BioSystems 49 (1999) 83–103

An evolutionary analytical model of a complementary circular code

Didier G. Arquès [a,*], Jean-Paul Fallot [b], Laurent Marsan [a], Christian J. Michel [b]

[a] *Equipe de Biologie Théorique, Université de Marne la Vallée, Institut Gaspard Monge, 2 rue de la Butte Verte, 93160 Noisy le Grand, France*

[b] *Equipe de Biologie Théorique, Institut Polytechnique de Sévenans, Rue du Château à Sévenans, 90010 Belfort, France*

Received 22 July 1997; received in revised form 14 April 1998; accepted 19 May 1998

Abstract

The subset $X_0 = \{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,GAG,GAT,GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC\}$ of 20 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. This subset X_0 is a complementary maximal circular code with two permutated maximal circular codes X_1 and X_2 in the frames 1 and 2 respectively (frame 0 shifted by one and two nucleotides respectively in the 5′–3′ direction). X_0 is called a C^3 code (Arquès and Michel, 1997, J. Biosyst 44, 107–134). A quantitative study of these three subsets X_0, X_1 and X_2 in the three frames 0, 1 and 2 of eukaryotic protein genes shows that their occurrence frequencies are constant functions of the trinucleotide positions in the sequences. The frequencies of X_0, X_1 and X_2 in the frame 0 of eukaryotic protein genes are 48.5%, 29% and 22.5% respectively. These properties are not observed in the 5′ and 3′ regions of eukaryotes where X_0, X_1 and X_2 occur with variable frequencies around the random value (1/3). Several frequency asymmetries unexpectedly observed, e.g. the frequency difference between X_1 and X_2 in the frame 0, are related to a new property of the C^3 code X_0 involving substitutions. An evolutionary analytical model at three parameters (p, q, t) based on an independent mixing of the 20 codons (trinucleotides in the frame 0) of X_0 with equiprobability (1/20) 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 X_0, X_1 and X_2 observed in the three frames of protein genes and explains these asymmetries. The complex behaviour of these analytical curves is totally unexpected and a priori difficult to imagine. Finally, the evolutionary analytical method developed could be applied to the phylogenetic tree reconstruction and the DNA sequence alignment.

Keywords: Circular code; Protein coding gene; Frequency asymmetry; Evolutionary analytical model; Substitution; Trinucleotide probability

* Corresponding author. Tel: +1 49 329010; fax: +1 49 329138; e-mail: arques@univ-mlv.fr

0303-2647/99/$ - see front matter
PII S0303-2647(98)00038-0

1. Introduction

1.1. Previous results

Recently, a statistical analysis with 12288 autocorrelation functions has identified three subsets of 20 trinucleotides which occur preferentially in the three frames of protein (coding) genes (Arquès and Michel, 1997): $X_0 = \{AAC,AAT,ACC,ATC, ATT,CAG,CTC,CTG,GAA,GAC,GAG,GAT, GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC\}$ in frame 0 (reading frame established by the ATG start trinucleotide), X_1 and X_2 in the shifted frames 1 and 2 respectively with X_1 and X_2 defined in Table 1 (frames 1 and 2 being the frame 0 shifted by one and two nucleotides respectively in the 5′–3′ direction). The same three subsets X_0, X_1 and X_2 are retrieved with a very few exceptions for the two large protein gene populations of prokaryotes (13686 sequences, 4708758 trinucleotides) and eukaryotes (26757 sequences, 11397678 trinucleotides). These three subsets have several important properties (detailed in Arquès and Michel, 1997), in particular:

1. X_0, X_1 and X_2 are maximal (20 trinucleotides) circular codes: X_0 (resp. X_1, X_2) allows the automatic retrieval of the frame 0 (resp. 1, 2) in any region of a protein gene model formed by a series of trinucleotides of X_0 (resp. X_1, X_2). This property allows the avoidance of motifs for locating the reading frame, such as the start codon ATG in the actual protein genes;
2. the DNA complementarity property (Watson and Crick, 1953, e.g. the complementary trinucleotide of AAC is GTT): X_0 is self-complementary (ten trinucleotides of X_0 are complementary to the ten other trinucleotides of X_0) and, X_1 and X_2 are complementary to each other (the 20 trinucleotides of X_1 are complementary to the 20 trinucleotides of X_2). This property allows the two paired reading frames of a DNA double helix to code simultaneously for amino acids, in agreement with biological results (Zull and Smith, 1990; Konecny et al., 1993; Béland and Allen, 1994; Konecny et al., 1995);
3. the circular permutation property (e.g. the permutated trinucleotide of AAC is ACA): X_0 generates X_1 by one circular permutation and X_2 by another circular permutation (one and two circular permutations with each trinucleotide of X_0 lead to the trinucleotides of X_1 and X_2 respectively) implying that X_1 and X_2 can be deduced from X_0.

In summary, the subset X_0 of 20 trinucleotides identified in protein genes of prokaryotes and eukaryotes is a complementary maximal circular code with two permutated maximal circular codes X_1 and X_2 in the frames 1 and 2 respectively. X_0 is called a C^3 code (Arquès and Michel, 1997).

As the property (1) of circular code is important and also unusual, Section 1.2 introduces and explains the concept of circular code from a biological and computer point of view.

1.2. Concept of circular code

1.2.1. From a biological point of view

The biological concept of circular code, called code without commas, is presented according to a historical background. This concept was introduced by Crick et al. (1957). The code without commas is a code readable in only one frame and without start signal. Such a theoretical code 'without commas' is a set X of codons so that their concatenation (series of codons) leads to genes which have the interesting property to automatically retrieve the concatenation of codons of X without the usage of a start codon in the case of the trace of this initial concatenation is lost (the 'commas' dividing the series of nucleotides into groups of three for constituting the codons in the initial concatenation are lost). Such a code was proposed in order to explain how the reading of a series of nucleotides in the protein genes could code for the amino acids constituting the proteins. The two problems stressed were: why there are more codons than amino acids and how to choose the reading frame. For example, a series of nucleotides ...AGTCCGTACGA... can be read in three frames: ...AGT,CCG,TAC,GA..., ...A,GTC,CGT, ACG,A... and ...AG,TCC,GTA,CGA,... Crick et al. (1957) have then proposed that only 20 among 64 codons code for the 20 amino acids. However,

Table 1
Identification of three subsets of 20 trinucleotides in the protein coding genes of both prokaryotes and eukaryotes (Arquès and Michel, 1997): X_0 in frame 0, X_1 in frame 1 and X_2 in frame 2

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

Table 2
Mean frequencies $P(X_g, F_f)$ (%) of X_0, X_1 and X_2 in the three frames $f = 0, 1, 2$ of protein coding genes (5′ and 3′ parts) of primates, rodents, mammals and vertebrates (PRMV) (with rounded averages)

	Frame 0			Frame 1			Frame 2		
	5′ Parts P5_ $PRMV_0$	3′ Parts P3_ $PRMV_0$	Average P_ $PRMV_0$	5′ Parts P5_ $PRMV_1$	3′ Parts P3_ $PRMV_1$	Average P_ $PRMV_1$	5′ Parts P5_ $PRMV_2$	3′ Parts P3_ $PRMV_2$	Average P_ $PRMV_2$
X_0	48.7	48.3	48.5	25.3	25.4	25.5	31.1	31.2	31.0
X_1	28.9	29.0	29.0	43.6	43.6	43.5	22.5	22.7	22.5
X_2	22.4	22.7	22.5	31.1	31.0	31.0	46.4	46.1	46.5

the determination of a set of 20 codons forming a code X without commas depends on a great number of constraints. For example, the four codons with identical nucleotides AAA, CCC, GGG and TTT must be excluded from such a code. Indeed, the concatenation of AAA, for example, with itself does not allow to retrieve the frame: ...AAA,AAA,AAA,..., ...A,AAA,AAA,AA... and ...AA,AAA,AAA,A... Similarly, two codons related to circular permutations, e.g. AAC and ACA (or CAA), cannot belong at the same time to such a code. Indeed, the concatenation of AAC, for example, with itself leads to the concatenation of ACA (or CAA) with itself in another frame, making the frame determination impossible. Therefore, by excluding AAA, CCC, GGG and TTT and by gathering the 60 remaining codons in 20 classes of three codons so that, in each class, the three codons are deduced from each other by circular permutations, e.g. AAC, ACA and CAA, a code X 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).

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

1.2.2. From a computer point of view

Let B be a genetic alphabet, e.g. $B_4 = \{A,C,G,T\}$. B* denotes the words on B of finite length including the empty word of length 0. B^+ denotes the words on B of finite length ≥ 1. Let w_1w_2 be the concatenation of the two words w_1 and w_2.

