An Evolutionary Model of a Complementary Circular Code

Didier G. Arquès*, Jean-Paul Fallot† and Christian J. Michel†‡

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

(*Received on* 20 *June* 1996, *Accepted in revised form on* 23 *October* 1996)

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 frame 0 (a reading frame established by the ATG start trinucleotide) of protein (coding) genes of both prokaryotes and eukaryotes. This subset X_0 has the rarity property (6×10^{-8}) to be a complementary maximal circular code with two permutated maximal circular codes X_1 and X_2 in 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.

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 frame 0 of the 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 model at three parameters (p, q, k) based on an independent mixing of the 20 codons (trinucleotides in frame 0) of X_0 with equiprobability (1/20) followed by $k \approx 5$ substitutions per codon in the three codon sites in proportions $p \approx 0.1$, $q \approx 0.1$ and $r = 1 - p - q \approx 0.8$ respectively, retrieves the frequencies of X_0, X_1 and X_2 observed in the three frames of protein genes and explains these asymmetries.

1. Introduction

1.1. Historical Background

The concept of code without commas has been introduced by Crick *et al.* (1957) in order to explain how the reading of a series of nucleotides in the protein (coding) genes could code for the amino acids constituting the proteins. The two problems stressed were: why are there more codons than amino acids and how to choose the reading frame? For example, a series of nucleotides . . .AGTCCGTACGA. . . can be read in three frames: . . .AGT,CCG,TAC,GA. . ., . . .A,GTC,CGT,ACG,A. . . and . . .AG,TCC,GTA, CGA,. . . Crick *et al.* (1957) have then proposed that only 20 among 64 codons, code for the 20 amino acids. However, the determination of a set of 20 codons forming a code without commas depends on a great number of constraints. For example, the four codons with identical nucleotides AAA, CCC, GGG and TTT must be excluded from such a code. Indeed, the concatenation of AAA, for example, with itself does not allow to retrieve the frame: . . .AAA,AAA,AAA,. . . . . .A,AAA,AAA,AA. . . and . . .AA,AAA,AAA,A. . . Similarly, two codons related to circular 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)

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

0022–5193/97/060241 + 13 $25.00/0/jt960305

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 interesting property has naturally led to propose a code without commas assigning one codon per amino acid (Crick *et al.*, 1957).

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

The two discoveries that the codon TTT, an "excluded" codon in the concept of code without commas, codes for phenylalanine (Nirenberg & Matthaei, 1961) and that the protein genes are placed in the reading frame with a particular codon, namely the start codon ATG, have led to 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 & Schuster, 1978).

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. The Sections 1.2 and 1.3 are the necessary reminders of the results obtained and detailed in Arquès & Michel (1996). The 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 B be a genetic alphabet, $B_2 = \{R,Y\}$ and $B_4 = \{A,C,G,T\}$. B^* denotes the words on B of finite length including the empty word of length 0. B^+ denotes the words on B of finite length $\geqslant 1$. Let $w_1 w_2$ be the concatenation of the two words w_1 and w_2.

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

The DNA double helix is formed of two nucleotide sequences s_1 and s_2 connected with the nucleotide pairing (hydrogen bonds) according to the complementarity rule C: the nucleotide A (resp. C,G,T) in s_1 pairs with the complementary nucleotide $C(A) = T$ (resp. $C(C) = G$, $C(G) = C$, $C(T) = A$) in s_2. The extension of this rule to the alphabet B_2 leads to $C(R) = Y$ and $C(Y) = R$. The two nucleotide sequences s_1 and s_2 run in opposite directions (called antiparallel) in the DNA double helix: the trinucleotide $w = l_1 l_2 l_3$, $l_1, l_2, l_3 \in B$, in s_1 pairs with the complementary trinucleotide $C(w) = C(l_3)C(l_2)C(l_1)$ in s_2, e.g. $C(AAC) = GTT$, $C(RRY) = RYY$.

Recall of the trinucleotide circular permutation

The circular permutation P of the trinucleotide $w = l_1 l_2 l_3$, is the permutated trinucleotide $P(w) = l_2 l_3 l_1$, e.g. $P(AAC) = ACA$, $P(RRY) = RYR$.

Definition of a circular code

A subset X of B^+ is a circular code if for all $n,m \geqslant 1$ and $x_1, x_2, \ldots, x_n \in X$, $y_1, y_2, \ldots, y_m \in X$ and $p \in B^*$, $s \in B^+$, the equalities $s x_2 x_3 \ldots x_n p = y_1 y_2 \ldots y_m$ and $x_1 = ps$ imply $n = m$, $p = 1$ and $x_i = y_i$, $1 \leqslant i \leqslant n$ [Béal, 1993, Berstel & Perrin, 1985 and Fig. 1(a)]. In other terms, every word on B "written on a circle" has at most one factorization (decomposition) over X. In the following, X will be a set of words of length three letters 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 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 $B_2 = \{R,Y\}$, there are nine potential maximal (sets of two trinucleotides) circular codes (Arquès & Michel, 1996). Two among these nine sets, $X_a = \{RRY,RYY\} = RNY$ and $\{YRR,YYR\} = YNR$, are complementary maximal circular codes with two permutated maximal circular codes (called C^3 codes; Arquès & Michel, 1996). The concept of circular code is presented with the set X_a which is associated with the biological model of RNY codons (Eigen & 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 (X_a) in frame 0 (reading frame), one type of trinucleotide NRY (X_b) in frame

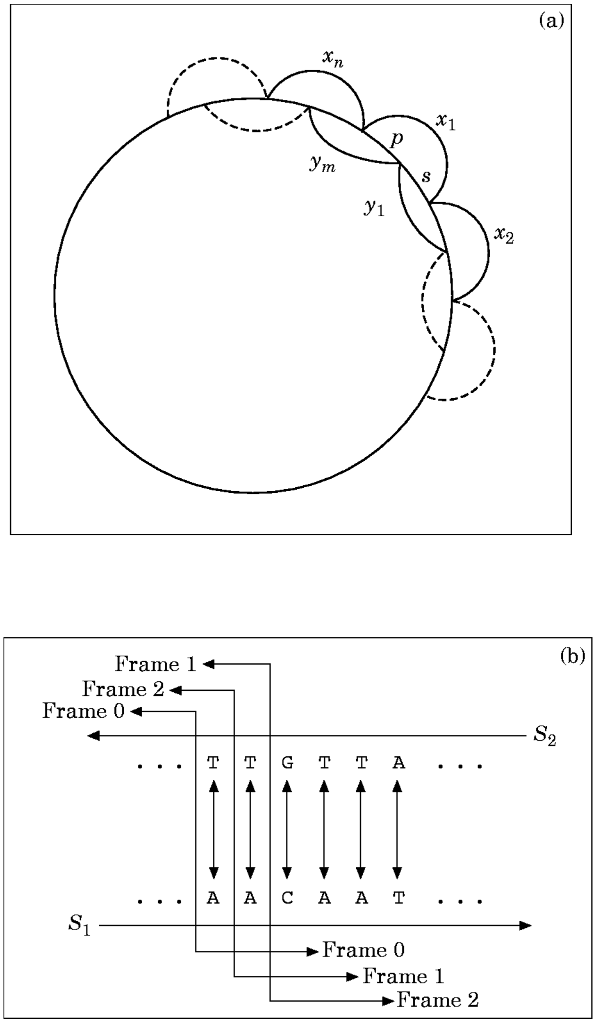

FIG. 1. (a) A representation of the definition of a circular code. (b) The complementarity property of the C^3 code X_0 implies several symmetries with the occurrence frequencies of X_0, X_1 and X_2 in the three frames, in particular the same frequency of X_1 and X_2 in the frame 0.

