q/3) - r/3)^3(-(\lambda - p + q/3) + r)]^3 \\
& \times (-(\lambda - p - q) - r/3)^3(-(\lambda - p - q) + r) \\
= {} & (-\lambda - p/3 - q/3 - r/3)^{27}(-\lambda - p/3 - q/3 + r)^9 \\
& \times (-\lambda - p/3 + q - r/3)^9(-\lambda + p - q/3 - r/3)^9 \\
& \times (-\lambda - p/3 + q + r)^3(-\lambda + p - q/3 + r)^3 \\
& \times (-\lambda + p + q - r/3)^3(-\lambda + 1).
\end{aligned}$$

Therefore, there are eight characteristic roots λ_k : $\lambda_1 = 1$ (order of the associated space eigenvector: 1), $\lambda_2 = p + q - r/3$ (order of the associated space eigenvector: 3), $\lambda_3 = p - q/3 + r$ (order of the associated space eigenvector: 3) $\lambda_4 = -p/3 + q + r$ (order of the associated space eigenvector: 3), $\lambda_5 = p - q/3 - r/3$ (order of the associated space eigenvector: 9), $\lambda_6 = -p/3 + q - r/3$ (order of the associated space eigenvector: 9), $\lambda_7 = -p/3 - q/3 + r$ (order of the associated space eigenvector: 9) and $\lambda_8 = -p/3 - q/3 - r/3 = -1/3$ (order of the associated space eigenvector: 27). The eigenvectors associated with these eight characteristic roots λ_k computed by formal calculus are independent of p, q, r (data not shown).

The independent mixing of the 22 trinucleotides of $\mathcal{T}_0$ with equiprobability (1/22) leads to the following initial vector $P(0) = [1/22, 1/22, 0, 1/22, 0, 1/22, 0, 0, 0, 0, 0, 0, 0, 1/22, 0, 1/22, 0, 0, 1/22, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/22, 1/22, 0, 1/22, 1/22, 1/22, 1/22, 0, 1/22, 0, 0, 0, 1/22, 0, 1/22, 1/22, 1/22, 0, 1/22, 0 , 1/22, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/22, 0, 1/22]$. The formula $P(t) = Q \cdot e^{Dt} \cdot Q^{-1} \cdot P(0)$ with the 64 trinucleotide probabilities $P_j(0)$ before the substitution process ($t = 0$), the diagonal matrix of characteristic roots $e^{(\lambda_k - 1)t}$, the eigenvector matrix Q and its inverse Q^{-1}, allows deduction of the 64 trinucleotide probabilities $P_i(t)$ in frame 0 after t substitutions in function of the three substitution rates p, q and $r = 1 - p - q$ of the three codon sites respectively (the indexes i or j representing the trinucleotides AAA, ..., TTT in

the alphabetical order). For example with the trinucleotide AAA ($i = 1$),

$$\begin{aligned}\mathrm{P}_1(t) = \frac{1}{704}[&9e^{-\frac{4}{3}t} + e^{-\frac{4}{3}pt} + 3e^{-\frac{4}{3}(1-p)t} + 7e^{-\frac{4}{3}qt} \\ &+e^{-\frac{4}{3}(1-q)t} + 5e^{-\frac{4}{3}(p+q)t} - 5e^{-\frac{4}{3}(1-p-q)t} + 11].\end{aligned}$$

NOTE. $\lim_{t\to\infty} \mathrm{P}_1(t) = 11/704 = 1/64$.

Therefore, the occurrence probabilities $P(\mathcal{T}_g,\ f = 0, t)$ of the subsets $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ of trinucleotides in the frame $f = 0$ at the substitution step t are equal to $P(\mathcal{T}_g, 0, t) = \sum_{i\in\mathcal{T}_g} P_i(t)$.

After simplification,

$$\begin{aligned}P(\mathcal{T}_0, 0, t) = \frac{1}{352}[&61e^{-\frac{4}{3}t} + 33e^{-\frac{4}{3}pt} + 13e^{-\frac{4}{3}(1-p)t} + 49e^{-\frac{4}{3}qt} \\ &+29e^{-\frac{4}{3}(1-q)t} + 13e^{-\frac{4}{3}(p+q)t} + 33e^{-\frac{4}{3}(1-p-q)t} + 121] \\ P(\mathcal{T}_1, 0, t) = \frac{1}{704}[&-61e^{-\frac{4}{3}t} - 29e^{-\frac{4}{3}pt} - 25e^{-\frac{4}{3}(1-p)t} - 49e^{-\frac{4}{3}qt} \\ &-29e^{-\frac{4}{3}(1-q)t} - e^{-\frac{4}{3}(p+q)t} - 37e^{-\frac{4}{3}(1-p-q)t} + 231] \\ P(\mathcal{T}_2, 0, t) = \frac{1}{704}[&-61e^{-\frac{4}{3}t} - 37e^{-\frac{4}{3}pt} - e^{-\frac{4}{3}(1-p)t} - 49e^{-\frac{4}{3}qt} \\ &-29e^{-\frac{4}{3}(1-q)t} - 25e^{-\frac{4}{3}(p+q)t} - 29e^{-\frac{4}{3}(1-p-q)t} + 231]\end{aligned}$$

NOTES.

$$\sum_{g=0,1,2} P(\mathcal{T}_g, 0, t) = 1$$

$$\lim_{t\to\infty} P(\mathcal{T}_0, 0, t) = 121/352 = 22/64$$

and

$$\lim_{t\to\infty} P(\mathcal{T}_1, 0, t) = \lim_{t\to\infty} P(\mathcal{T}_2, 0, t) = 231/704 = 21/64$$

(remember that the subsets $\mathcal{T}_0$, $\mathcal{T}_1$ and $\mathcal{T}_2$ have 22, 21 and 21 trinucleotides respectively).