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

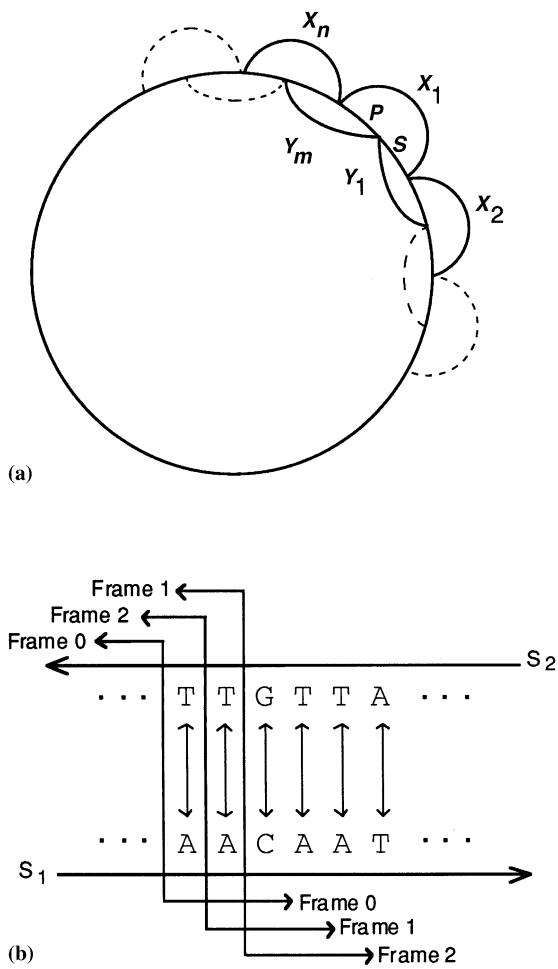


Fig. 1. (a) A representation of the definition of a circular code. (b) The complementarity property of the C^3 code X_0 implies several symmetries with the occurrence frequencies of X_0, X_1 and X_2 in the three frames, in particular the same frequency of X_1 and X_2 in frame 0.

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

On the alphabet $B_2 = \{R, Y\}$, the circular codes can be completely studied by hand (Arquès and Michel, 1997). On the alphabet $B_4 = \{A, C, G, T\}$, the identification of a circular code is obviously more complex and difficult as there are ≈ 3.5 milliard potential maximal circular codes of 20 trinucleotides (Arquès and Michel, 1997, Table 3d).

1.3. Aims of this study

1.3.1. A quantitative analysis of the circular code

The first aim in Section 2 consists in quantifying the three circular codes X_0, X_1 and X_2 of 20 trinucleotides in the three frames of protein genes of eukaryotes. Therefore, the occurrence frequencies of X_0, X_1 and X_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).

This analysis will show that the global mean frequency of X_0 in frame 0 of eukaryotic protein genes is greater than the global mean frequencies of X_1 or X_2 in frame 0, as expected by the preferential occurrence of X_0 in frame 0 according to the Table 1. Similarly, the global mean frequency of X_1 (resp. X_2) in frame 1 (resp. 2) of eukaryotic protein genes will be demonstrated to be greater than the global mean frequencies of X_0 or X_2 (resp. X_0 or X_1) in frame 1 (resp. 2), as expected by the preferential occurrence of X_1 (resp. X_2) in frame 1 (resp. 2) shown in Table 1.

The complementarity property of the C^3 code X_0 would imply several symmetries with the frequencies of X_0, X_1 and X_2 in the three frames, in particular the same frequency of X_1 and X_2 in frame 0 (Fig. 1b). Unexpectedly, this analysis will not observed this symmetry in eukaryotic protein genes. Several frequency asymmetries in contradiction with the complementarity property will be identified.

1.3.2. An evolutionary model of the circular code

The second aim in Section 3 is the development of an evolutionary model in order to propose an explanation for the existence of the three circular codes X_0, X_1 and X_2 in the three frames of actual eukaryotic protein genes and their asymmetries.

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), i.e. the three sets X_0, X_1 and X_2 of trinucleotides occur in frame 0 (also in frames 1 and 2). However, these three sets do not occur with the same frequency. Indeed, the results obtained in Section 2

2. the evolutionary process based on random substitutions in the three codon sites will transform the simulated primitive genes into simulated actual genes. Substitutions with different rates in the codon sites of the circular code X_0 allow the generation of the circular codes X_1 and X_2 according to a non-balanced way. The simulated actual genes obtained by this process have the three circular codes and their frequency asymmetries, as observed in the three frames of real actual protein genes. Furthermore, the measure of these asymmetries in the model allows the determination of a number of substitutions in the evolutionary process. According to the degeneracy of the genetic code, the highest substitution rate is expected to occur in the third codon site, which will be observed in the model. Indeed, after ≈ 4 substitutions per codon in the three codon sites in proportions ≈ 0.1, ≈ 0.1, and ≈ 0.8 respectively, the simulated actual genes are correlated with the real actual protein genes. This model leads to the identification of a new property, namely an evolutionary property, of the C^3 code X_0 which has to be added to the previous ones (Arquès and Michel, 1997).

2. A quantitative study of the C^3 code X_0 in the protein coding genes of eukaryotes

2.1. Method

Let w be a trinucleotide in {AAA,...,TTT} (64 trinucleotides). Let $f \in \{0, 1, 2\}$ be a frame determined by a series of trinucleotides in a protein (coding) 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. By choosing the stop trinucleotide TAA as an example, $f = 0$ is the following frame ATG,NNN,...,NNN,TAA and $f = 1$, A,TGN,...,NNT,AA and $f = 2$, AT, GNN,...,NTA,A (N being any nucleotide). Therefore, the population F containing the protein 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 after ATG (5′ parts of protein genes, $d > 0$) or before a stop trinucleotide (3′ parts of protein genes, $d < 0$). 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 10th codon after ATG (resp. TGN, GNN) and $d = -10$ in $f = 0$ (resp. $f = 1$, $f = 2$) is the 10th codon 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 X_g be the subset of 20 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

$$\delta_d(X_g) = \begin{cases} 1 \text{ if } w_d \in X_g \\ 0 \text{ if } w_d \notin X_g \end{cases}$$

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

$$P_d(X_g, F_f) = \sum_{s \in F_f} \delta_d(X_g) \Big/ \sum_{g=0,1,2} \sum_{s \in F_f} \delta_d(X_g) \qquad (1a)$$

This probability function is represented as a curve as follows: the abscissa shows the position d in trinucleotides, by varying d in the ranges [2, 200] (5′ parts of protein genes) and $[-200, -2]$ (3′ parts of protein genes), and the ordinate gives the occurrence probability of $P_d(X_0, F_f)$, $P_d(X_1, F_f)$ and $P_d(X_2, F_f)$ in a protein gene population F_f.

Remarks:

1. for readibility reasons, the ATG, the stop and the first ($d = 1$ and $d = -1$) conserved trinucleotides are not represented in the curves;
2. the four trinucleotides AAA, CCC, GGG and TTT are excluded from the statistical analysis (see the denominator of Eq. (1a)) as X_0, X_1 and X_2 are circular codes and thereby, containing no trinucleotide with identical nucleotides. As X_0, X_1 and X_2 have the same

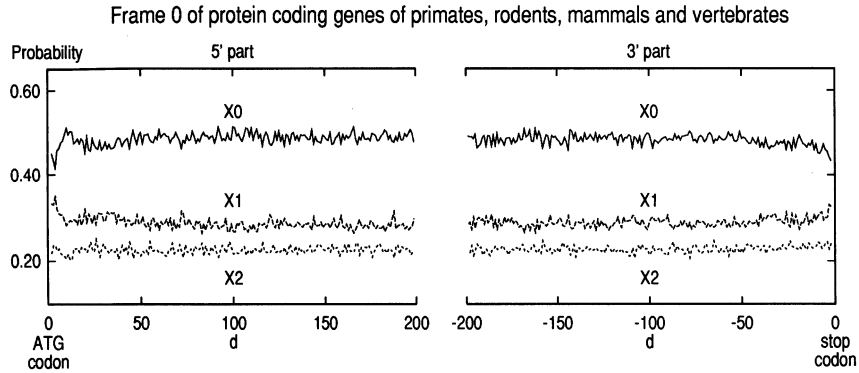


Fig. 2. Probability $P_d(X_g, F_0)$ of X_0, X_1 and X_2 at the trinucleotide position d in frame 0 of protein coding genes (5′ parts $F_0 = P5\_PRMV_0$ and 3′ parts $F_0 = P3\_PRMV_0$) of primates, rodents, mammals and vertebrates ($PRMV_0$). Three distinct horizontal curves X_0, X_1, X_2 in decreasing probabilities occur in the protein coding genes.

number of trinucleotides (20), their occurrence probabilities can be directly compared;

3. $P_d(X_f, F_f) > P_d(X_g, F_f)$, $f, g \in \{0, 1, 2\}$ 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 X_0, X_1 and X_2 have a preferential occurrence in the frames 0, 1 and 2 respectively (Table 1);
4. $P_d(X_g, R_f) = 1/3$, $f, g \in \{0, 1, 2\}$, for any position d in the three frames f of a random gene population R_f generated with an independent mixing of the four nucleotides A, C, G and T with equiprobability (1/4).