1 and one type of trinucleotide YRN (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.

The set $X_a = \{\text{RRY,RYY}\} = \text{RNY}$ is a maximal circular code. Indeed, the concatenation of two trinucleotides of X_a, . . .RRYRRY. . ., . . .RRYRYY. . ., . . .RYYRRY. . . and . . .RYYRYY. . ., leads to only one factorization over X_a as the eventual decomposition in frame 1 always has an R in the third position but no trinucleotide of X_a ends with R and as the eventual decomposition in frame 2 always has an Y in the first position but no trinucleotide of X_a begins with Y. Furthermore, X_a is self-complementary, i.e. $C(X_a) = X_a$, as RRY and RYY are complementary. Any subset of X_a is also a circular code but not maximal. Therefore, the RRY model (Crick *et al.*, 1976) is a non-maximal circular code. The two sets $X_b = P(X_a) = \{\text{RYR,YYR}\} = \text{NYR}$ and $X_c = P(X_b) = \{\text{YRR,YRY}\} = \text{YRN}$ obtained by circular permutations of X_a are also maximal circular codes (identical proof). Furthermore, X_b and X_c are complementary to each other, i.e. $C(X_b) = X_c$ and $C(X_c) = X_b$, as RYR (resp. YYR) and YRY (resp. YRR) are complementary. In summary, the set $X_a = \text{RNY}$ is a complementary maximal circular code with two permutated maximal circular codes $X_b = \text{NYR}$ and $X_c = \text{YRN}$ (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 $B_2 = \{\text{R,Y}\}$ where the circular codes can be completely studied by hand, the identification of a circular code on the alphabet $B_4 = \{\text{A,C,G,T}\}$ is obviously more complex and difficult as there are ≈3.5 milliard potential maximal (sets of 20 trinucleotides) circular codes (table 2d in Arquès & Michel, 1996). Unexpectedly, a simple method computing the occurrence frequencies of the 64 trinucleotides AAA, . . ., TTT in the 3 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: $T_0 = X_0 \cup \{\text{AAA,TTT}\}$ with $X_0 = \{\text{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,GAG,GAT,GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC}\}$ in frame 0, $T_1 = X_1 \cup \{\text{CCC}\}$ and $T_2 = X_2 \cup \{\text{GGG}\}$ in the shifted frames 1 and 2 respectively, with X_1 and X_2 defined in Table 1. Furthermore, the same three subsets T_0, T_1 and T_2 are retrieved with a very few exceptions for the

TABLE 1
Identification of three subsets of 20 *trinucleotides in the protein coding genes of both prokaryotes and eukaryotes*

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_1:	AGA	AGT	CAA	CAC	CAT	CCT	CGA	CGC	CGG	CGT	CTA	CTT	GCA	GCT	GGA	TAA	TAT	TGA	TGG	TGT

(Arquès & Michel, 1996): X_0 in frame 0, X_1 in frame 1 and X_2 in frame 2 (i.e. the trinucleotides of X_0 occur preferentially in frame 0 compared with frames 1 and 2, the trinucleotides of X_1 in frame 1 compared with frames 0 and 2, and the trinucleotides of X_2 in frame 2 compared with frames 0 and 1).

two large protein gene populations of prokaryotes (13 686 sequences, 4 708 758 trinucleotides) and eukaryotes (26 757 sequences, 11 397 678 trinucleotides).

The three subsets X_0, X_1 and X_2 of 20 trinucleotides have five important properties (detailed in Arquès & Michel, 1996):

(i) the property of maximal (20 trinucleotides) circular code for X_0 allowing to retrieve automatically the frame 0 in any region of a protein gene model formed by a series of trinucleotides of 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, X_1 and X_2 are also maximal circular codes;
(ii) the DNA complementarity property C: $C(X_0) = X_0$ (X_0 is self-complementary: ten trinucleotides of X_0 are complementary to the ten other trinucleotides of X_0), $C(X_1) = X_2$ and $C(X_2) = X_1$ (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) allowing the two paired reading frames of a DNA double helix simultaneously to code for amino acids, in agreement with biological results (Zull & Smith, 1990; Konecny *et al*., 1993, 1995; Béland & Allen, 1994);
(iii) the circular permutation property P: $P(X_0) = X_1$ and $P(X_1) = X_2$ (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 the two subsets X_1 and X_2 can be deduced from X_0;
(iv) the rarity property: the occurrence probability of X_0 equal to 6×10^{-8} (table 2d in Arquès & Michel, 1996) is very low and therefore, non-random in protein genes. In addition, this code 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 & Michel, 1996, Section 3.7) implying that the code X_0 has flexibility properties.

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 (C^3 code) and concatenation properties, which retrieves on the alphabet {A,C,G,T} the properties both of the code without commas on the alphabet {R,Y} (Crick *et al*., 1976; Eigen & Schuster, 1978) and of the flexible code (Dounce, 1952). Several consequences of the C^3 code X_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 & Michel, 1996). A new property, precisely an evolutionary property, of the C^3 code X_0 is identified in this paper.

1.4. IDENTIFICATION OF UNEXPECTED FREQUENCY ASYMMETRIES

As the codons in X_0 (resp. X_1 and X_2) have a preferential occurrence in the frame 0 (resp. 1 and 2) (Table 1), the global mean frequency of X_0 (resp. X_1 and X_2) in frame 0 (resp. 1 and 2) will be expected to be greater than the global mean frequencies of X_1 and X_2 (resp. X_0 and X_2, and X_0 and X_1) in frame 0 (resp. 1 and 2). Furthermore, 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. 1(b) and detailed in Section 3]. In Section 2, in order to verify these quantitative consequences of the C^3 code X_0, 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). The strong statistical properties associated with the C^3 code X_0, e.g. the preferential occurrence of X_0 (resp. X_1 and X_2) in the frame 0 (resp. 1 and 2), are indeed observed in eukaryotic protein genes. Unexpectedly, several frequency asymmetries are identified in contradiction with the complementarity property of the C^3 code X_0.