The occurrence probabilities $P(\mathcal{T}_g,\ f = 1, t)$ (resp. $P(\mathcal{T}_g,\ f = 2, t)$) of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the shifted frame $f = 1$ (resp. $f = 2$) at the substitution step t are obtained by determining the 64 trinucleotide probabilities $P(i,\ f = 1, t)$ (resp. $P(i,\ f = 2, t)$) in the shifted frame $f = 1$ (resp. $f = 2$). The probability $P(i, 1, t)$ (or $P(i, 2, t)$) of a trinucleotide i given in alphabetical order in frame 1 (or 2) is obtained from the product of the two probabilities $P(j, 0, t)$ and $P(k, 0, t)$ associated with the concatenation of the two trinucleotides j and k in frame 0 generating the trinucleotide i in frame 1 (or 2). For example, the

trinucleotide AAA ($i = 1$) in frame $f = 1$ is obtained by the concatenation of the two types of trinucleotides NAA and ANN in frame 0. Therefore, the probability $P(i = 1, 1, t)$ of the trinucleotide AAA ($i = 1$) in the frame $f = 1$ is equal to the product of the probability of the trinucleotides NAA, i.e. AAA, CAA, GAA or TAA ($j = 1, 17, 33, 49$) in frame 0 and the probability of the trinucleotides ANN ($k = 1, \ldots, 16$) in frame 0. Similarly, the probability $P(i = 1, 2, t)$ of the trinucleotide AAA ($i = 1$) in the frame $f = 2$ is equal to the product of the probability of the trinucleotides NNA ($j = 1 + 4 \times j'$ with $j' = 0, \ldots, 15$) in frame 0 and the probability of the trinucleotides AAN ($k = 1, \ldots, 4$) in frame 0:

$$P(2, 1, t) = \sum_{j=1,17,33,49} P(j, 0, t) \times \sum_{k=1,\ldots,16} P(k, 0, t)$$

and

$$P(2, 2, t) = \sum_{\substack{j=1+4\times j' \\ j'=0,\ldots,15}} P(j, 0, t) \times \sum_{k=1,\ldots,4} P(k, 0, t)$$

After simplification,

$$\begin{aligned}
P(\mathcal{T}_0, 1, t) &= \frac{1}{3872}[-15e^{-\frac{4}{3}t} - 275e^{-\frac{4}{3}pt} + 33e^{-\frac{4}{3}(1-p)t} - 154e^{-\frac{4}{3}qt} \\
&\quad -20e^{-\frac{4}{3}(1-q)t} - 98e^{-\frac{4}{3}(p+q)t} - 154e^{-\frac{4}{3}(1-p-q)t} + 1331] \\
P(\mathcal{T}_1, 1, t) &= \frac{1}{7744}[21e^{-\frac{4}{3}t} + 715e^{-\frac{4}{3}pt} + 231e^{-\frac{4}{3}(1-p)t} + 847e^{-\frac{4}{3}qt} \\
&\quad +115e^{-\frac{4}{3}(1-q)t} + 133e^{-\frac{4}{3}(p+q)t} + 869e^{-\frac{4}{3}(1-p-q)t} + 2541] \\
P(\mathcal{T}_2, 1, t) &= \frac{1}{7744}[9e^{-\frac{4}{3}t} - 165e^{-\frac{4}{3}pt} - 297e^{-\frac{4}{3}(1-p)t} - 539e^{-\frac{4}{3}qt} \\
&\quad -75e^{-\frac{4}{3}(1-q)t} + 63e^{-\frac{4}{3}(p+q)t} - 561e^{-\frac{4}{3}(1-p-q)t} + 2541] \\
P(\mathcal{T}_0, 2, t) &= \frac{1}{3872}[-15e^{-\frac{4}{3}t} - 154e^{-\frac{4}{3}pt} - 98e^{-\frac{4}{3}(1-p)t} - 154e^{-\frac{4}{3}qt} \\
&\quad -20e^{-\frac{4}{3}(1-q)t} + 33e^{-\frac{4}{3}(p+q)t} - 275e^{-\frac{4}{3}(1-p-q)t} + 1331] \\
P(\mathcal{T}_1, 2, t) &= \frac{1}{7744}[9e^{-\frac{4}{3}t} - 561e^{-\frac{4}{3}pt} + 63e^{-\frac{4}{3}(1-p)t} - 539e^{-\frac{4}{3}qt} \\
&\quad -75e^{-\frac{4}{3}(1-q)t} - 297e^{-\frac{4}{3}(p+q)t} - 165e^{-\frac{4}{3}(1-p-q)t} + 2541] \\
P(\mathcal{T}_2, 2, t) &= \frac{1}{7744}[21e^{-\frac{4}{3}t} + 869e^{-\frac{4}{3}pt} + 133e^{-\frac{4}{3}(1-p)t} + 847e^{-\frac{4}{3}qt} \\
&\quad +115e^{-\frac{4}{3}(1-q)t} + 231e^{-\frac{4}{3}(p+q)t} + 715e^{-\frac{4}{3}(1-p-q)t} + 2541].
\end{aligned}$$