The large protein gene populations F of higher eukaryotes analysed here are the protein (coding) genes of:

1. primates, rodents, mammals and vertebrates: 17072 sequences PRMV with the 5′ parts $F =$ P5_PRMV and 3′ parts $F =$ P3_PRMV;
2. primates: 6455 sequences PRI with the 5′ parts $F =$ P5_PRI and the 3′ parts $F =$ P3_PRI;
3. rodents: 6409 sequences ROD with the 5′ parts $F =$ P5_ROD and the 3′ parts $F =$ P3_ROD;
4. (other) mammals: 1991 sequences MAM with the 5′ parts $F =$ P5_MAM and the 3′ parts $F =$ P3_MAM;
5. (other) vertebrates: 2217 sequences VRT with the 5′ parts $F =$ P5_VRT and the 3′ parts $F =$ P3_VRT.

These large populations, obtained from the release 45 (December 1995) of the EMBL nucleotide sequence data library in the same way as described in previous studies (see e.g. Arquès and Michel, 1987, 1990 for a description of data acquisitions), allow to have stable frequencies (consequence of the law of large numbers, Arquès and Michel, 1990, Section 2.3.3).

2.2. Results

Fig. 2 shows that the probability curve X_0 is, as expected, greater than the curves X_1 or X_2, for any trinucleotide position d in frame 0 of 5′ parts ($P_d(X_g, P5\_PRMV_0)$) and 3′ parts ($P_d(X_g, P3\_PRMV_0)$) of protein genes of primates, rodents, mammals and vertebrates. The curve X_0 is globally horizontal with an average frequency around 48.5% in P_PRMV_0 ($P5\_PRMV_0$ and $P3\_PRMV_0$) (Table 2). The two curves X_1 and X_2 are also globally horizontal but unexpectedly, distinct (Fig. 2). Indeed, the average frequency around 29% of X_1 is greater than the frequency around 22.5% of X_2 in P_PRMV_0 (Table 2). The probability difference $P_d(X_1, P\_PRMV_0) - P_d(X_2, P\_PRMV_0) \approx 0.065$ in frame 0 cannot be explained by the difference $1/60 \approx 0.017$ consequent on the fact that X_1 has one stop trinucleotide less than X_2 ($TAG \in X_1$ and $TAA, TGA \in X_2$, Table 1). This difference is a first contradiction with the expected equality resulting from the complementarity property of the C^3 code X_0 (self-complementarity of X_0 and complementarity of X_1 and X_2). The probabilities in frame 0 can be

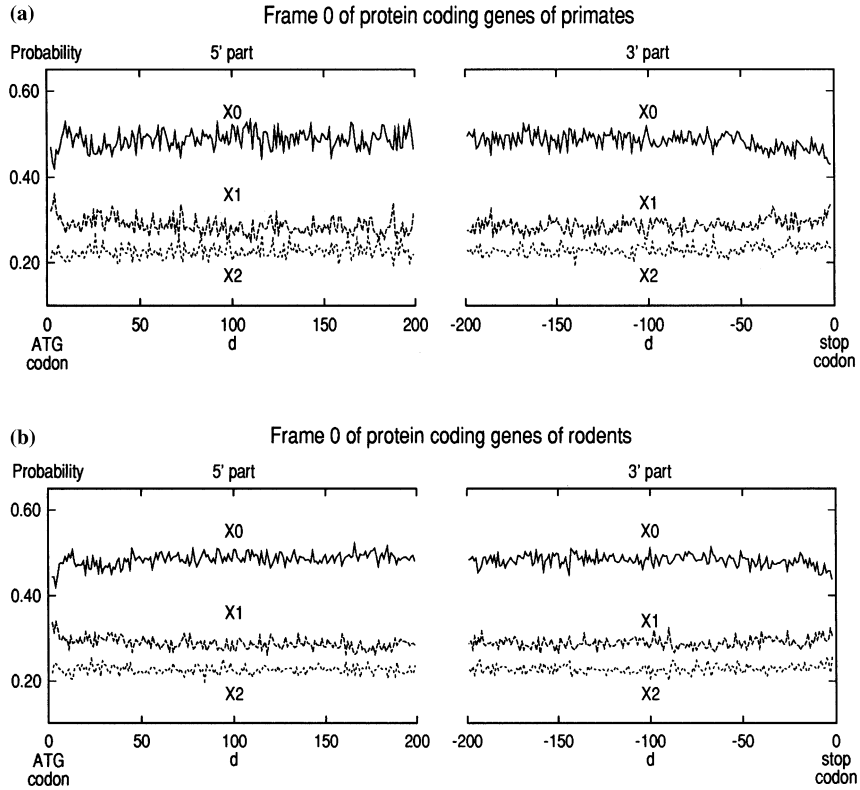


Fig. 3. (a) probability $P_d(X_g, F_0)$ of X_0, X_1 and X_2 at the trinucleotide position d in frame 0 of protein coding genes (5′ parts $F_0 = P5\_PRI_0$ and 3′ parts $F_0 = P3\_PRI_0$) of primates (PRI_0). Three distinct horizontal curves X_0, X_1, X_2 in decreasing probabilities occur in the protein coding genes of primates; (b) probability $P_d(X_g, F_0)$ of X_0, X_1 and X_2 at the trinucleotide position d in frame 0 of protein coding genes (5′ parts $F_0 = P5\_ROD_0$ and 3′ parts $F_0 = P3\_ROD_0$) of rodents (ROD_0). Three distinct horizontal curves X_0, X_1, X_2 in decreasing probabilities occur in the protein coding genes of rodents; (c) probability $P_d(X_g, F_0)$ of X_0, X_1 and X_2 at the trinucleotide position d in frame 0 of protein coding genes (5′ parts $F_0 = P5\_MAM_0$ and 3′ parts $F_0 = P3\_MAM_0$) of mammals (MAM_0). Three distinct horizontal curves X_0, X_1, X_2 in decreasing probabilities occur in the protein coding genes of mammals; (d) probability $P_d(X_g, F_0)$ of X_0, X_1 and X_2 at the trinucleotide position d in frame 0 of protein coding genes (5′ parts $F_0 = P5\_VRT_0$ and 3′ parts $F_0 = P3\_VRT_0$) of vertebrates (VRT_0). Three distinct horizontal curves X_0, X_1, X_2 in decreasing probabilities occur in the protein coding genes of vertebrates.

represented by the following set Q_0 of inequalities: $P_d(X_0, P\_PRMV_0) > P_d(X_1, P\_PRMV_0) > P_d(X_2, P\_PRMV_0)$.

The horizontality 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 frame 0 of protein genes of primates, rodents, mammals and vertebrates (data not shown).

These results in frame 0 are also observed with each protein gene subpopulations: primates (Fig. 3a: $P_d(X_g, P5\_PRI_0)$ and $P_d(X_g, P3\_PRI_0)$), rodents (Fig. 3b: $P_d(X_g, P5\_ROD_0)$ and $P_d(X_g, P3\_ROD_0)$), mammals (Fig. 3c: $P_d(X_g, P5\_MAM_0)$ and $P_d(X_g, P3\_MAM_0)$) and vertebrates (Fig. 3d: $P_d(X_g, P5\_VRT_0)$ and $P_d(X_g, P3\_VRT_0)$).

Fig. 4 (resp. 5) shows that the probability curve X_1 (resp. X_2) is, as expected, greater than the curves X_2 or X_0 (resp. X_0 or X_1), for any trinucleotide position d in frame 1 (resp. 2) of 5′ and 3′ parts of protein genes of primates, rodents, mammals and vertebrates (Fig. 4: $P_d(X_g, P5\_PRMV_1)$ and $P_d(X_g, P3\_PRMV_1)$) (resp. Fig. 5: $P_d(X_g, P5\_PRMV_2)$ and $P_d(X_g, P3\_PRMV_2)$). In frame 1, the three curves X_1, X_2 and X_0 are globally horizontal with an average frequency

