An Evolutionary Model of a Complementary Circular Code

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

* Équipe de Biologie Théorique, Université de Marne-la-Vallée, Institut Gaspard Monge, 2 rue de la butte verte, 93160 Noisy le Grand, and † Équipe 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 ($6 \times 10^{-8}$) to be a complementary maximal circular code with two permuted 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 C3 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 C3 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 $k$ 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.

© 1997 Academic Press Limited

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: . . .AG, TCC, GT, ACG, A. . . and . . .A, GTCCGTACGA. . . can be read in three frames: . . .AG, TCC, GT, ACG, A. . . and . . .AG, TCC, GT, ACG, A. . . Thus, 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 . . .A, 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) . . .A, GTC, CGT, ACG, A. . . and . . .AG, TCC, GTA, CGA, . . .
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 “written on a circle” has at most one factorization (decomposition) over $B^+$ (resp. $C(R)$ = Y and $C(G)$ = C, $C(T)$ = A) in $s_2$. The extension of this rule to the alphabet $B_1$ 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 $w = l_1l_2l_3$, $l_1,l_2,l_3 \in B_1$ 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 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 $B_1$ 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 $w = l_1l_2l_3$, $l_1,l_2,l_3 \in B_1$ 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 $w = l_1l_2l_3$, $l_1,l_2,l_3 \in B_1$ is the permuted trinucleotide $P(w) = l_2l_1l_3$, e.g. $P(AAC) = ACA, P(RRY) = RYY$.

Definition of a circular code

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 $sx_1x_2\ldots x_np = y_1y_2\ldots y_m$ and $x_1 = ps$ imply $n = m$, $p = 1$, and $s_1 = s_2 \leq n$ (Beal, 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 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 $B_1 = \{R,Y\}$, there are nine potential maximal (sets of two trinucleotides) circular codes (Arqués & Michel, 1996). Two among these nine sets, $X_0 = \{ RRY,RYY \} = RNY$ and $\{ YRR,YRY \} = YNR$, are complementary maximal circular codes with two permuted maximal circular codes (called $C^*$ codes; Arqués & Michel, 1996). The concept of circular code is presented with the set $X_0$ 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 ($X_0$) in frame 0 (reading frame), one type of trinucleotide NRY ($X_0$) in frame...
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 $X_c = \{\text{RYY}, \text{RYR}\} = \text{RNY}$ is a maximal circular code. Indeed, the concatenation of two trinucleotides of $X_c$, $\ldots \text{RRYRYY} \ldots$, leads to only one factorization over $X_c$ as the eventual decomposition in frame 1 always has an R in the third position but no trinucleotide of $X_c$ ends with R and as the eventual decomposition in frame 2 always has an Y in the first position but no trinucleotide of $X_c$ begins with Y. Furthermore, $X_c$ is self-complementary, i.e. $C(X_c) = X_c$ as RRY and RYY are complementary. Any subset of $X_c$ 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 $X_c = P(X_c) = \{\text{YRR, YRY}\} = \text{YRN}$ 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 permuted 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 $B_2 = \{R,Y\}$ where the circular codes can be completely studied by hand, the identification of a circular code on the alphabet $B_4 = \{A,C,G,T\}$ is obviously more complex and difficult as there are $\approx 3.5$ milliard potential maximal (sets of 20 trinucleotides) circular codes (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 $X_0 = \{\text{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,GAG,GAT,GCC,GGC,GGT,GTG,TA}, \text{GTC,GT}T, \text{TAC,TTC}\}$ 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

<table>
<thead>
<tr>
<th>Identification of three subsets of 20 trinucleotides in the protein coding genes of both prokaryotes and eukaryotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_0$: AAC AAT ACC ATC ATT CAG CTC CTG GAA GAC GAG GAT GCC GGC GGT GTA GTC GTT TAC TCC</td>
</tr>
<tr>
<td>$X_1$: AAG ACA ACG ACT AGT ATA ATG CCA CCG GCC GGT GGA GTA TCA TCC TCG TGT TGT</td>
</tr>
<tr>
<td>$X_2$: AGA AGT CAA CAC CAT CCT CGA CGG CGT GTA CTA TGG GGA TAA TAT TGA TGG TGT</td>
</tr>
</tbody>
</table>

(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 (13686 sequences, 4708758 trinucleotides) and eu-
karyotes (26757 sequences, 11397678 trinucleotides).