NOTES.

$$\sum_{g=0,1,2} P(\mathcal{T}_g, 1, t) = \sum_{g=0,1,2} P(\mathcal{T}_g, 2, t) = 1$$

$$\lim_{\substack{t\to\infty \\ f=1,2}} P(\mathcal{T}_0,\ f, t) = 1331/3872 = 22/64$$

and

$$\lim_{\substack{t\to\infty \\ f=1,2}} P(\mathcal{T}_1,\ f, t) = \lim_{\substack{t\to\infty \\ f=1,2}} P(\mathcal{T}_2,\ f, t) = 2541/7744 = 21/64.$$

The numerical results obtained with these analytical solutions have been verified by computer simulation (simulation of random substitutions in simulated sequences, see Section 4).

The model (p, q, t) has a solution if, for given values of the two site substitution parameters p and q, there are values of the codon substitution parameter t verifying the three inequality sets $\mathcal{Q}(\mathrm{P}_0)$: $P(\mathcal{T}_0, \mathrm{P}_0) > P(\mathcal{T}_1, \mathrm{P}_0) > P(\mathcal{T}_2, \mathrm{P}_0)$, $\mathcal{Q}(\mathrm{P}_1)$: $P(\mathcal{T}_1, \mathrm{P}_1) > P(\mathcal{T}_2, \mathrm{P}_1) > P(\mathcal{T}_0, \mathrm{P}_1)$, $\mathcal{Q}(\mathrm{P}_2)$: $P(\mathcal{T}_2, \mathrm{P}_2) > P(\mathcal{T}_0, \mathrm{P}_2) > P(\mathcal{T}_1, \mathrm{P}_2)$ and the frequency order of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ associated with the three frames of actual protein genes and if there are higher values of t verifying the inequality set $\mathcal{Q}(\mathrm{R}_f)$: $P(\mathcal{T}_0, \mathrm{R}_f) > P(\mathcal{T}_1, \mathrm{R}_f) \approx P(\mathcal{T}_2, \mathrm{R}_f)$ and the frequency order of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ with any frame f of actual 5′ and 3′ regions (by assuming that the 5′ and 3′ regions have a greater number of substitutions compared with the protein genes).

3.3. *Results.*

3.3.1. *Simulation of protein coding genes of eukaryotes.* By varying the two parameters p and q in the range [0, 1] with a step of 0.05 and the parameter t in the range [0, 30] with a step of 0.1, the model (p, q, t) retrieves the three inequality sets $\mathcal{Q}(\mathrm{P}_0)$, $\mathcal{Q}(\mathrm{P}_1)$, $\mathcal{Q}(\mathrm{P}_2)$ of actual protein genes and the inequality set $\mathcal{Q}(\mathrm{R}_f)$ of the actual 5′ and 3′ regions when $p = 0.1 \pm 0.05, q = 0.1 \pm 0.05$ and $r = 1 - p - q = 0.8 \pm 0.1$.

At the construction process ($t = 0$), the model ($p = 0.1, q = 0.1, t = 0$) leads to the following expected probabilities, in frame 0: $P(\mathcal{T}_0, 0, 0) = 1$ and $P(\mathcal{T}_1, 0, 0,) = P(\mathcal{T}_2, 0, 0) = 0$ (Fig. 6a), in frame 1: $P(\mathcal{T}_1, 1, 0) = 0.708$, $P(\mathcal{T}_0, 1, 0) = 0.167$ and $P(\mathcal{T}_2, 1, 0) = 0.125$ (Fig. 6b) and in frame 2: $P(\mathcal{T}_2, 2, 0) = 0.708$, $P(\mathcal{T}_0, 2, 0) = 0.167$ and $P(\mathcal{T}_1, 2, 0) = 0.125$ (Fig. 6c), i.e. to the following inequalities, in frame 0: $P(\mathcal{T}_0, 0, 0,) > P(\mathcal{T}_1, 0, 0) = P(\mathcal{T}_2, 0, 0)$, in frame 1: $P(\mathcal{T}_1, 1, 0) > P(\mathcal{T}_0, 1, 0) > P(\mathcal{T}_2, 1, 0)$ and in frame 2: $P(\mathcal{T}_2, 2, 0) > P(\mathcal{T}_0, 2, 0) > P(\mathcal{T}_1, 2, 0)$ which result from the complementarity property of the C^3 code $\mathcal{X}_0$. These probability inequalities before the substitution process differ from the inequalities $\mathcal{Q}(\mathrm{P}_0)$ and $\mathcal{Q}(\mathrm{P}_1)$ of protein genes and from the inequality set $\mathcal{Q}(\mathrm{R}_f)$ of the 5′ and 3′ regions. The construction process only simulates the inequality $\mathcal{Q}(\mathrm{P}_2)$ in frame 2 of protein genes.