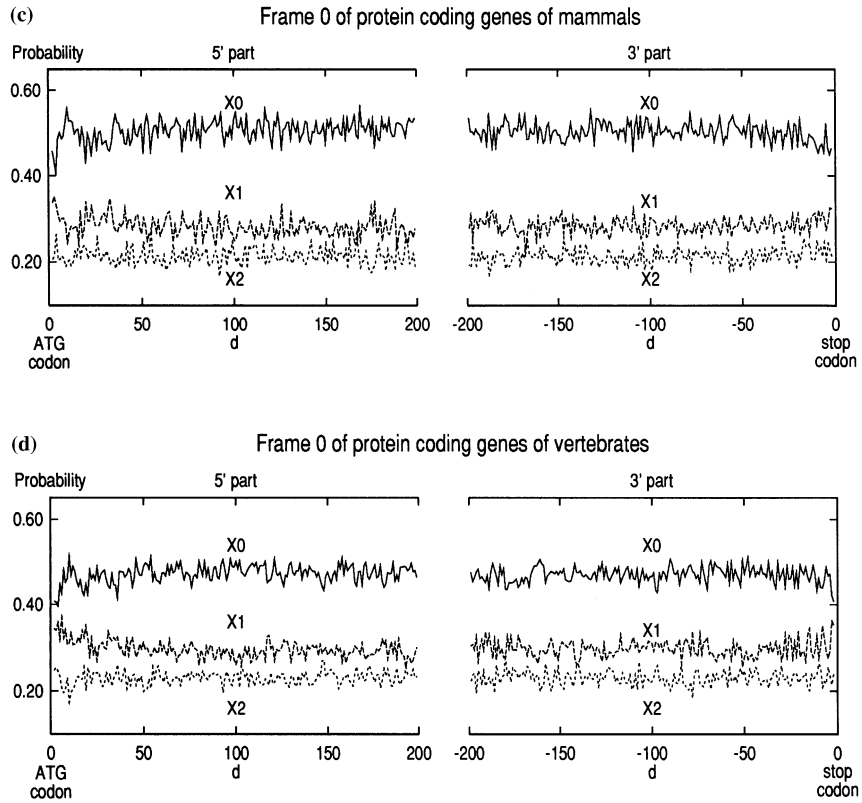


Fig. 3. (*Continued*)

around 43.5%, 31% and 25.5% respectively (Table 2). The probabilities in frame 1 can be represented by the following set Q_1 of inequalities: $P_d(X_1, P\_PRMV_1) > P_d(X_2, P\_PRMV_1) > P_d(X_0, P\_PRMV_1)$. In frame 2, the three curves X_2, X_0 and X_1 are globally horizontal with an average frequency around 46.5%, 31% and 22.5% respectively (Table 2). The probabilities in frame 2 can be represented

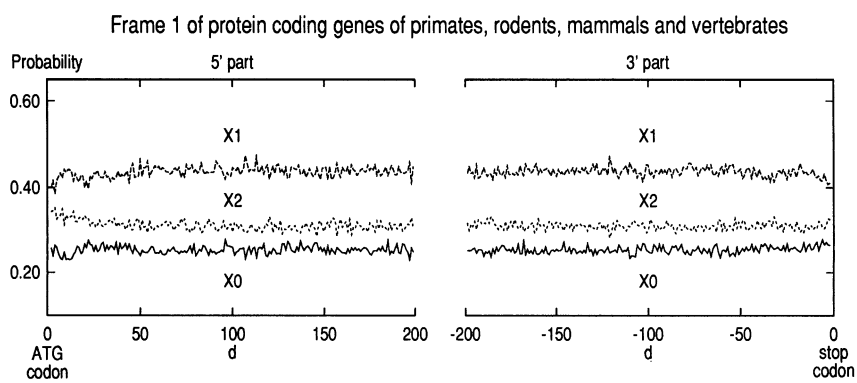


Fig. 4. Probability $P_d(X_g, F_1)$ of X_0, X_1 and X_2 at the trinucleotide position d in frame 1 of protein coding genes (5′ parts $F_1 = P5\_PRMV_1$ and 3′ parts $F_1 = P3\_PRMV_1$) of primates, rodents, mammals and vertebrates ($PRMV_1$). Three distinct horizontal curves X_1, X_2, X_0 in decreasing probabilities occur in the protein coding genes.

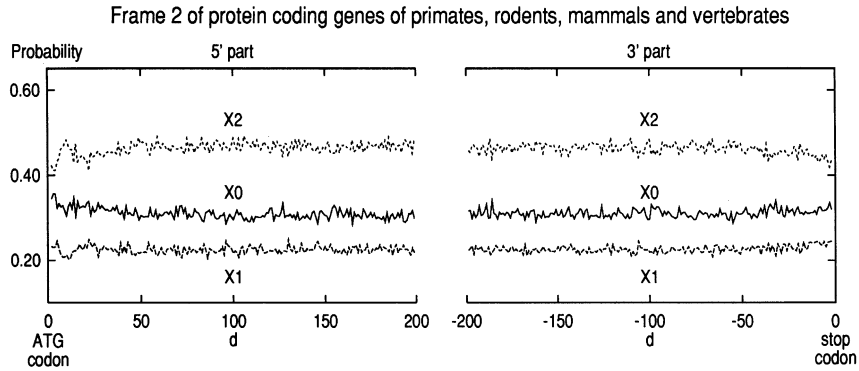


Fig. 5. Probability $P_d(X_g, F_2)$ of X_0, X_1 and X_2 at the trinucleotide position d in frame 2 of protein coding genes (5′ parts $F_2 = P5\_PRMV_2$ and 3′ parts $F_2 = P3\_PRMV_2$) of primates, rodents, mammals and vertebrates ($PRMV_2$). Three distinct horizontal curves X_2, X_0, X_1 in decreasing probabilities occur in the protein coding genes.

by the following set Q_2 of inequalities: $P_d(X_2, P\_PRMV_2) > P_d(X_0, P\_PRMV_2) > P_d(X_1, P\_PRMV_2)$. These inequalities in frame 2 constitute a second contradiction with the complementarity property of the C^3 code X_0 which would have led to: $P_d(X_2, P\_PRMV_2) > P_d(X_1, P\_PRMV_2) > P_d(X_0, P\_PRMV_2)$. Indeed, a simulated population S generated, for example, according to an independent mixing of the 20 trinucleotides of X_0 with equiprobability (1/20) leads to the following inequalities, in frame 0: $P(X_0, S_0) > P(X_1, S_0) = P(X_2, S_0)$, in frame 1: $P(X_1, S_1) > P(X_2, S_1) > P(X_0, S_1)$ and in frame 2: $P(X_2, S_2) > P(X_1, S_2) > P(X_0, S_2)$.

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

3. An evolutionary analytical model of the C^3 code X_0

3.1. Presentation of the model

The three subsets X_0, X_1 and X_2 of trinucleotides in the three frames of eukaryotic protein genes ($F = P$) present several statistical properties. The nine probability curves are constant functions of the trinucleotide position (horizontal curves) in the three frames of eukaryotic protein genes. Therefore, these curves can be characterized by their probabilities $P(X_g, P_f)$ (instead of $P_d(X_g, P_f)$) (Table 2). These probabilities are highly statistically significant as they are computed in a large population (PRMV) and retrieved in its subpopulations (PRI, ROD, MAM, VRT). These probabilities $P(X_g, F_f)$ of X_0, X_1 and X_2 in the three frames $f = 0,1,2$ of protein genes P are non-random (values different from 1/3) and can be represented by three sets of inequalities, Q_0 in frame 0: $P(X_0, P_0) > P(X_1, P_0) > P(X_2, P_0)$, Q_1 in frame 1: $P(X_1, P_1) > P(X_2, P_1) > P(X_0, P_1)$ and Q_2 in frame 2: $P(X_2, P_2) > P(X_0, P_2) > P(X_1, P_2)$ (Table 2). As detailed in Section 2.2, these probability inequalities seem to be in contradiction with the complementarity property of the C^3 code X_0.

A new property of the C^3 code X_0 related to evolution by random substitutions is studied in this section. Precisely, the problem investigated here is whether an evolutionary analytical model can explain the properties of the C^3 code X_0 observed in actual protein genes and in particular their asymmetries. The main evolutionary process of protein genes is determined by random substitutions of nucleotides. RNA editing by insertions and deletions of nucleotides, is an evolutionary process only observed in particular protein genes as it destroys the reading frame (Benne et al., 1986; Benne, 1989). The different remarks presented in Section 1.3.2 and in particular the two

facts that the actual protein genes have a preferential occurrence of the subset X_0 (in frame 0; note also that $P(X_0, P_0) > P(X_2, P_2) > P(X_1, P_1)$ in Table 2) and that the main process of gene evolution is determined by nucleotide substitutions, lead to investigate a model based on two processes:

1. a construction process based on an independent mixing of the 20 trinucleotides of the circular code X_0 with equiprobability (1/20);
2. an evolutionary process based on different random substitutions in the three trinucleotide sites.

A solution of this model is obtained when the three sets Q_0, Q_1 and Q_2 of inequalities are verified and when X_0, X_1 and X_2 have frequencies in their three frames similar to those observed in the actual protein genes (given in Table 2).