1.5. AN EVOLUTIONARY C^3 CODE

An evolutionary process by substitutions of nucleotides will allow to explain 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 20 trinucleotides of X_0 with equiprobability (1/20). These primitive genes have 20 among 64 trinucleotides and frequency symmetries (due to the complementarity property of the C^3 code X_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 X_0 allow to generate the trinucle-

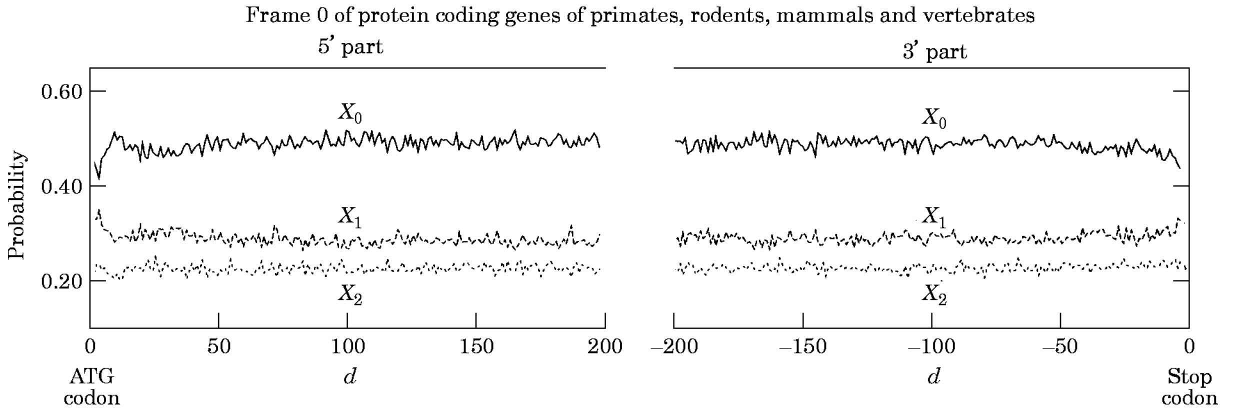

FIG. 2. Probability $P_d(X_g, F_0)$ of X_0, X_1 and X_2 at the trinucleotide distance d in the frame 0 of protein coding genes (5′ parts F_0 = P5_PRMV$_0$ and 3′ parts F_0 = P3_PRMV$_0$) of primates, rodents, mammals and vertebrates (PRMV$_0$). Three distinct horizontal curves X_0, X_1, X_2 in decreasing probabilities occur in the protein coding genes.

otides of X_1 and X_2 according to a non-balanced way and to retrieve 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 X_0. The measure of these asymmetries in the model allows to quantify 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 ≈5 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 genes.

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 $T = \{$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 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 distance d in a given frame f is defined as being the number of trinucleotides after ATG (5′ parts of protein genes) and before a stop trinucleotide (3′ parts of protein genes). A positive (resp. negative) distance 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) and $d = -10$ in $f = 0$ (resp. $f = 1, f = 2$) is the tenth codon before the chosen stop trinucleotide TAA (resp. NNT,NTA). A trinucleotide w at the distance 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 trinucleotide distance d belongs or not to X_g with $g = 0,1,2$. Then, the occurence probability $P_d(X_g, F_f)$ of a subset X_g at the trinucleotide distance d in a protein gene population F_f, is:

$$P_d(X_g, F_f) = \sum_{s \in F_f} \delta_d(X_g) \Bigg/ \sum_{g=0,1,2} \sum_{s \in F_f} \delta_d(X_g) \quad (1).$$

This probability function is represented as a curve as follows: the abscissa shows the distance d in trinucleotides, by varying d in the ranges [2200] (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:

(i) the ATG, the stop and the first ($d = -1$ and $d = 1$) conserved trinucleotides are not represented;
(ii) The four trinucleotides AAA, CCC, GGG and TTT are excluded from the statistical analysis [see the

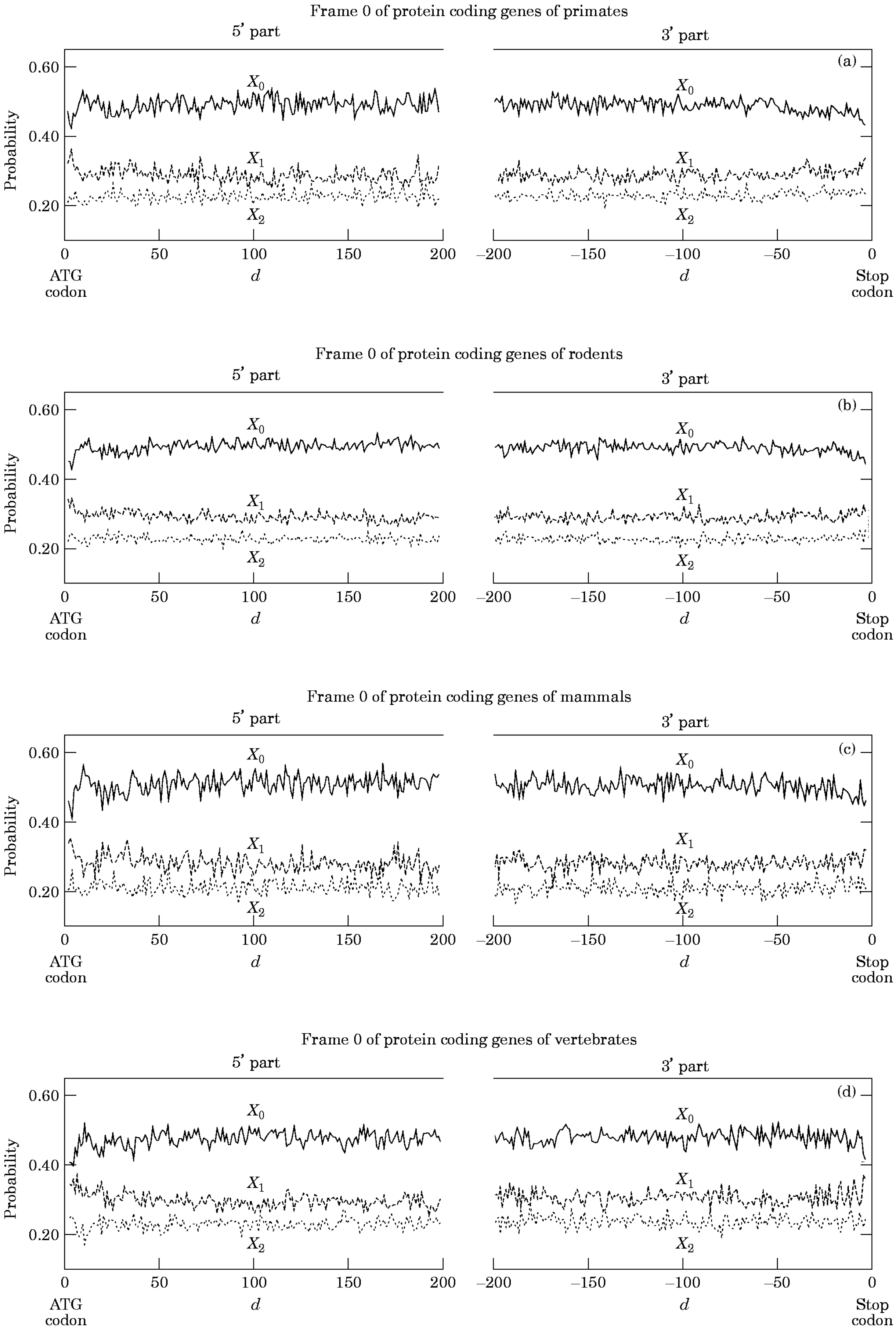

Fig. 3—(*caption opposite*)