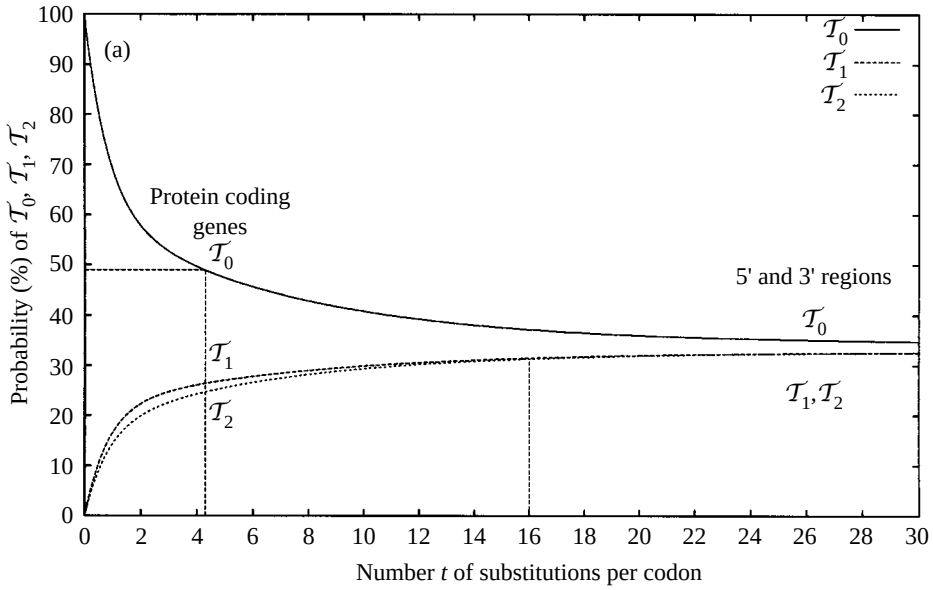

Figure 6. (a) Probability $P(\mathcal{T}_g, 0, t)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ in frame 0 generated with an independent mixing of 22 trinucleotides of $\mathcal{T}_0$ with equiprobability (1/22) and subjected to t substitutions per codon according to the proportions $p = 0.1$ $q = 0.1$ and $r = 1 - p - q = 0.8$ in the three codon sites respectively. The inequality $\mathcal{Q}(\mathrm{P}_0)$ in frame 0 of protein coding genes is verified for a substitution number t in the range $[0, \approx 16]$. At $t = 4.3$ substitutions, the three analytical curves have the occurrence probability orders of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ similar to those observed in frame 0 of protein coding genes. The inequality $\mathcal{Q}(\mathrm{R}_0)$ in frame 0 of the $5'$ and $3'$ regions is verified for a substitution number $t >\approx 16$. (b) Probability $P(\mathcal{T}_g, 1, t)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ in frame 1 generated with an independent mixing of the 22 trinucleotides of $\mathcal{T}_0$ with equiprobability (1/22) and subjected to t substitutions per codon according to the proportions $p = 0.1$ $q = 0.1$ and $r = 1 - p - q = 0.8$ in the three codon sites respectively. The inequality $\mathcal{Q}(\mathrm{P}_1)$ in frame 1 of protein-coding genes is verified for a substitution number t in the range $]0.9, 5.9]$. At $t = 4.3$ substitutions, the three analytical curves have the occurrence probability orders of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ similar to those observed in frame 1 of protein coding genes. The inequality $\mathcal{Q}(\mathrm{R}_1)$ in frame 1 of the $5'$ and $3'$ regions is verified for a substitution number $t > 22.5$. (c) Probability $P(\mathcal{T}_g, 2, t)$ of $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ in frame 2 generated with an independent mixing of the 22 trinucleotides of $\mathcal{T}_0$ with equiprobability (1/22) and subjected to t substitutions per codon according to the proportions $p = 0.1$ $q = 0.1$ and $r = 1 - p - q = 0.8$ in the three codon sites respectively. The inequality $\mathcal{Q}(\mathrm{P}_2)$ in frame 2 of protein coding genes is verified for a substitution number t in the range [0, 22.2]. At $t = 4.3$ substitutions, the three analytical curves have the occurrence probability orders of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ similar to those observed in frame 2 of protein coding genes. The inequality $\mathcal{Q}(\mathrm{R}_2)$ in frame 2 of the $5'$ and $3'$ regions is verified for a substitution number $t > 22.2$.