Such models based on two successive processes, construction and evolution (random substitutions, random 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 X_0, X_1 and X_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 $[t, t + dt]$:

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

where α is the probability that a trinucleotide is subjected to one substitution during an 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 different from i and, j and i differ more than one nucleotide as dt is assumed to be enough small that a codon cannot substitute successively two times during $[t, t + dt]$) otherwise it is given in 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 $i = 1$, $P(\mathrm{AAC} \to \mathrm{AAA}) = P(\mathrm{AAG} \to \mathrm{AAA}) = P(\mathrm{AAT} \to \mathrm{AAA}) = r/3$, $P(\mathrm{ACA} \to \mathrm{AAA}) = P(\mathrm{AGA} \to \mathrm{AAA}) = P(\mathrm{ATA} \to \mathrm{AAA}) = q/3$, $P(\mathrm{CAA} \to \mathrm{AAA}) = P(\mathrm{GAA} \to \mathrm{AAA}) = P(\mathrm{TAA} \to \mathrm{AAA}) = p/3$ and $P(j \to \mathrm{AAA}) = 0$ with

Table 3
Substitution matrix **A** (64, 64) of the 4096 trinucleotide substitutions given in function of the three substitution rates p, q and $r = 1 - p - q$ in the three codon sites respectively

	1:AAA…16:ATT	17:CAA…32:CTT	33:GAA…48:GTT	49:TAA…64:TTT
1:AAA…16:ATT	**M**	(p/3)**Id**	(p/3)**Id**	(p/3)**Id**
17:CAA…32:CTT	(p/3)**Id**	**M**	(p/3)**Id**	(p/3)**Id**
33:GAA…48:GTT	(p/3)**Id**	(p/3)**Id**	**M**	(p/3)**Id**
49:TAA…64:TTT	(p/3)**Id**	(p/3)**Id**	(p/3)**Id**	**M**

The lines and the columns correspond to the trinucleotides given in the alphabetical order. The matrix **A** (64, 64) is symmetrical, i.e. the lines or the columns can correspond to the trinucleotides before substitutions 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 the 12 non-diagonal elements are formed by 12 identical square submatrices (p/3)**Id** (16, 16).

$j \notin$ {AAC,AAG,AAT,ACA,AGA,ATA,CAA,GAA,TAA}. The 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.

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 the formula (1) becomes

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

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 Eq. (2) can be represented by the following matrix equation:

$$P'(t) = -P(t) + \mathbf{A}.P(t) = (\mathbf{A} - \mathbf{Id}).P(t)$$

where the symbol . represents the matrix product.

The real matrix **A** is symmetrical (see Table 3 and Appendix). Therefore, the real matrix $\mathbf{A} - \mathbf{Id}$ is symmetrical. Then, it exists 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 $\mathbf{A} - \mathbf{Id} = \mathbf{Q}.\mathbf{D}.\mathbf{Q}^{-1}$. Then,

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

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

This equation has the classical solution

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

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

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

The characteristic roots $(\lambda_k - 1)$ of the matrix $\mathbf{A} - \mathbf{Id}$ are deduced from the characteristic roots λ_k of the matrix **A** which can be obtained by determining the roots of the characteristic equation $\det(\mathbf{A} - \lambda\mathbf{Id}) = 0$ of **A**. The determinant for the matrix $\mathbf{A} - \lambda\mathbf{Id}$ is equal to (proof given in appendix):

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

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 20 trinucleotides of X_0 with equiprobability (1/20) leads to the following initial vector $P(0) = [0, 1/20, 0, 1/20, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 1/20, 0, 1/20, 0, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/20, 1/20, 0, 1/20, 1/20, 1/20, 1/20, 0, 1/20, 0, 0, 0, 1/20, 0, 1/20, 1/20, 1/20, 0, 1/20, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/20, 0, 0]$. The formula $P(t) = \mathbf{Q}.e^{\mathbf{D}t}.\mathbf{Q}^{-1}.P(0)$ 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 $\mathbf{Q}$ and its inverse $\mathbf{Q}^{-1}$, allows to deduce the 64 trinucleotide probabilities $P_i(t)$ in frame 0 after t substitutions in function of the three substitution rates p, q and $r = 1 - p - q$ in the three codon sites respectively (the indexes i or j representing the trinucleotides AAA,...,TTT in the alphabetical order). For example with the trinucleotide AAC ($i = 2$),

$$P_2(t) = \frac{1}{320}\,[2e^{-\frac{4}{3}t} - e^{-\frac{4}{3}(1-p)t} + 3e^{-\frac{4}{3}qt} + 2e^{-\frac{4}{3}(1-q)t} + 5e^{-\frac{4}{3}(1-p-q)t} + 5]$$

Note: $\lim_{t \to \infty} P_2(t) = 5/320 = 1/64$

Therefore, the occurrence probabilities $P(X_g, f = 0, t)$ of the subsets X_0, X_1 and X_2 of trinucleotides in the frame $f = 0$ at the substitution step t are equal to

$$P(X_g, 0, t) = \sum_{i \in X_g} P_i(t) \Big/ \sum_{i \in X_0 \cup X_1 \cup X_2} P_i(t)$$

After simplification,

$$P(X_0, 0, t) = \frac{1}{2 \times D_0}\,[607e^{-\frac{4}{3}t} + 408e^{-\frac{4}{3}pt} + 111e^{-\frac{4}{3}(1-p)t} + 402e^{-\frac{4}{3}qt} + 349e^{-\frac{4}{3}(1-q)t} + 108e^{-\frac{4}{3}(p+q)t} + 413e^{-\frac{4}{3}(1-p-q)t} + 1090]$$

$$P(X_1, 0, t) = \frac{1}{2 \times D_0}\,[-261e^{-\frac{4}{3}t} - 192e^{-\frac{4}{3}pt} - 117e^{-\frac{4}{3}(1-p)t} - 194e^{-\frac{4}{3}qt} - 111e^{-\frac{4}{3}(1-q)t} - 215e^{-\frac{4}{3}(1-p-q)t} + 1090]$$

$$P(X_2, 0, t) = \frac{1}{D_0}\,[-131e^{-\frac{4}{3}t} - 108e^{-\frac{4}{3}pt} + 2e^{-\frac{4}{3}(1-p)t} - 104e^{-\frac{4}{3}qt} - 54e^{-\frac{4}{3}(1-q)t} - 51e^{-\frac{4}{3}(p+q)t} - 99e^{-\frac{4}{3}(1-p-q)t} + 545]$$

with $D_0 = 42e^{-\frac{4}{3}t} - e^{-\frac{4}{3}(1-p)t} + 65e^{-\frac{4}{3}(1-q)t} + 3e^{-\frac{4}{3}(p+q)t} + 1635$

Notes:

$$\sum_{g=0,1,2} P(X_g, 0, t) = 1$$

$$\lim_{\substack{t \to \infty \\ g=0,1,2}} P(X_g, 0, t) = 1090/(2 \times 1635) = 545/1635 = 1/3$$

(remember that the three subsets X_0, X_1 and X_2 have the same number of trinucleotides).

The occurrence probabilities $P(X_g, f = 1, t)$ (resp. $P(X_g, f = 2, t)$) of X_0, X_1 and X_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 the 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 AAC ($i = 2$) in the frame $f = 1$ is obtained by the concatenation of the two types of trinucleotides NAA and CNN in frame 0. Therefore, the probability $P(i = 2, 1, t)$ of the trinucleotide AAC ($i = 2$) 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 CNN ($k = 17,...,32$) in frame 0. Similarly, the probability $P(i = 2, 2, t)$ of the trinucleotide AAC ($i = 2$) in the frame $f = 2$ is equal to the product of the probability of the trinucleotides NNA ($j = 1 + 4 \times j'$ with $j' = 0,...,15$) in frame 0 and the probability of the trinucleotides ACN ($k = 5,...,8$) in frame 0:

$$P(2, 1, t) = \sum_{j = 1,17,33,49} P(j, 0, t) \times \sum_{k = 17,...,32} P(k, 0, t)$$

$$P(2, 2, t) = \sum_{\substack{j = 1 + 4 \times j' \\ j' = 0,...,15}} P(j, 0, t) \times \sum_{k = 5, ...,8} P(k, 0, t)$$