denominator of formula (1)] 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_s have the same number of trinucleotides (20), their occurrence probabilities can be directly compared;

(iii) $P_d(X_g, R_f) = 1/3$ for any distance d with f, $g \in \{0, 1, 2\}$ in a random gene population R_f generated with an independent mixing of the four nucleotides A, C, G and T with equiprobability (1/4);

(iv) $P_d(X_f, F_f) > P_d(X_g, F_f)$ for any distance d with f, $g \in \{0, 1, 2\}$ and $g \neq f$ in a protein gene population F_f as X_0, X_1 and X_2 have been defined to have a preferential occurrence in the frames 0, 1 and 2 respectively (Table 1). The aim of the study in Section 2 consists in quantifying these probabilities.

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

(i) primates, rodents, mammals and vertebrates (PRMV): 5′ parts $F = \mathrm{P5\_PRMV}$ (17072 sequences, 2962462 trinucleotides) and 3′ parts $F = \mathrm{P3\_PRMV}$ (16972 sequences, 3144807 trinucleotides);

(ii) primates (PRI): 5′ parts $F = \mathrm{P5\_PRI}$ (6455 sequences, 1124292 trinucleotides) and 3′ parts $F = \mathrm{P3\_PRI}$ (6399 sequences, 1183741) trinucleotides);

(iii) rodents (ROD): 5′ parts $F = \mathrm{P5\_ROD}$ (6409 sequences, 1121880 trinucleotides) and 3′ parts $F = \mathrm{P3\_ROD}$ (6382 sequences, 1184633 trinucleotides);

(iv) mammals (MAM): 5′ parts $F = \mathrm{P5\_MAM}$ (1991 sequences, 339888 trinucleotides) and 3′ parts $F = \mathrm{P3\_MAM}$ (1983 sequences, 367183 trinucleotides);

(v) vertebrates (VRT): 5′ parts $F = \mathrm{P5\_VRT}$ (2217 sequences, 376402 trinucleotides) and 3′ parts $F = \mathrm{P3\_VRT}$ (2208 sequences, 409250 trinucleotides).

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 & Michel (1987, 1990) for a description of data acquisitions], allow to have stable frequencies [consequence of the law of large numbers (Arquès & Michel, 1990) Section 2.3.3].

2.2. RESULTS

Figure 2 shows that the probability curve X_0 is, as expected, greater than the two curves X_1 and X_2, for any distance trinucleotide d in the frame 0 of 5′ parts $[P_d(X_g, \mathrm{P5\_PRMV}_0)]$ and 3′ parts $[P_d(X_g, \mathrm{P3\_PRMV}_0)]$ of protein genes of primates, rodents, mammals and vertebrates. The curve X_0 is globally horizontal with an average frequency around 48.5% in $\mathrm{P\_PRMV}_0$ ($\mathrm{P5\_PRMV}_0$ and $\mathrm{P3\_PRMV}_0$) (Table 2). The two curves X_1 and X_2 are also globally horizontal but unexpectedly, distinct (Fig. 2) with an average frequency around 29% of X_1 greater than the X_2 frequency around 22.5% in $\mathrm{P\_PRMV}_0$ (Table 2). The probability difference $P_d(X_1, \mathrm{P\_PRMV}_0)\text{-}P_d(X_2, \mathrm{P\_PRMV}_0) \approx 0.065$ in frame 0 can be explained neither by the difference $1/60 \approx 0.017$ consequent on the fact that X_1 has one stop trinucleotide less than X_2 ($\mathrm{TAG} \in X_1$ and $\mathrm{TAA,TGA} \in X_2$, Table 1) nor by the complementarity property of the C^3 code X_0. The probabilities in frame 0 can be represented by the following set Q_0 of inequalities: $P_d(X_0, \mathrm{P\_PRMV}_0) > P_d(X_1, \mathrm{P\_PRMV}_0) > P_d(X_2, \mathrm{P\_PRMV}_0)$.

The horizontality as well as the frequency of the three curves are retrieved by increasing the distance trinucleotide d, e.g. [2500] and $[-500, -2]$, in the frame 0 of protein genes of primates, rodents, mammals and vertebrates (data not shown).

These results are similar in the frame 0 of each protein gene subpopulations: primates [Fig. 3(a): $P_d(X_g, \mathrm{P5\_PRI}_0)$ and $P_d(X_g, \mathrm{P3\_PRI}_0)$], rodents [Fig. 3(b): $P_d(X_g, \mathrm{P5\_ROD}_0)$ and $P_d(X_g, \mathrm{P3\_ROD}_0)$], mammals [Fig. 3(c): $P_d(X_g, \mathrm{P5\_MAM}_0)$ and $P_d(X_g, \mathrm{P3\_MAM}_0)$] and vertebrates [Fig. 3(d): $P_d(X_g, \mathrm{P5\_VRT}_0)$ and $P_d(X_g, \mathrm{P3\_VRT}_0)$].

Figure 4 (resp. 5) shows that the probability curve X_1 (resp. X_2) is, as expected, greater than the two curves X_2 and X_0 (resp. X_0 and X_1), for any trinucleotide distance d in the frame 1 (resp. 2) of 5′ and 3′ parts of protein genes of primates, rodents, mammals and vertebrates [Fig. 4: $P_d(X_g, \mathrm{P5\_PRMV}_1)$ and $Pd(X_g, \mathrm{P3\_PRMV}_1)$] [resp. Fig. 5:

FIG. 3 (a) Probability $P_d(X_g, F_0)$ of X_0, X_1 and X_2 at the trinucleotide distance d in the frame 0 of protein coding genes (5′ parts $F_0 = \mathrm{P5\_PRI}_0$ and 3′ parts $F_0 = \mathrm{P3\_PRI}_0$) 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 distance d in the frame 0 of protein coding genes (5′ parts $F_0 = \mathrm{P5\_ROD}_0$ and 3′ parts $F_0 = \mathrm{P3\_ROD}_0$) 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 distance d in the frame 0 of protein coding genes (5′ parts $F_0 = \mathrm{P5\_MAM}_0$ and 3′ parts $F_0 = \mathrm{P3\_MAM}_0$) 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 distance d in the frame 0 of protein coding genes (5′ parts $F_0 = \mathrm{P5\_VRT}_0$ and 3′ parts $F_0 = \mathrm{P3\_VRT}_0$) of vertebrates (VRT_0). Three distinct horizontal curves X_0, X_1, X_2 at decreasing probabilities occur in the protein coding genes of vertebrates.