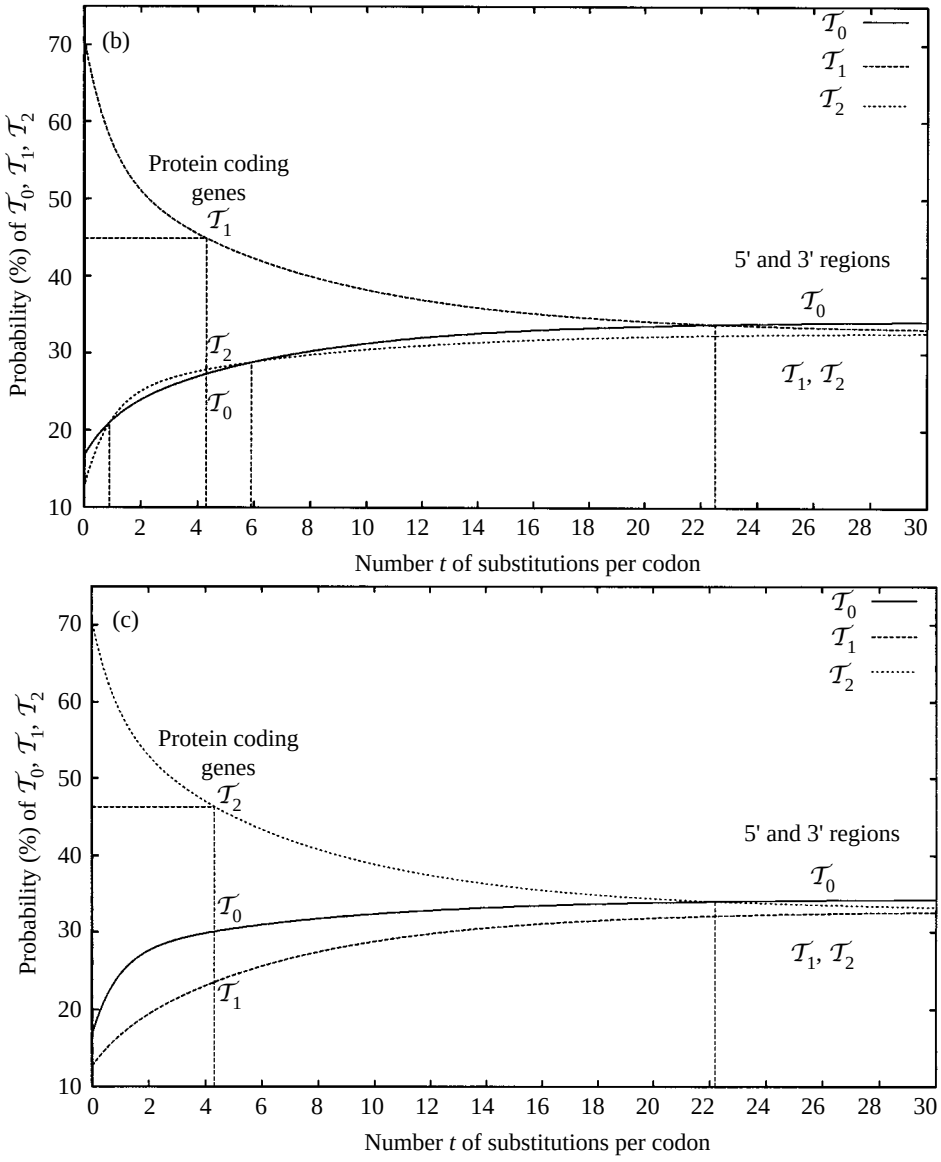

Figure 6. Continued.

The inequality $\mathcal{Q}(\mathrm{P}_0)$ in frame 0 of protein genes exists in the model $(p = 0.1, q = 0.1, t)$ for a substitution number t in the range $]0, \approx 16]$ (Fig. 6a). The inequality $\mathcal{Q}(\mathrm{P}_1)$ in frame 1 of protein genes exists in the model $(0.1, 0.1, t)$ for t in the range]0.9, 5.9] (Fig. 6b). The inequality $\mathcal{Q}(\mathrm{P}_2)$ in frame 2 of protein genes exists in the model $(0.1, 0.1, t)$ for t in the range [0, 22.2] (Fig. 6c). Therefore, the range of substitutions simultaneously verifying these three inequalities in the three frames of protein genes is determined by the range of substitutions in frame 1, i.e. t in the range]0.9, 5.9]. Furthermore, the model $(0.1, 0.1, t)$ retrieves the probability order of $\mathcal{T}_0$ in frame 0, $\mathcal{T}_1$ in frame 1 and $\mathcal{T}_2$ in frame 2 at $t = 4.3 : P(\mathcal{T}_0, \mathrm{P}_0) = 0.49$ (Table 2a) and $P(\mathcal{T}_0, 0, 4.3) \approx 0.49$ (Fig. 6a), $P(\mathcal{T}_1, \mathrm{P}_1) = 0.43$ (Table 2a) and $P(\mathcal{T}_1, 1, 4.3) \approx 0.45$ (Fig. 6b) and $P(\mathcal{T}_2, \mathrm{P}_2) = 0.45$ (Table 2a) and $P(\mathcal{T}_2, 2, 4.3) \approx 0.462$ (Fig. 6c). Note that the model $(0.1, 0.1, t)$ allows simulation of another inequality observed in protein genes: $P(\mathcal{T}_0, \mathrm{P}_0) > P(\mathcal{T}_2, \mathrm{P}_2) > P(\mathcal{T}_1, \mathrm{P}_1)$.