The three subsets \( X_0, X_1 \) and \( X_2 \) of 20 trinucleotides have five important properties (detailed in Arqués & 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 coding 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; Konceny 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 \( 6 \times 10^{-8} \) (table 2d in Arqués & 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: 13686 sequences and eukaryotes: 26757 sequences);

(v) three concatenation properties (Arqués & 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 permuted maximal circular codes (\( C^0 \) 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 \( C^1 \) 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 (Arqués & Michel, 1996). A new property, precisely an evolutionary property, of the \( C^1 \) code \( X_0 \) is identified in this paper.

1.4. IDENTIFICATION OF UNEXPECTED FREQUENCY ASYM-
METRIES

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 \( C^1 \) 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 \( C^1 \) 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 \( C^1 \) 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 \( C^1 \) code \( X_0 \).

1.5. AN EVOLUTIONARY \( C^1 \) 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 \( C^1 \) 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-
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 $\approx 5$ substitutions per codon in the three codon sites in proportions $\approx 0.1$, $\approx 0.1$ and $\approx 0.8$ 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 $T = \{\text{AAA}, \ldots, \text{TTT}\}$ (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,NNT,NNN,TAA and $f=1, A,TGN, \ldots, NNT,AA$ and $f=2, AT,GNN, \ldots, NTA,AA$ (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 $X_f$ 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

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

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

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

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 $P_d(X_f, F_f)$, $P_d(X_s, F_s)$ and $P_d(X_z, F_z)$ 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
Frame 0 of protein coding genes of primates

Frame 0 of protein coding genes of rodents

Frame 0 of protein coding genes of mammals

Frame 0 of protein coding genes of vertebrates

Fig. 3—(caption opposite)
denominator of formula (1)] as $X_a$, $X_1$ and $X_2$ are circular codes and thereby, containing no trinucleotide with identical nucleotides. As $X_a$, $X_1$ and $X_2$ have the same number of trinucleotides (20), their occurrence probabilities can be directly compared;

(iii) $P_d(X_f, R_f) = 1/3$ 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) $P_d(X_f, F_f) > P_d(X_g, F_g)$ for any distance $d$ with $f, g \in \{0, 1, 2\}$ and $g \neq f$ in a protein gene population $F_f$ as $X_a$, $X_1$ 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 $F = P_5 \text{PRMV}$ (17072 sequences, 2962462 trinucleotides) and 3' parts $F = P_3 \text{PRMV}$ (16972 sequences, 3144807 trinucleotides);

(ii) primates (PRI): 5' parts $F = P_5 \text{PRI}$ (16972 sequences, 3144807 trinucleotides) and 3' parts $F = P_3 \text{PRI}$ (6399 sequences, 1183741 trinucleotides);

(iii) rodents (ROD): 5' parts $F = P_5 \text{ROD}$ (6409 sequences, 1121880 trinucleotides) and 3' parts $F = P_3 \text{ROD}$ (6382 sequences, 1184633 trinucleotides);

(iv) mammals (MAM): 5' parts $F = P_5 \text{MAM}$ (1919 sequences, 339888 trinucleotides) and 3' parts $F = P_3 \text{MAM}$ (1983 sequences, 367183 trinucleotides);

(v) vertebrates (VRT): 5' parts $F = P_5 \text{VRT}$ (2217 sequences, 376402 trinucleotides) and 3' parts $F = P_3 \text{VRT}$ (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 $[P_d(X_0, P_5 \text{PRMV}_0)]$ and 3' parts $[P_d(X_0, P_3 \text{PRMV}_0)]$ of protein genes of primates, rodents, mammals and vertebrates. The curve $X_0$ is globally horizontal with an average frequency around 48.5% in $P_5 \text{PRMV}_0$ ($P_5 \text{PRMV}_0$ and $P_3 \text{PRMV}_0$) (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 $P_5 \text{PRMV}_0$ (Table 2). The probability difference $P_d(X_0, P_5 \text{PRMV}_0) - P_d(X_1, P_5 \text{PRMV}_0) \approx 0.065$ 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 $X_2$ (TAG $\in X_1$ and TAA, TGA $\not\in X_2$, Table 1) nor by the complementarity property of the C' code $X_0$. The probabilities in frame 0 can be represented by the following set $Q_0$ of inequalities: $P_d(X_0, P_5 \text{PRMV}_0) > P_d(X_1, P_5 \text{PRMV}_0) > P_d(X_2, P_5 \text{PRMV}_0)$.