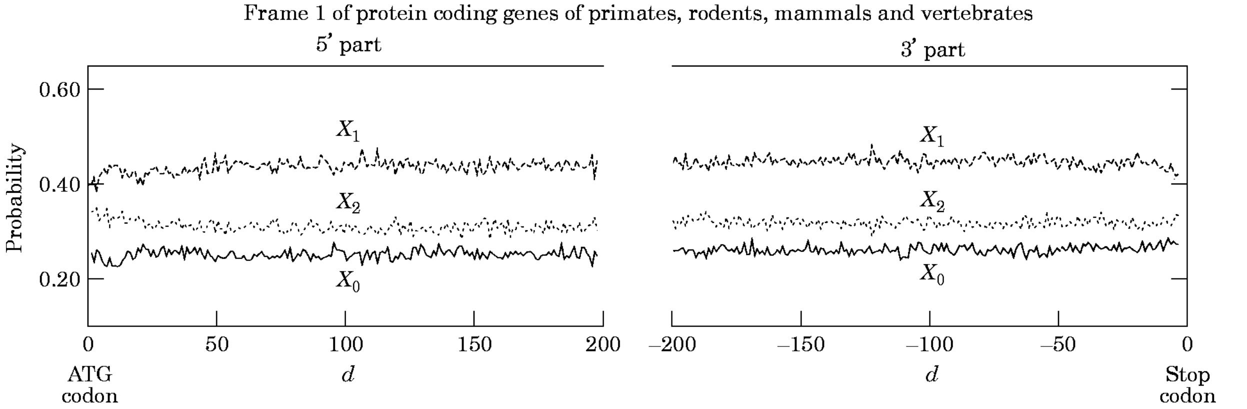

FIG. 4. Probability $P_d(X_g, F_1)$ of X_0, X_1 and X_2 at the trinucleotide distance d in the frame 1 of protein coding genes (5′ parts F_1 = P5_PRMV$_1$ and 3′ parts F_1 = P3_PRMV$_1$) of primates, rodents, mammals and vertebrates (PRMV$_1$). Three distinct horizontal curves X_1, X_2, X_0 in decreasing probabilities occur in the protein coding genes.

$P_d(X_g$, P5_PRMV$_2)$ and $P_d(X_g$, P3_PRMV$_2)$]. In the frame 1, the three curves, X_1, X_2 and X_0 are globally horizontal with an average frequency 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 the 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 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$_2)$ which is unexpected as the complementarity property of the C^3 code X_0 would have led to $P_d(X_2$, P_PRMV$_2)$ > $P_d(X_1$, P_PRMV$_2)$ > $P_d(X_0$,P_PRMV$_2)$.

The horizontality as well as the frequency of the three curves are retrieved by increasing the distance trinucleotide d in the frames 1 and 2 of protein genes of primates, rodents, mammals and vertebrates and of its subpopulations (data not shown).

3. An Evolutionary 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 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 $P(X_g, F_f)$ [instead of $P_d(X_g, F_f)$] (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 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_PRMV$_0)$ > $P(X_1$, P_PRMV$_0)$ > $P(X_2$, P_PRMV$_0)$, Q_1 in frame 1: $P(X_1$, P_PRMV$_1)$ > $P(X_2$, P_PRMV$_1)$ > $P(X_0$, P_PRMV$_1)$

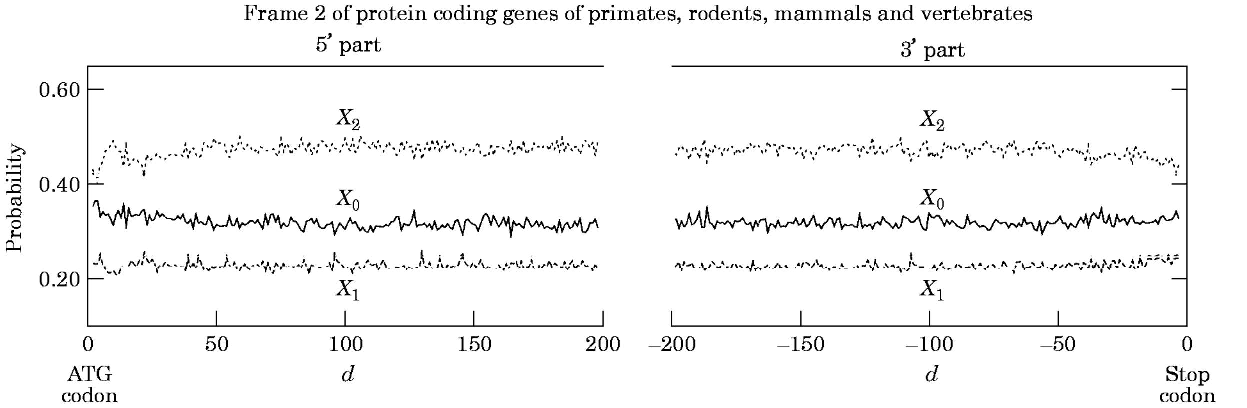

FIG. 5. Probability $P_d(X_g, F_2)$ of X_0, X_1 and X_2 at the trinucleotide distance d in the frame 2 of protein coding genes (5′ parts F_2 = P5_PRMV$_2$ and 3′ parts F_2 = P3_PRMV$_2$) of primates, rodents, mammals and vertebrates (PRMV$_2$). Three distinct horizontal curves X_2, X_0, X_1 in decreasing probabilities occur in the protein coding genes.

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* (*with rounded averages*)

	Protein coding genes of primates, rodents, mammals and vertebrates (PRMV)								
	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_1	22.4	22.7	22.5	31.1	31.0	31.0	46.4	46.1	46.5

and Q_2 in frame 2: $P(X_2, \text{P\_PRMV}_2) > P(X_0, \text{P\_PRMV}_2) > P(X_1, \text{P\_PRMV}_2)$ (Table 2). As detailed in Section 2.2, these probability inequalities are in contradiction with the complementarity property of the C^3 code X_0. 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 $P(X_2, S_2) > P(X_1, S_2) > P(X_0, S_2)$.