3.3.2. *Simulation of the* 5′ *and* 3′ *regions of eukaryotes.* By increasing the number t of substitutions, the model $(0.1, 0.1, t)$ allows simulation of the inequality $\mathcal{Q}(\mathrm{R}_f) : P(\mathcal{T}_0, \mathrm{R}_f) > P(\mathcal{T}_1, \mathrm{R}_f) \approx P(\mathcal{T}_2, \mathrm{R}_f)$ observed in the three frames of the actual 5′ and 3′ regions.

(i) In frame 0: $P(\mathcal{T}_1, 0, t) \approx P(\mathcal{T}_2, 0, t)$ for $t > \approx 16$ (Fig. 6a). As $P(\mathcal{T}_0, 0, t) > P(\mathcal{T}_1, 0, t)$ and $P(\mathcal{T}_0, 0, t) > P(\mathcal{T}_2, 0, t)$ whatever t, the inequality $\mathcal{Q}(\mathrm{R}_0)$ in frame 0 of the 5′ and 3′ regions is simulated with the model $(0.1, 0.1, t)$ for $t > \approx 16$.

(ii) In frame 1: $P(\mathcal{T}_0, 1, t) > P(\mathcal{T}_1, 1, t)$ for $t > 22.5$ (Fig. 6b). Therefore, the inequality $\mathcal{Q}(\mathrm{R}_1)$ in frame 1 of the 5′ and 3′ regions is simulated with the model $(0.1, 0.1, t)$ for $t > 22.5$.

(iii) In frame 2: $P(\mathcal{T}_0, 2, t) > P(\mathcal{T}_2, 2, t)$ for $t > 22.2$ (Fig. 6c). Therefore, the inequality $\mathcal{Q}(\mathrm{R}_2)$ in frame 2 of the 5′ and 3′ regions is simulated with the model $(0.1, 0.1, t)$ for $t > 22.2$.

The range of substitutions simultaneously verifying these three inequalities in the three frames of the 5′ and 3′ regions is determined by the range of substitutions in frames 1 and 2, i.e. $t > \approx 22$.

4. DISCUSSION

The subset $\mathcal{T}_0$ of trinucleotides (Table 1) has a preferential occurrence in protein genes (frame 0) of prokaryotes and eukaryotes, and the rarity property (6×10^{-8}) to contain a complementary maximal circular code $\mathcal{X}_0$ with two permutated maximal circular codes $\mathcal{X}_1$ and $\mathcal{X}_2$ (C^3 code, Section 1.3). The quantitative study of the three subsets $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the three frames 0, 1, 2 of protein genes, and the 5′ and 3′ regions of eukaryotes, has shown that their occurrence frequencies are

constant for each codon position in the sequences. In protein genes, the frequencies of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in frame 0 are 49, 28.5 and 22.5% respectively (Table 2a). In the 5′ (resp. 3′) regions, these frequencies are 35.5% (resp. 38%), 32.5% (resp. 31%) and 32% (resp. 31%) (Tables 2b, c). This property leads to an application at the sequence level. Each sequence in a population is classified in $\mathcal{T}_0, \mathcal{T}_1$ or $\mathcal{T}_2$ according to its greatest number of codons belonging to $\mathcal{T}_0, \mathcal{T}_1$ or $\mathcal{T}_2$. In the eukaryotic protein gene population, 94% of sequences are classified in $\mathcal{T}_0$, 5% of sequences in $\mathcal{T}_1$ and 1% of sequences in $\mathcal{T}_2$. In contrast, in the eukaryotic 5′ (resp. 3′) region populations, 44% (resp. 60%) of sequences are classified in $\mathcal{T}_0$, 34% (resp. 20%) of sequences in $\mathcal{T}_1$ and 22% (resp. 20%) of sequences in $\mathcal{T}_2$. By including these values in frame 0 and also those which can be obtained in the shifted frames (to improve the significance) in some statistical tests, this application could be used to discriminate protein coding and non-coding genes and could be added to the other discriminating tests (e.g. Shulman *et al.*, 1981, Shepherd, 1981; Staden and McLachlan, 1982; Fickett, 1982; Smith *et al.*, 1983; Blaisdell, 1983).