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): $P_d(X_g, P_5 \text{PRI}_0)$ and $P_d(X_g, P_3 \text{PRI}_0)$], rodents [Fig. 3(b): $P_d(X_g, P_5 \text{ROD}_0)$ and $P_d(X_g, P_3 \text{ROD}_0)$], mammals [Fig. 3(c): $P_d(X_g, P_5 \text{MAM}_0)$ and $P_d(X_g, P_3 \text{MAM}_0)$] and vertebrates [Fig. 3(d): $P_d(X_g, P_5 \text{VRT}_0)$ and $P_d(X_g, P_3 \text{VRT}_0)$].

Figure 4 (resp. 5) shows that the probability curve $X_1$ (resp. $X_2$) is, as expected, greater than the two curves $X_1$ 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: $P_d(X_g, P_5 \text{PRMV}_1)$ and $P_d(X_g, P_3 \text{PRMV}_1)$] [resp. Fig. 5:
Frame 1 of protein coding genes of primates, rodents, mammals and vertebrates

\[
\begin{align*}
\text{5' part} & \\
\text{3' part} & \\
\end{align*}
\]

Frame 2 of protein coding genes of primates, rodents, mammals and vertebrates

\[
\begin{align*}
\text{5' part} & \\
\text{3' part} & \\
\end{align*}
\]

3. An Evolutionary Model of the C³ Code \(X_o\)

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 \(P(X_o, F_i)\) [instead of \(P_d(X_o, F_i)\)] (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 \(P(X_o, F_i)\) 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: \(P(X_o, P_{PRMV0}) > P(X_1, P_{PRMV0}) > P(X_2, P_{PRMV0})\), \(Q_1\) in frame 1: \(P(X_1, P_{PRMV1}) > P(X_2, P_{PRMV1}) > P(X_0, P_{PRMV1})\), \(Q_2\) in frame 2: \(P(X_2, P_{PRMV2}) > P(X_0, P_{PRMV2}) > P(X_1, P_{PRMV2})\).
and \( Q_2 \) in frame 2: \( P(X_5, P \text{ PRMV}_2) > P(X_6, P \text{ PRMV}_2) > P(X_5, P \text{ PRMV}_1) \) (Table 2). As detailed in Section 2.2, these probability inequalities are in contradiction with the complementarity property of the C\(^1\) 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\(^1\) 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\(^1\) 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 \( P(X_0, P \text{ PRMV}_0) > P(X_1, P \text{ PRMV}_0) > P(X_2, P \text{ PRMV}_0) \) 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).

### Table 2

<table>
<thead>
<tr>
<th>Protein coding genes of primates, rodents, mammals and vertebrates (PRMV)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Frame 0</strong></td>
</tr>
<tr>
<td>5' parts</td>
</tr>
<tr>
<td>( P_5 \text{ PRMV}_0 )</td>
</tr>
<tr>
<td>( X_0 )</td>
</tr>
<tr>
<td>( X_1 )</td>
</tr>
<tr>
<td>( X_2 )</td>
</tr>
</tbody>
</table>

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 \( 100 \times 0.1 \times 0.2 = 2 \) 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. 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 \( P(X_i, S_i, k) \) 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 $P_d(X_0, F_s)$ for $X_0$, $X_1$ and $X_2$ at the trinucleotide distance $d$ in the extended frame 0 of 5' regions ($F_s = R5$ PRMVv) and 5' parts of protein coding genes ($F_s = P5$ PRMVv) of primates, rodents, mammals and vertebrates (PRMVv). Three distinct horizontal curves $X_0$, $X_1$ and $X_2$ in decreasing probabilities occur in the protein coding genes while the three curves are mixed in the 5' regions. (b) Probability $P_d(X_0, F_3)$ at the trinucleotide distance $d$ in the extended frame 0 of 3' parts of protein coding genes ($F_s = P3$ PRMVv) and 3' regions ($F_s = R3$ PRMVv) of primates, rodents, mammals and vertebrates (PRMVv). Three distinct horizontal curves $X_0$, $X_1$ and $X_2$ in decreasing probabilities occur in the protein coding genes while the three curves are mixed in the 3' regions.