A new property of the C^3 code X_0 related to substitutions is studied in this section. Precisely, the problem investigated here is whether an evolutionary model can explain the actual properties of the C^3 code X_0 observed in protein genes. The main evolutionary process of protein genes is determined by substitutions of nucleotides. RNA editing (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). As the actual protein genes have a preferential occurrence of the subset X_0 [in frame 0; note also that $P(X_0, \text{P\_PRMV}_0) > P(X_2, \text{P\_PRMV}_2) > P(X_1, \text{P\_PRMV}_1)$ in Table 2], the model which will be tested, is based on two processes:

(i) a construction process generating simulated genes according to an independent mixing of the 20 trinucleotides of X_0 with equiprobability (1/20). Note: a construction process, less simple, based on a Markov mixing of trinucleotides could also have been considered;
(ii) an evolutionary process based on substitutions in the three trinucleotide sites transforming the simulated genes into evolutionary simulated genes.

A solution of this model is obtained when the evolutionary simulated genes verify the three sets Q_0, Q_1 and Q_2 of inequalities, and have frequencies of X_0, X_1 and X_2 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 (substitutions, insertions and deletions of nucleotides), have already been developed on the purine/pyrimidine alphabet (Arquès & Michel, 1990, 1992, 1993).

3.2. METHOD

The substitution process in the model is characterized by a mean number $k \in [0, 20]$ of substitutions per codon (trinucleotide in frame 0) in the three codon sites in proportions p, q and $r = 1 - p - q$ with $p, q, r \in [0, 1]$. The substitution sites are randomly chosen in the sequences. For example, a sequence of 100 codons after $k = 0.1$ substitutions per codon in site proportions $p = 0.2$, $q = 0.3$ and $r = 1 - 0.2 - 0.3 = 0.5$ means that $100 \times 0.1 =$ ten codons randomly chosen in the sequence have mutated, $10 \times 0.2 =$ two codons in the first site, $10 \times 0.3 =$ three codons in the second site and $10 \times 0.5 =$ five codons in the third site. The type of substitutions is random but the generation of a stop trinucleotide TAA, TAG or TGA in frame 0 of simulated genes is not allowed during the substitution process.

A simulated population S, having 500 sequences of 3000 base length, is generated according to an independent mixing of the 20 trinucleotides of X_0 with equiprobability (1/20) (construction process, i.e. $k = 0$). The computations obtained with such a sample of 1.5 million bases are precise. Note: a sample having 100 sequences of 3000 base length leads to similar results. Then, for given site proportions p and q (r being the complement to 1) of substitutions, this population S is subjected to k substitutions per codon which are randomly applied to each sequence of S (substitution process, i.e. $k > 0$). At each substitution step k, the occurrence probabilities $P(X_g, S_f, k)$ of X_0, X_1 and X_2 in the three frames $f = 0$, 1, 2 of simulated

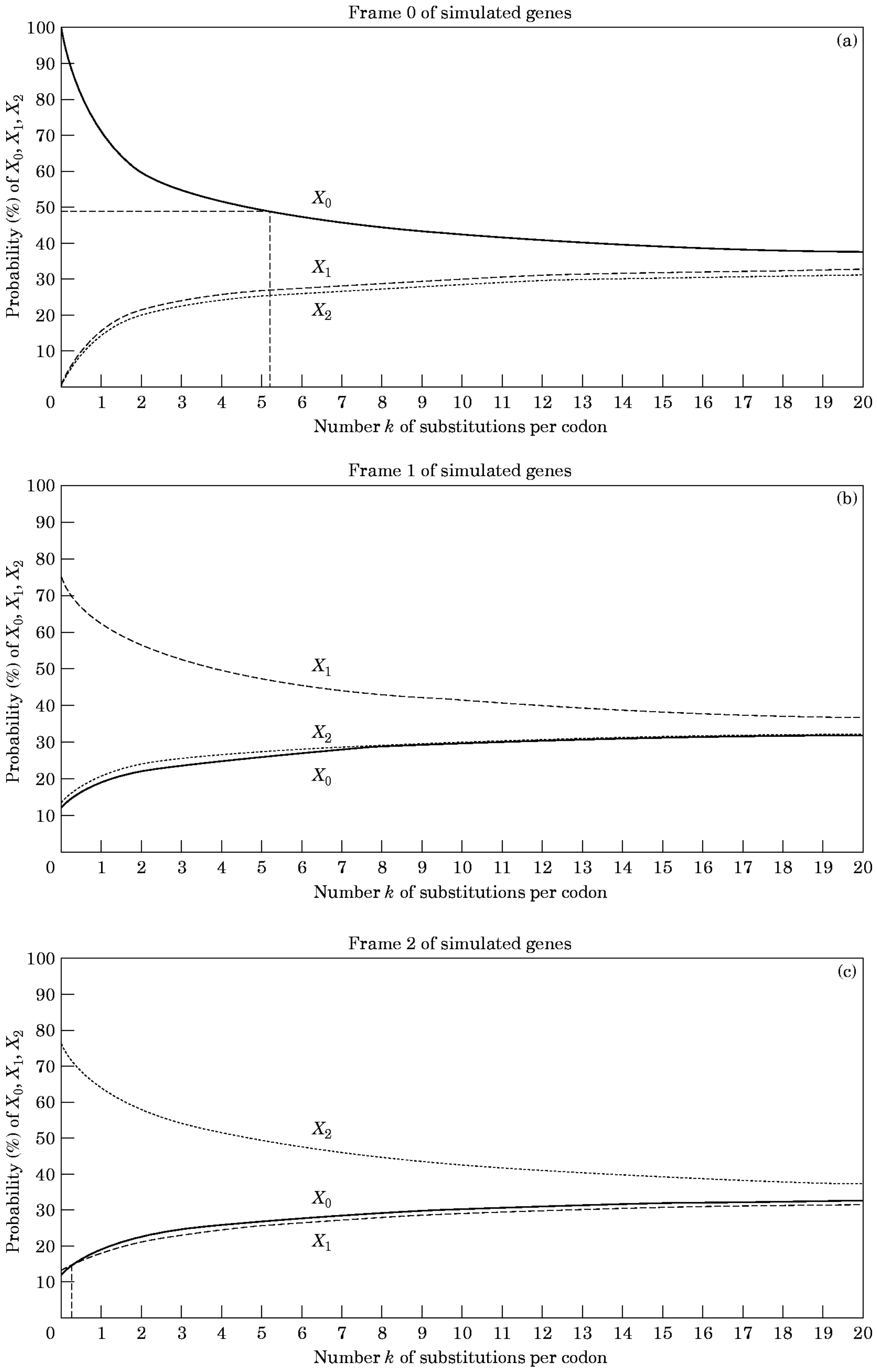

Fig. 6.—(*caption opposite*)