The evolutionary model tested has a solution correlated with the reality observed in the protein genes, and the 5′ and 3′ regions. Its biological meaning would suggest that these three types of sequences before substitution ($t = 0$) are constructed by trinucleotides. Only 22 among 64 trinucleotides would have been necessary. The 22 types of trinucleotides as well as the type of their concatenation are determined in the model. Indeed, the 22 trinucleotides are defined by the subset $\mathcal{T}_0$ which contains a C^3 code with concatenation properties of flexibility (Section 1.3) allowing its evolution. The independent concatenation of these 22 trinucleotides with equiprobability is the simplest type of conatenation and therefore, compatible with a primitive stage of gene evolution. A Markov concatenation of trinucleotides would have been too complex at this time. The model also demonstrates that a substitution process ($t > 0$) must follow the construction process in order to simulate the protein genes, and the 5′ and 3′ regions. The substitution process allows generation of new and totally unexpected inequalities, e.g. the substitution process in frame 1 creates four successive inequalities: $P(\mathcal{T}_1, 1, t) > P(\mathcal{T}_0, 1, t) > P(\mathcal{T}_2, 1, t)$, $P(\mathcal{T}_1, 1, t) > P(\mathcal{T}_2, 1, t) > P(\mathcal{T}_0, 1, t)$, then again $P(\mathcal{T}_1, 1, t) > P(\mathcal{T}_0, 1, t) > P(\mathcal{T}_2, 1, t)$ and finally $P(\mathcal{T}_0, 1, t) > P(\mathcal{T}_1, 1, t) \approx P(\mathcal{T}_2, 1, t)$ (Fig. 6b). Furthermore, the substitution process decreases the initial probabilities of $P(\mathcal{T}_0, 0, t)$, $P(\mathcal{T}_1, 1, t)$, $P(\mathcal{T}_2, 2, t)$ in the frames 0, 1, 2 respectively.

The evolutionary model ($p = 0.1, q = 0.1, t = 4.3$) allows simulation of the protein genes by retrieving not only the three sets $\mathcal{Q}(\mathrm{P}_0)$, $\mathcal{Q}(\mathrm{P}_1)$ and $\mathcal{Q}(\mathrm{P}_2)$ of inequalities and the frequency order of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the three frames respectively but also several other sets of probability inequalities observed in protein genes (Table 2a): $P(\mathcal{T}_0, \mathrm{P}_0) > P(\mathcal{T}_2, \mathrm{P}_2) > P(\mathcal{T}_1, \mathrm{P}_1) > P(\mathcal{T}_0, \mathrm{P}_2) > P(\mathcal{T}_2, \mathrm{P}_1)$ and $(P(\mathcal{T}_0, \mathrm{P}_2) - P(\mathcal{T}_1, \mathrm{P}_2)) > (P(\mathcal{T}_1, \mathrm{P}_0) - P(\mathcal{T}_2, \mathrm{P}_0)) > (P(\mathcal{T}_2, \mathrm{P}_1) - P(\mathcal{T}_0, \mathrm{P}_1))$ (numerical results of the analytical solutions not shown). However, the model is

insufficient to simulate the inequalities associated with the four lowest probabilities $P(\mathcal{T}_1, \mathrm{P}_0)$, $P(\mathcal{T}_0, \mathrm{P}_1)$, $P(\mathcal{T}_2, \mathrm{P}_0)$ and $P(\mathcal{T}_1, \mathrm{P}_2)$. Similarly, all observed probabilities cannot exactly be simulated, e.g. the analytical probability $P(\mathcal{T}_1, 1, 4.3) \approx 0.448$ (Fig. 6b) is greater than the observed probability $P(\mathcal{T}_1, \mathrm{P}_1) = 0.43$ (Table 2a) which is obtained in the model at $t \approx 5.4$ (Fig. 6b). The model proposed can be improved, for example by forbidding the generation of a stop trinucleotide TAA, TAG or TGA in frame 0 or by suppressing the strong constraint of a constant proportion of substitutions in the three codon sites during all the substitution process. Nevertheless, the investigation of simple models in a first approach is essential for reducing the great number of possible combinations and secondly for obtaining properties which can be used afterwards to develop more general models containing the simple models. The first hypothesis of model improvement has been tested. An evolutionary model forbidding the generation of a stop trinucleotide TAA, TAG or TGA in frame 0, developed by a numerical model (with the substitution matrix (61, 61) there is no formula giving the eigenvectors and the characteristic roots in a function of the parameters p and q, i.e. the eigenvectors and the characteristic roots must be explicitly determined for each value of the doublet (p, q) in the scanning) or by computer simulation (simulation of random substitutions in simulated sequences), does not improve the results obtained with the analytical model developed here (data not shown).

The substitutions in the model ($p = 0.1, q = 0.1, t = 4.3$) occur with the highest rate in the third codon site (0.8 representing $4.3 \times 0.8 \approx 3.5$ transformations), as expected with the degeneracy of the genetic code. They must also occur in the first and second codon sites but at a weaker rate (0.1 representing $4.3 \times 0.1 \approx 0.5$ transformations for both sites). An evolutionary process with substitutions in the third codon site only ($p = q = 0$ and $r = 1 - p - q = 1$) does not lead to a similarity with the reality observed in protein genes (data not shown).

The evolutionary model ($p = 0.1, q = 0.1, t > \approx 22$) allows simulation of the 5′ and 3′ regions by retrieving the set $\mathcal{Q}(\mathrm{R}_f)$ of inequalities whatever the frame. According to this model, the 5′ and 3′ regions have a greater number of substitutions compared with the protein genes. The absence of statistical property associated with a frame observed in the study of the reality of the 5′ and 3′ regions, as well as the existence of a higher evolutionary process in this model simulating these regions, may be explained by the absence of a protein coding function in these regions. Finally, as the probability of $\mathcal{T}_0$ in the 5′ regions (0.355) is less than the probability of $\mathcal{T}_0$ in the 3′ regions (0.38) (Tables 2b, c), a greater number of substitutions may have occurred in the 5′ regions compared with the 3′ regions.