After simplification,

$$P(X_0, 1, t) = \frac{1}{D_1}\,[2e^{-\frac{4}{3}t} - 50e^{-\frac{4}{3}pt} - 45e^{-\frac{4}{3}qt} - 2e^{-\frac{4}{3}(1-q)t} - 18e^{-\frac{4}{3}(p+q)t} - 45^{-\frac{4}{3}(1-p-q)t} + 250]$$

$$P(X_1, 1, t) = \frac{1}{D_1}\,[-e^{-\frac{4}{3}t} + 95e^{-\frac{4}{3}pt} + 25e^{-\frac{4}{3}(1-p)t} + 90e^{-\frac{4}{3}qt} + 14e^{-\frac{4}{3}(1-q)t} + 12e^{-\frac{4}{3}(p+q)t} + 95e^{-\frac{4}{3}(1-p-q)t} + 250]$$

$$P(X_2, 1, t) = \frac{1}{D_1}\,[-45e^{-\frac{4}{3}pt} - 25e^{-\frac{4}{3}(1-p)t} - 45e^{-\frac{4}{3}qt} - 2e^{-\frac{4}{3}(1-q)t} + 15e^{-\frac{4}{3}(p+q)t} - 50e^{-\frac{4}{3}(1-p-q)t} + 250]$$

with $D_1 = e^{-\frac{4}{3}t} + 10e^{-\frac{4}{3}(1-q)t} + 9e^{-\frac{4}{3}(p+q)t} + 750$

$$P(X_0, 2, t) = \frac{1}{D_2}\,[2e^{-\frac{4}{3}t} - 45e^{-\frac{4}{3}pt} - 18^{-\frac{4}{3}(1-p)t} - 45e^{-\frac{4}{3}qt} - 2e^{-\frac{4}{3}(1-q)t} - 50e^{-\frac{4}{3}(1-p+q)t} + 250]$$

$$P(X_1, 2, t) = \frac{1}{D_2}\,[-50e^{-\frac{4}{3}pt} + 15e^{-\frac{4}{3}(1-p)t} - 45e^{-\frac{4}{3}qt} - 2e^{-\frac{4}{3}(1-q)t} - 25e^{-\frac{4}{3}(p+q)t} - 45e^{-\frac{4}{3}(1-p-q)t} + 250]$$

$$P(X_2, 2, t) = \frac{1}{D_2}\,[-e^{-\frac{4}{3}t} + 95e^{-\frac{4}{3}pt} + 12e^{-\frac{4}{3}(1-p)t} + 90e^{-\frac{4}{3}qt} + 14e^{-\frac{4}{3}(1-q)t} + 25e^{-\frac{4}{3}(p+q)t} + 95e^{-\frac{4}{3}(1-p-q)t} + 250]$$

with $D_2 = e^{-\frac{4}{3}t} + 9e^{-\frac{4}{3}(1-p)t} + 10e^{-\frac{4}{3}(1-q)t} + 750$

Notes:

$$\sum_{g=0,1,2} P(X_g,1,t) = \sum_{g=0,1,2} P(X_g,2,t) = 1$$

$$\lim_{\substack{t\to 0\\ g=0,1,2\\ f=1,2}} P(X_g,f,t) = 250/750 = 1/3$$

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 there are values for the three parameters p, q and t verifying both the three inequality sets Q_0 in frame 0: $P(X_0, P_0) > P(X_1, P_0) > P(X_2, P_0)$, Q_1 in frame 1: $P(X_1, P_1) > P(X_2, P_1) > P(X_0, P_1)$ and Q_2 in frame 2: $P(X_2, P_2) > P(X_0, P_2) > P(X_1, P_2)$ (Section 3.1) and the frequency orders of X_0, X_1 and X_2 associated with the three frames of actual protein genes (Table 2).

3.3. Results

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, 20] with a step of 0.1, the model (p, q, t) retrieves the three inequality sets Q_0, Q_1 and Q_2 of actual protein genes when $p = 0.1 \pm 0.05$, $q = 0.1 \pm 0.05$ ($r = 1 - p - q = 0.8 \pm 0.1$) and $t \geq 0.3$ (Fig. 6a–c). Furthermore, $P(X_0, 0, t) = 0.485$ at $t \approx 4.3$ (Fig. 6a).

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: $P(X_0, 0, 0) = 1$ and $P(X_1, 0, 0) = P(X_2, 0, 0) = 0$ (Fig. 6a), in frame 1: $P(X_1, 1, 0) = 0.754$, $P(X_2, 1, 0) = 0.127$ and $P(X_0, 1, 0) = 0.119$ (Fig. 6b) and in frame 2: $P(X_2, 2, 0) = 0.754$, $P(X_1, 2, 0) = 0.127$ and $P(X_0, 2, 0) = 0.119$ (Fig. 6c). The expected probabilities in frame 0 result from the absence of X_1 and X_2 in frame 0. The expected probabilities in the frames 1 and 2 are the consequence of misplaced trinucleotides in these shifted frames (detailed in Arquès and Michel, 1997). These probabilities lead to the following inequalities, in frame 0: $P(X_0, 0, 0) > P(X_1, 0, 0) = P(X_2, 0, 0)$, in frame 1: $P(X_1, 1, 0) > P(X_2, 1, 0) > P(X_0, 1, 0)$ and in frame 2: $P(X_2, 2, 0) > P(X_1, 2, 0) > P(X_0, 2, 0)$ which result from the complementarity property of the C^3 code X_0. The construction process only simulates the inequality Q_1 in frame 1 of protein genes.

Unexpectedly, the substitution process ($t > 0$) allows simultaneously the simulation of the three inequalities Q_0, Q_1 and Q_2 in frames 0, 1 and 2 respectively of protein genes. The inequality Q_0 including $P(X_1, P_0) > P(X_2, P_0)$ in frame 0 of protein genes exists in the model ($p = 0.1$, $q = 0.1$, t) for a substitution number $t > 0$ (Fig. 6a).

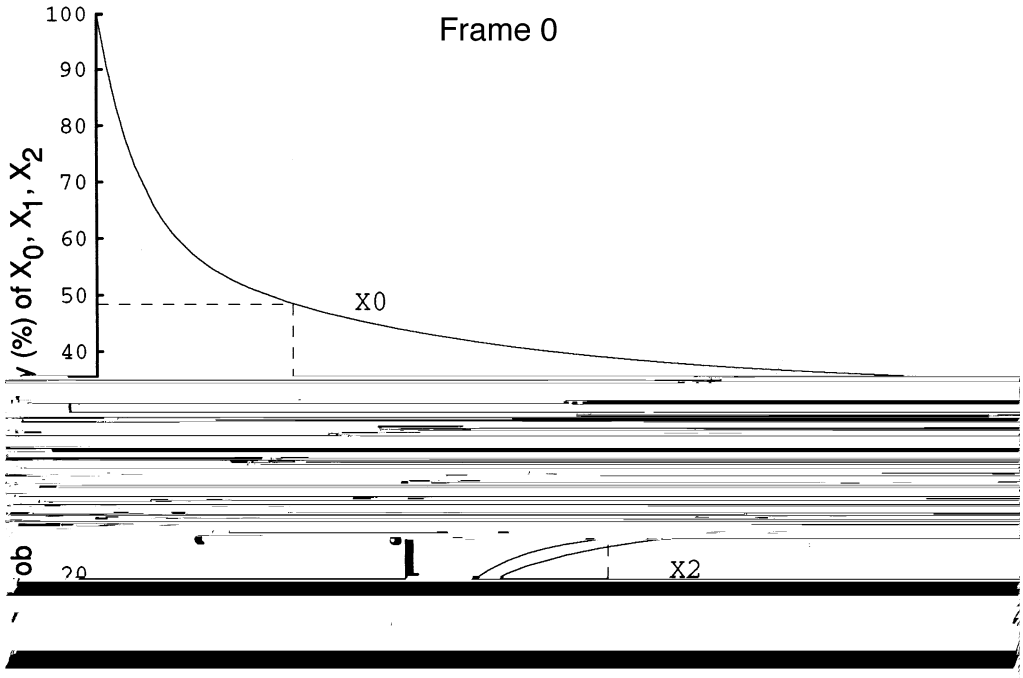


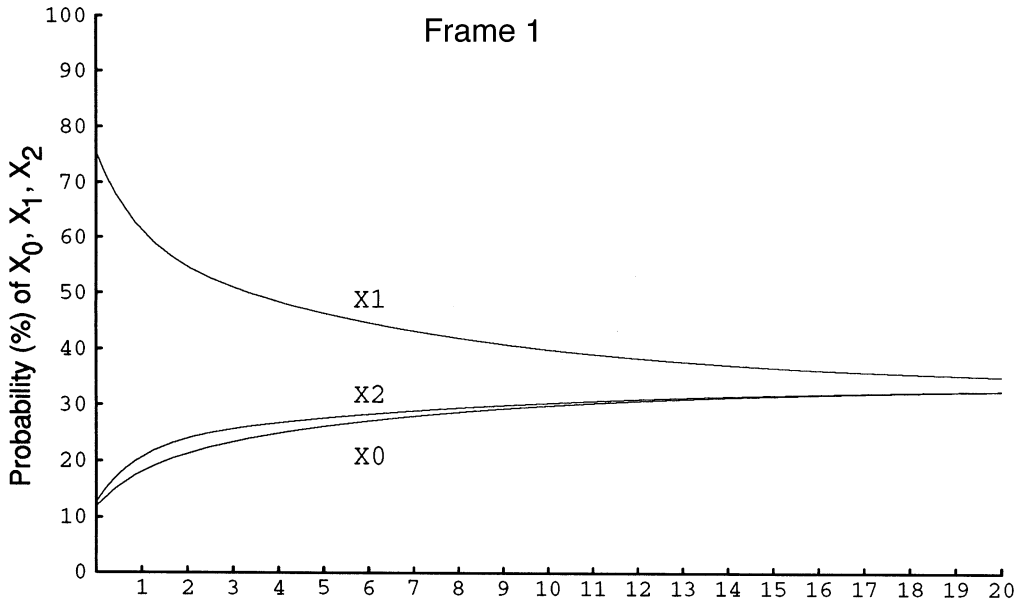
Fig. 6. (a) probability $P(X_g, 0, t)$ of X_0, X_1 and X_2 in frame 0 generated with an independent mixing of the 20 trinucleotides of X_0 with equiprobability (1/20) and subjected to t substitutions per codon according to the proportions $p = 0.1$, $q = 0.1$ and $r = 1 - p - q = 0.8$ in the three codon sites respectively. The inequality Q_0 in frame 0 of protein coding genes is verified for a substitution number $t > 0$. At $t \approx 4.3$ substitutions, the three analytical curves have the occurrence probability orders of X_0, X_1, X_2 similar to those observed in frame 0 of protein coding genes; (b) probability $P(X_g, 1, t)$ of X_0, X_1 and X_2 in frame 1 generated with an independent mixing of the 20 trinucleotides of X_0 with equiprobability (1/20) 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 Q_1 in frame 1 of protein coding genes is verified for a substitution number $t \geq 0$. At $t \approx 4.3$ substitutions, the three analytical curves have the occurrence probability orders of X_1, X_2, X_0 similar to those observed in frame 1 of protein coding genes; (c) probability $P(X_g, 2, t)$ of X_0, X_1 and X_2 in frame 2 generated with an independent mixing of the 20 trinucleotides of X_0 with equiprobability (1/20) 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. At $t \approx 0.2$ substitutions, the two analytical curves X_0 and X_1 cross and retrieve the inequality Q_2 in frame 2 of protein coding genes (Q_2 is not verified for $t < 0.2$). At $t \approx 4.3$ substitutions, the three analytical curves have the occurrence probability orders of X_2, X_0, X_1 similar to those observed in frame 2 of protein coding genes.