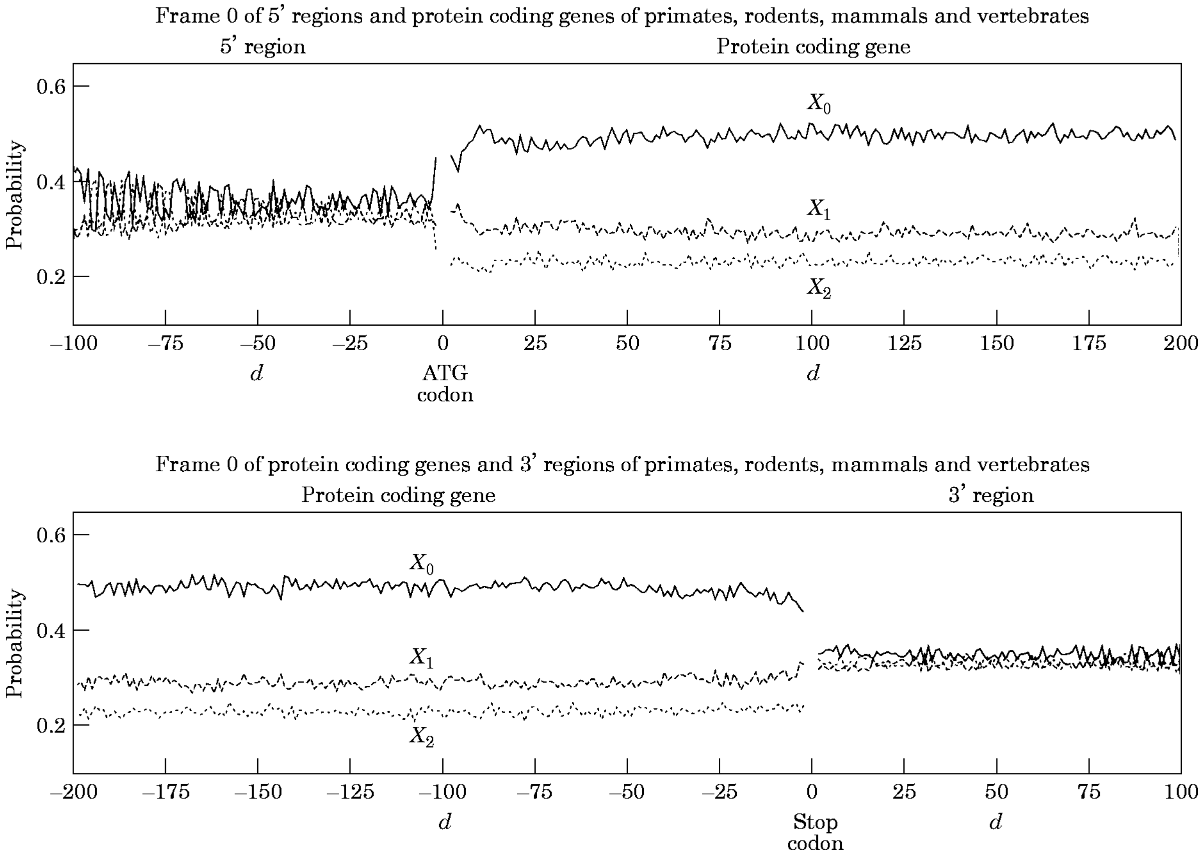

FIG. 7. (a) Probability $P_d(X_g, F_0)$ for X_0, X_1 and X_2 at the trinucleotide distance d in the extended frame 0 of 5′ regions (F_0 = R5_PRMV$_0$) and 5′ parts of protein coding genes (F_0 = P5_PRMV$_0$) of primates, rodents, mammals and vertebrates (PRMV$_0$). Three distinct horizontal curves X_0, X_1, X_2 in decreasing probabilities occur in the protein coding genes while the three curves are mixed in the 5′ regions. (b) Probability $P_d(X_g, F_0)$ of X_0, X_1 and X_2 at the trinucleotide distance 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.

genes ($F = S_f$) are computed. The model (p, q, k) has a solution if there are values for the three parameters p, q and k verifying the three sets Q_0, Q_1 and Q_2 of inequalities and leading to frequencies of X_0, X_1 and X_2 in the three frames in the order of those observed in the actual protein genes (Table 2 and Section 3.1).

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 k in the range [0, 20] with a step of 0.1, the model (p, q, k) retrieves the three sets Q_0, Q_1 and Q_2 of inequalities observed in 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 $k \geqslant 0.3$ [Fig. 6(a–c)]. Furthermore, $P(X_0, S_0, k) \approx 0.485$ at $k \approx 5.2$ [Fig. 6(a)].

At the construction process ($k = 0$), the model ($p = 0.1, q = 0.1, k = 0$) leads to the following probabilities, in frame 0 $P(X_0, S_0, 0) = 1$ and $P(X_1, S_0, 0) = P(X_2, S_0, 0) = 0$ [Fig. 6(a)], in frame 1

FIG. 6. (a) Probability $P(X_g, S_0, k)$ of X_0, X_1 and X_2 in the frame 0 of simulated genes (S_0) generated with an independent mixing of the 20 trinucleotides of X_0 with equiprobability (1/20) and subjected to k substitutions per codon in the three codon sites in proportions $p = 0.1$ $q = 0.1$ and $r = 1 - p - q = 0.8$ respectively. At $k \approx 5.2$ substitutions, the three simulated curves show occurrence probabilities of X_0, X_1, X_2 similar to those observed in frame 0 of protein coding genes. (b) Probability $P(X_g, S_1, k)$ of X_0, X_1 and X_2 in the frame 1 of simulated genes (S_1) generated with an independent mixing of the 20 trinucleotides of X_0 with equiprobability (1/20) and subjected to k substitutions per codon in the three codon sites in proportions $p = 0.1$ $q = 0.1$ and $r = 1 - p - q = 0.8$ respectively. At $k \approx 5.2$ substitutions, the three simulated curves show occurrence probabilities of X_1, X_2, X_0 similar to those observed in frame 1 of protein coding genes. (c) Probability $P(X_g, S_2, k)$ of X_0, X_1 and X_2 in the frame 2 of simulated genes (S_2) generated with an independent mixing of the 20 trinucleotides of X_0 with equiprobability (1/20) and subjected to k substitutions per codon in the three codon sites in proportions $p = 0.1$ $q = 0.1$ and $r = 1 - p - q = 0.8$ respectively. At $k \approx 0.3$ substitutions, the two simulated curves X_0 and X_1 cross and retrieve the probability inequality Q_2 observed in frame 2 of protein coding genes. The inequality Q_2 is not verified for $k < 0.3$. At $k \approx 5.2$ substitutions, the three simulated curves show occurrence probabilities of X_2, X_0, X_1 similar to those observed in frame 2 of protein coding genes.