The complex behaviour of these analytical curves giving the trinucleotide probabilities under a random evolutionary process, see for example Fig. 6b, is totally unexpected and implies two remarks. It is impossible to predict the relative variations of trinucleotides after substitutions without modelling. On the other hand, even after a great number of substitutions, e.g. 4 and 22 substitutions per

codon for the protein genes and the regions respectively (Fig. 6a–c), the trace of primitive disparities between the trinucleotide probabilities are conserved in the actual genes simulated with the model ($p = 0.1, q = 0.1, t$) and correlated with the real actual genes. With other parameters p and q, the behaviour of the analytical curves is completely different and does not lead to a correlation with the reality observed (data not shown).

As mentioned in Section 3.2, the time t is equivalent to a mean number of substitutions per codon. Therefore, the analytical probabilities of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the three frames after t substitutions can be approximated by computing the occurrence probabilities of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the three frames in a simulated population S (having, for example, 100 sequences of 3000 base length to get significant statistical results) which is generated according to an independent mixing of the 22 trinucleotides of $\mathcal{T}_0$ with equiprobability (1/22) (construction process, i.e. $t = 0$) and subjected to t substitutions per codon according to the given site proportions p and q (r being the complement to 1) randomly applied to each sequence of S (substitution process, i.e. $t > 0$).

Furthermore, the replacement of t by $-t$ in the analytical probabilities allows inversion of the evolutionary sense (from the present to the past), i.e. to analyze the probabilities of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ in the three frames after back substitutions. In this case, the trinucleotide probabilities $P_i(0)$ before the substitution process ($t = 0$) are the trinucleotide probabilities of actual genes. These probabilities are known and can be obtained from gene databases. It should also be stressed that the trinucleotide probabilities after back substitutions can only be obtained by analytical solution and not by computer simulation. Indeed, as the site, the type and the order of previous substitutions are unknown, it is impossible to reproduce by simulation the effects of back substitutions in the nucleotide series of actual genes (detailed in Arquès and Michel, 1994). This approach analyzing the analytical probabilities of $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2$ after back substitutions in protein genes, is currently in investigation. Finally, these analytical solutions could be used in the phylogenetic tree reconstruction and the sequence alignment.

ACKNOWLEDGEMENTS

We thank the referees for their advice. This work was supported by GIP GREG grant (Groupement d'Intérêt Public, Groupement de Recherches et d'Etudes sur les Génomes) and Mr Jean-Marc Vassards (Director of the society RVH, Mulhouse).

REFERENCES

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

Arquès, D. G. and C. J. Michel (1990). 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. and C. J. Michel (1992). A simulation of the genetic periodicities modulo 2 and 3 with processes of nucleotide insertions and deletions. *J. Theor. Biol.* **156**, 113–127.

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

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

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

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

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

Benne, R. (1989). RNA-editing in trypanosome mitochondria. *Biochem. Biophys. Acta* **1007**, 131–139.

Benne, R., J. Van Den Burg, J. P. J. Brakenhoff, P. Sloof, J. H. Van Boom and M. C. Tromp (1986). Major transcript of the frameshifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**, 819–826.

Berstel, J. and D. Perrin (1985). *Theory of Codes*. New York: Academic Press.

Blaisdell, B. E. (1983). A prevalent persistent nonrandomness that distinguishes coding and non-coding eukaryotic nuclear DNA sequences. *J. Mol. Evol.* **19**, 122–133.

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

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

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

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

Feagin, J. E. (1990). RNA editing in kinetoplastid mitochondria. *J. Biol. Chem.* **265**, 19373–19376.

Feagin, J. E., J. M. Abraham and K. Stuart (1988). Extensive editing of the cytochrome c oxidase III transcript in trypanosoma brucei. *Cell* **53**, 413–422.

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

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

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

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

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

Shaw, J. M., J. E. Feagin, K. Stuart and L. Simpson (1988). Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. *Cell* **53**, 401–411.

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

Shulman, M. J., C. M. Steinberg and N. Westmoreland (1981). The coding function of nucleotide sequences can be discerned by statistical analysis. *J. Theor. Biol.* **88**, 409–420.

Simpson, L. (1990). RNA editing—A novel genetic phenomenon? *Science* **250**, 512–513.

Smith, T. F., M. S. Waterman and J. R. Sadler (1983). Statistical characterization of nucleic acid sequence functional domains. *Nucl. Acids Res.* **11**, 2205–2220.

Staden, R. and A. D. McLachlan (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucl. Acids Res.* **10**, 141–156.

Stuart, K. (1991). RNA editing in mitochondrial mRNA of trypanosomatids. *Trends Biochem. Sci.* **16**, 68–72.

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

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

Received 16 February 1997 and accepted 19 November 1997