The inequality Q_1 including $P(X_2, P_1) > P(X_0, P_1)$ in frame 1 of protein genes exists in the model ($p = 0.1$, $q = 0.1$, t) for a substitution number $t \geq 0$ (Fig. 6b). The inequality Q_2 including $P(X_0, P_2) > P(X_1, P_2)$ in frame 2 of protein genes exists in the model ($p = 0.1$, $q = 0.1$, t) for a substitution number $t \geq 0.2$ (note that $P(X_1, 2, t) > P(X_0, 2, t)$ for $t < 0.2$) (Fig. 6c).

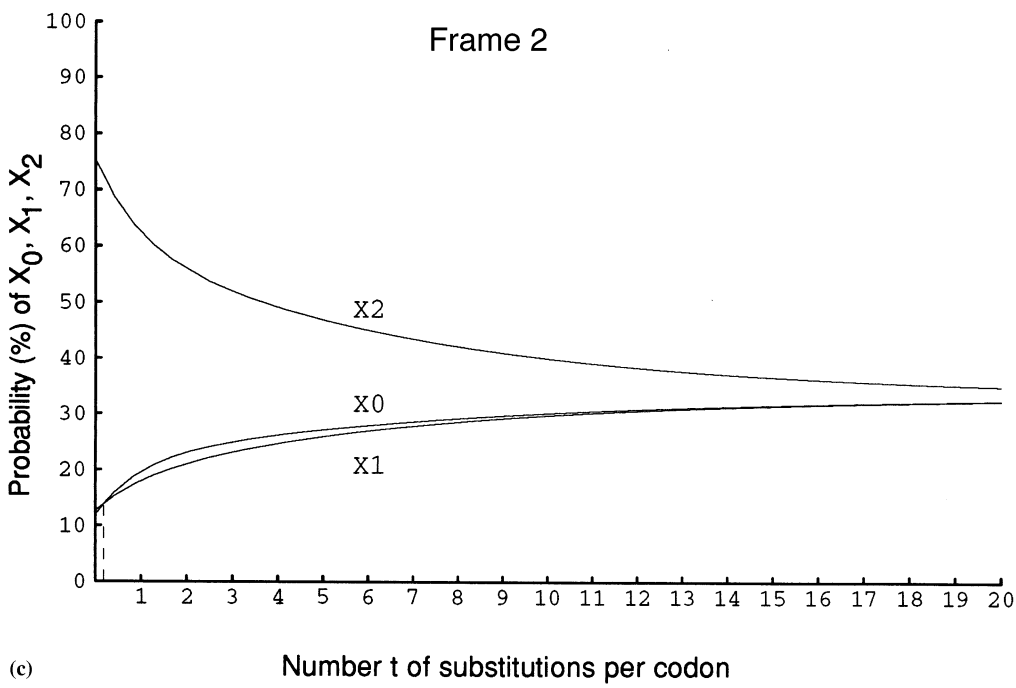
In summary, the substitution process generates the two inequalities $P(X_1, 0, t) > P(X_2, 0, t)$ and $P(X_0, 2, t) > P(X_1, 2, t)$ in frames 0 and 2 respectively and increases the amplitude of the inequality $P(X_2, 1, 0) > P(X_0, 1, 0)$ in frame 1 at $t = 0$. The domain of solution of the model (p, q, t) verifying the inequalities and the frequency orders observed in protein genes, is small: $p = 0.1 \pm 0.05$, $q = 0.1 \pm 0.05$ and $t \approx 4.3$. Outside these ranges of values, the model has no correlation with the genetic reality observed.

4. Discussion

The subset X_0 of trinucleotides (Table 1) has a preferential occurrence in protein genes (frame 0) of prokaryotes and eukaryotes, and the property



(b)



(c)

Fig. 6. (*Continued*)

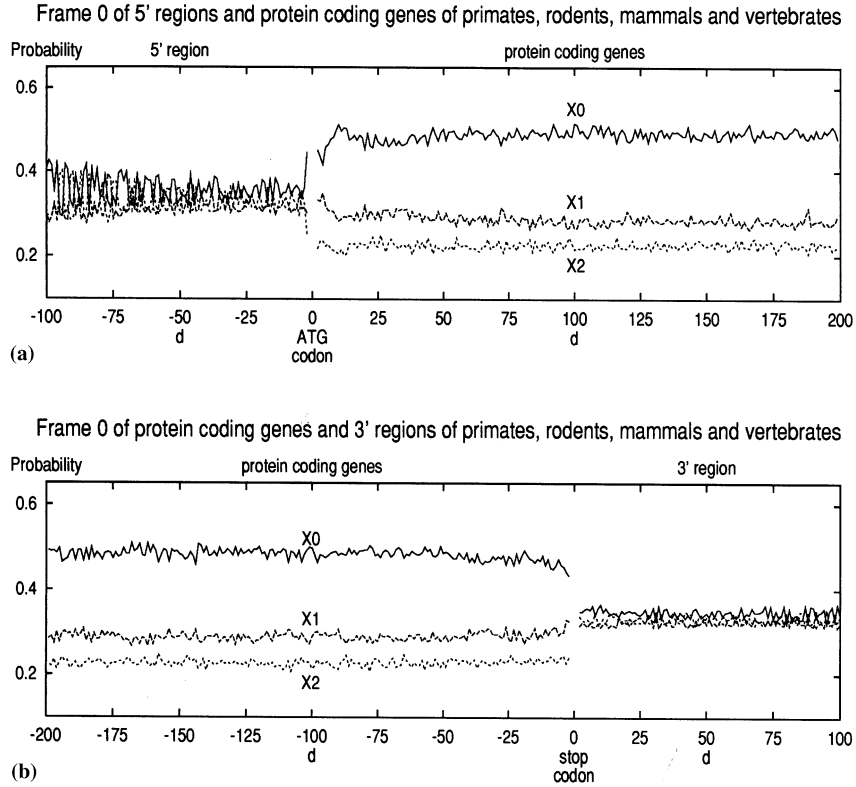


Fig. 7. (a) probability $P_d(X_g, F_0)$ of X_0, X_1 and X_2 at the trinucleotide position d in the extended frame 0 of 5′ regions ($F_0 = R5\_PRMV_0$) and 5′ parts of protein coding genes ($F_0 = P5\_PRMV_0$) of primates, rodents, mammals and vertebrates ($PRMV_0$). Three distinct horizontal curves X_0, X_1, X_2 in decreasing probabilities occur in the protein coding genes while the three curves are mixed in the 5′ regions around the random value 1/3; (b) probability $P_d(X_g, F_0)$ of X_0, X_1 and X_2 at the trinucleotide position d in the extended frame 0 of 3′ parts of protein coding genes ($F_0 = P3\_PRMV_0$) and 3′ regions ($F_0 = R3\_PRMV_0$) of primates, rodents, mammals and vertebrates ($PRMV_0$). Three distinct horizontal curves X_0, X_1, X_2 in decreasing probabilities occur in the protein coding genes while the three curves are mixed in the 3′ regions around the random value 1/3.