$P(X_1, S_1, 0) = 0.754$, $P(X_2, S_1, 0) = 0.127$ and $P(X_0, S_1, 0) = 0.119$ [Fig. 6(b)] and in frame 2 $P(X_2, S_2, 0) = 0.754$, $P(X_1, S_2, 0) = 0.127$ and $P(X_0, S_2, 0) = 0.119$ [Fig. 6(c)], i.e. to the following inequalities, in frame 0 $P(X_0, S_0, 0) > P(X_1, S_0, 0) = P(X_2, S_0, 0)$, in frame 1 $P(X_1, S_1, 0) > P(X_2, S_1, 0) > P(X_0, S_1, 0)$ and in frame 2 $P(X_2, S_2, 0) > P(X_1, S_2, 0) > P(X_0, S_2, 0)$ which result from the complementarity property of the C^3 code X_0. These probability inequalities in simulated genes before the substitution process differ from the inequalities Q_0, Q_1 and Q_2 observed in protein genes. The inequality $P(X_1, \text{P\_PRMV}_0) > P(X_2, \text{P\_PRMV}_0)$ observed in frame 0 of protein genes exists in the model $(p = 0.1, q = 0.1, k)$ for a substitution number $k > 0$ [Fig. 6(a)]. The inequality $P(X_2, \text{P\_PRMV}_1) > P(X_0, \text{P\_PRMV}_1)$ observed in frame 1 of protein genes exists in the model $(p = 0.1, q = 0.1, k)$ for a substitution number $k \geqslant 0$ [Fig. 6(b)]. The inequality $P(X_0, \text{P\_PRMV}_2) > P(X_1, \text{P\_PRMV}_2)$ observed in frame 2 of protein genes exists in the model $(p = 0.1, q = 0.1, k)$ for a substitution number $k \geqslant 0.3$ (note that $P(X_1, S_2, k) > P(X_0, S_2, k)$ for $k < 0.3$) [Fig. 6(c)]. In summary, the construction process $(k = 0)$ generates only the inequality $P(X_2, S_1, 0) > P(X_0, S_1, 0)$ in frame 1 of simulated genes. The substitution process $(k > 0)$ generates the two inequalities $P(X_1, S_0, k) > P(X_2, S_0, k)$ and $P(X_0, S_2, k) > P(X_1, S_2, k)$ in frames 0 and 2, respectively, of simulated genes and increases the amplitude of the inequality $P(X_2, S_1, 0) > P(X_0, S_1, 0)$ in frame 1 of simulated genes.

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 rarity property (6×10^{-8}) to be a complementary maximal circular code with two permutated maximal circular codes X_1 and X_2 (C^3 code, Section 1.3). The quantitative study of the three subsets X_0, X_1 and X_2 in the three frames 0, 1, 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 the 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 $F = \text{R5\_PRMV}$ (6997 sequences, 443 351 trinucleotides) and 3′ regions $F = \text{R3\_PRMV}$ (14 315 sequences, 1 098 651 trinucleotides) of primates, rodents, mammals and vertebrates where the three subsets X_0, X_1 and X_2 occur with variable frequencies around the random value (1/3) [Fig. 7(a) and (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. By considering these values and those obtained in the shifted frames with 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 & McLachlan, 1982; Fickett, 1982; Smith *et al.*, 1983; Blaisdell, 1983, etc).

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 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 with concatenation properties of flexibility (Section 1.3) allowing its evolution. 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 would have been too complex at this time. The model also demonstrates that a substitution process $(k > 0)$ must follow the construction process, e.g. for generating the two inequalities $P(X_1, S_0, k) > P(X_2, S_0, k)$ and $P(X_0, S_2, k) > P(X_1, S_2, k)$ in frames 0 and 2 respectively, or for decreasing the initial probabilities of $P(X_0, S_0, k)$, $P(X_1, S_1, k)$ and $P(X_2, S_2, k)$ in frames 0, 1 and 2 respectively. The substitutions in the model occur with the highest rate in the third codon site (around 0.8 representing $5.2 \times 0.8 \approx 4$ 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 (around 0.1 representing $5.2 \times 0.1 \approx 1/2$ 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).

This simple model ($p = 0.1, q = 0.1, k = 5.2$) developed allows to retrieve not only the three sets Q_0, Q_1 and Q_2 of inequalities (Section 2.2) and the frequency order of X_0, X_1 and X_2 in the three frames but also several other sets of probability inequalities observed in protein genes (Table 2): $P(X_0, \text{P\_PRMV}_0) > P\ (X_2, \text{P\_PRMV}_2) > P(X_1, \text{P\_PRMV}_1) > P(X_2, \text{P\_PRMV}_1) \approx P(X_0, \text{P\_PRMV}_2) > P(X_1, \text{P\_PRMV}_0) > P(X_0, \text{P\_PRMV}_1) > P(X_2, \text{P\_PRMV}_0)$ $P(X_1, \text{P\_PRMV}_2)$ and $[P(X_0, \text{P\_PRMV}_2) - P(X_1, \text{P\_PRMV}_2)] > [P(X_1,\text{P\_PRMV}_0) - P(X_2,\text{P\_PRMV}_0)] > [P(X_2, \text{P\_PRMV}_1) - P(X_0, \text{P\_PRMV}_1)]$ (numerical results of the simulation not shown). The model $(0.1, 0.1, 5.2)$ leads to a simulated probability $P(X_0, S_0, 5.2) = 0.485$ equal to the observed probability $P(X_0, \text{P\_PRMV}_0) \approx 0.485$. All observed frequencies cannot exactly be simulated, e.g. the simulated probability $P(X_1, S_1, 5.2) = 0.470$ is equal to the observed probability $P(X_1, \text{P\_PRMV}_1) \approx 0.435$ (Table 2) at $k \approx 7.6$ [Fig. 6(b)]. A simple model cannot simulate completely the biological reality depending on a great number of factors, in the same way that the first terms of development of a function in series cannot reveal the totality of the function. The model proposed can be improved, e.g. by suppressing the strong constraint of a constant proportion of substitutions in the three codon sites during all the substitution process. The analytical solutions giving the probabilities of the eight codons on the alphabet {R,Y} after substitutions (Arquès & Michel, 1994) are currently generalized to the 64 codons on the alphabet {A,C,G,T} in order to determine the exact probabilities of X_0, X_1 and X_2 after substitutions. Nevertheless, the investigation of simple models 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.

This work was supported by GIP GREG grant (Groupement d'Intérêt Public, Groupement de Recherches et d'Etudes sur les Génomes), INSERM grant (Contrat de Recherche Externe No 930101) and Mr Jean-Marc Vassards (Director of the Society RVH, Mulhouse).

REFERENCES

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

Arquès, D. G. & Michel, C. J. (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. Biosc.* **123,** 103–125.

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

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

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

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 *coxII* 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.

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.* **43,** 416–421.

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

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

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

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

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

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

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 dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci.* **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.

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

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.

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

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.