to be a complementary maximal circular code with two permutated maximal circular codes X_1 and X_2 (C^3 code, Section 1.1). The quantitative study of the three subsets X_0, X_1 and X_2 in the three frames 0, 1 and 2 of eukaryotic protein genes has showed that their occurrence frequencies are constant for each codon position in the sequences. The frequencies of X_0, X_1 and X_2 in frame 0 of eukaryotic protein genes are 48.5%, 29% and 22.5% respectively. This strong property is not observed in the extended frame 0 of 5′ regions $F = R5\_PRMV$ (6997 sequences, 443351 trinucleotides) and 3′ regions $F = R3\_PRMV$ (14315 sequences, 1098651 trinucleotides) of primates, rodents, mammals and vertebrates. Indeed, the three subsets X_0, X_1 and X_2 occur with variable frequencies around the random value (1/3) (Fig. 7a–b), as expected with the absence of reading frame in the 5′ and 3′ regions. This property leads to an application at the sequence level. Each sequence in a population is classified in X_0, X_1 or X_2 according to its greatest number of codons belonging to X_0, X_1 or X_2. In the protein genes of primates, rodents, mammals and vertebrates, 93% sequences are classified in X_0, 5% sequences in X_1 and 2% sequences in X_2. In contrast, in the 5′ (resp. 3′) regions of primates, rodents, mammals and vertebrates, 42% (resp. 45%) sequences are classified in X_0, 28% (resp. 29%) sequences in X_1 and 30% (resp. 26%) sequences in X_2. These val-

ues in frame 0 and also those which can be obtained in the shifted frames (to improve the significance) could be introduced in some statistical tests in order to discriminate protein coding and non-coding genes and this application 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 protein genes. Its biological meaning would suggest that the protein genes before substitutions ($t = 0$) are constructed by trinucleotides. Only 20 among 64 trinucleotides would have been necessary. The 20 types of trinucleotides as well as the type of their concatenation are determined in the model. Indeed, the 20 trinucleotides are defined by the subset X_0 which is a C^3 code (Section 1.1). The independent concatenation of these 20 trinucleotides with equiprobability is the simplest type of concatenation and therefore, compatible with a primitive stage of gene evolution. A Markov concatenation of trinucleotides (based on a matrix of probabilities) 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. The substitution process allows the generation of new and totally unexpected inequalities, e.g. the two inequalities $P(X_1, 0, t) > P(X_2, 0, t)$ and $P(X_0, 2, t) > P(X_1, 2, t)$ in the frames 0 and 2 respectively. Furthermore, the substitution process decreases the initial ($t = 0$) probabilities $P(X_0, 0, t)$, $P(X_1, 1, t)$ and $P(X_2, 2, t)$ in the frames 0, 1 and 2 respectively.

The evolutionary model ($p = 0.1$, $q = 0.1$, $t = 4.3$) allows the simulation of the protein genes (P) by retrieving not only the three sets Q_0, Q_1 and Q_2 of inequalities and the frequency orders of X_0, X_1 and X_2 in the three frames respectively but also several other sets of probability inequalities observed in different frames of protein genes (Table 2): $P(X_0, P_0) > P(X_2, P_2) > P(X_1, P_1) > P(X_2, P_1) \approx P(X_0, P_2) > P(X_1, P_0) > P(X_0, P_1) > P(X_2, P_0) \approx P(X_1, P_2)$ and $(P(X_0, P_2) - P(X_1, P_2)) > (P(X_1, P_0) - P(X_2, P_0)) > (P(X_2, P_1) - P(X_0, P_1))$ (numerical results of the analytical solutions not shown). The model (0.1, 0.1, 4.3) leads to the analytical probability $P(X_0, 0, 4.3) = 0.485$ (Fig. 6a) equal to the observed probability $P(X_0, P_0) \approx 0.485$. However, this model cannot simulated all observed frequencies exactly, e.g. the analytical probability $P(X_1, 1, 4.3) = 0.477$ (Fig. 6b) is greater than the observed probability $P(X_1, P_1) \approx 0.435$ (Table 2) which is obtained in the model at $t \approx 6.8$ (Fig. 6b). The model proposed can be improved, e.g. 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 at first is essential for reducing the great number of possible combinations and also 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 (see the Note below) 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).

Note: With the substitution matrix (61, 61) there is no formula giving the eigenvectors and the characteristic roots in 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.

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). For example, an evolutionary process with substitutions in the third codon site only ($p = q = 0$ and $r = 1 - p - q = 1$) is irrelevant (data not shown). The correlation between model and genetic reality is only verified for limited ranges of values for the parameters p, q and t (given in Section 3.3). Outside these ranges, the

behaviour of the analytical curves is completely different and has no similarity with the reality observed in protein genes (data not shown).

The complex behaviour of these analytical curves (a sum of negative exponential functions at three parameters p, q and t) giving the trinucleotide probabilities after a random evolutionary process is totally unexpected and implies two remarks. It is impossible to predict the relative variations of trinucleotides after random substitutions without modelling. On the other hand, even after a great number of substitutions, e.g. four substitutions per codon (Fig. 6a–c), the trace of primitive variations of trinucleotides (differences of probabilities) is conserved in the actual genes simulated with the model ($p = 0.1$, $q = 0.1$, $t = 4.3$) and correlated with the real actual protein genes.

As mentioned in the Section 3.2, the time t is equivalent to a mean number of substitutions per codon. Therefore, the analytical probabilities of X_0, X_1 and X_2 in the three frames after t substitutions can be approximated by computing the occurrence probabilities of X_0, X_1 and X_2 in the three frames in a simulated population S (having e.g. 100 sequences of 3000 base length to get significant statistical results) which is generated according to an independent mixing of the 20 trinucleotides of X_0 with equiprobability (1/20) (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 the inversion of the evolutionary sense (from the present to the past), i.e. to analyse the probabilities of X_0, X_1 and X_2 in the three frames after back substitutions. In this case, the initial ($t = 0$) trinucleotide probabilities $P_i(0)$ 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 analysing the probabilities of X_0, X_1 and X_2 after back substitutions in protein genes, is currently in investigation. Finally, the analytical solutions computing the trinucleotide probabilities after substitutions and after back substitutions constitute a new evolutionary method which could be applied to 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 Jean-Marc Vassards (Director of the society RVH, Mulhouse).

Appendix A. Determinant for the matrix $\mathbf{A} - \lambda\mathbf{Id}$

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 $(p/3)$**Id** (16,16) (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)

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)

N	$(q/3)$**Id**	$(q/3)$**Id**	$(q/3)$**Id**
$(q/3)$**Id**	**N**	$(q/3)$**Id**	$(q/3)$**Id**
$(q/3)$**Id**	$(q/3)$**Id**	**N**	$(q/3)$**Id**
$(q/3)$**Id**	$(q/3)$**Id**	$(q/3)$**Id**	**N**

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 $(q/3)$**Id** (4, 4) (Table 7).

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

M-(p/3)**Id**	0	0	0
0	**M**+p**Id**	0	0
0	(2p/3)**Id**	**M**-(p/3)**Id**	0
0	(p/3)**Id**	0	**M**-(p/3)**Id**

Table 7
Square submatrix (q/3)**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

leads to the following determinant for the matrix $\mathbf{A}-\lambda\mathbf{Id}$:

$$\det(\mathbf{A}-\lambda\mathbf{Id})$$
$$=\det^3(\mathbf{M}-\lambda\mathbf{Id}-(p/3)\mathbf{Id})\det(\mathbf{M}-\lambda\mathbf{Id}+p\mathbf{Id})$$
$$=\det^3(\mathbf{M}-(\lambda+p/3)\mathbf{Id})\det(\mathbf{M}-(\lambda-p)\mathbf{Id}) \quad (3)$$

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)**Id** (4,4) (Table 7). Therefore, the determinant for the matrix $\mathbf{M}-\mu\mathbf{Id}$ is as before:

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

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

$$\det(\mathbf{A}-\lambda\mathbf{Id})$$
$$=[\det^3(\mathbf{N}-(\lambda+p/3)\mathbf{Id}-(q/3)\mathbf{Id})$$
$$\times\det(\mathbf{N}-(\lambda+p/3)\mathbf{Id}+q\mathbf{Id})]^3$$
$$\times\det^3(\mathbf{N}-(\lambda-p)\mathbf{Id}-(q/3)\mathbf{Id})$$
$$\times\det(\mathbf{N}-(\lambda-p)\mathbf{Id}+q\mathbf{Id})$$

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

$$=\det^9(\mathbf{N}-(\lambda+p/3+q/3)\mathbf{Id})$$
$$\times\det^3(\mathbf{N}-(\lambda+p/3-q)\mathbf{Id})$$
$$\times\det^3(\mathbf{N}-(\lambda-p+q/3)\mathbf{Id})$$
$$\times\det(\mathbf{N}-(\lambda-p-q)\mathbf{Id}) \quad (4)$$

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

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

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

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

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

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

Arquès, D.G., Michel, C.J., 1997. A code in the protein coding genes. J. Biosystems 44, 107–134.

Béal, M.P., 1993. Codage symbolique. Masson, Paris.

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

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

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

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

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

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

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

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

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

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

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

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. USA 78, 1596–1600.

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

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

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

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

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