Fig. 6. (a) Probability $P(X_0, S_0, k)$ 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 $X_0$, $X_1$ and $X_2$ similar to those observed in frame 0 of protein coding genes. (b) Probability $P(X_1, S_1, k)$ of $X_1$ 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 $p = 1 - q = 0.1$ and $r = 1 - p - q = 0.8$ respectively. At $k \approx 5.2$ substitutions, the three simulated curves show occurrence probabilities of $X_1$, $X_2$ similar to those observed in frame 1 of protein coding genes. (c) Probability $P(X_2, S_2, k)$ of $X_2$ 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 $X_2$ 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 $X_1$, $X_2$, similar to those observed in frame 2 of protein coding genes.

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 $k \geq 0.3$ [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...
The inequality \( P(X_1, S_1, 0) = 0.754 \), \( P(X_2, S_1, 0) = 0.127 \) and \( P(X_3, S_1, 0) = 0.119 \) [Fig. 6(b)] and in frame 2 \( P(X_1, S_2, 0) = 0.754 \), \( P(X_2, S_2, 0) = 0.127 \) and \( P(X_3, S_2, 0) = 0.119 \) [Fig. 6(c)], i.e. to the following inequalities, in frame 0 \( P(X_2, S_0, 0) > P(X_1, S_0, 0) = P(X_3, S_0, 0) \), in frame 1 \( P(X_1, S_0, 0) > P(X_2, S_1, 0) > P(X_3, S_0, 0) \) and in frame 2 \( P(X_1, S_0, 0) > P(X_2, S_2, 0) > P(X_3, S_0, 0) \) which result from the complementarity property of the \( C^1 \) 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 \( P(X_1, P_{PRMV_0}) > P(X_2, P_{PRMV_0}) \) 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 \( P(X_1, P_{PRMV_1}) > P(X_2, P_{PRMV_1}) \) observed in frame 1 of protein genes in the model \( p = 0.1, q = 0.1, k \) for a substitution number \( k > 0 \) [Fig. 6(b)]. The inequality \( P(X_1, P_{PRMV_2}) > P(X_2, P_{PRMV_2}) \) observed in frame 2 of protein genes exists in the model \( p = 0.1, q = 0.1, k \) for a substitution number \( k > 0 \) [Fig. 6(c)]. In summary, the construction process \( (k = 0) \) generates only the inequality \( P(X_1, S_1, 0) > P(X_0, S_1, 0) \) in frame 1 of simulated genes. The substitution process \( (k > 0) \) generates the two inequalities \( P(X_1, S_1, k) > P(X_2, S_0, k) \) and \( P(X_3, S_2, k) > P(X_1, S_1, k) \) in frames 0 and 2, respectively, of simulated genes and increases the amplitude of the inequality \( P(X_1, S_1, 0) > P(X_0, S_1, 0) \) 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 \( (6 \times 10^{-4}) \) to be a complementary maximal circular code with two permuted maximal circular codes \( X_1 \) and \( X_2 \) \( (C^1 \) 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 \( F = R_5 \) PRMV \( (6997 \) sequences, 443351 trinucleotides) and 3’ regions \( F = R_3 \) PRMV \( (14315 \) sequences, 1098651 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, et al).

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 \( C^1 \) 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 \( P(X_3, S_1, k) > P(X_1, S_1, k) \) 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 \(X_0, X_1, X_2\) in the three frames but also several other sets of probability inequalities observed in protein genes (Table 2): 

\[
P(X_0, P_{PRMV_0}) > P(X_1, P_{PRMV_0}) > P(X_2, P_{PRMV_0})
\]

This simple model \((0.1, 0.1, 5.2)\) leads to a simulated probability 

\[
P(X_0, X_1, X_2) = 0.485 \text{ equal to the observed probability } P(X_0, P_{PRMV_0}) \approx 0.485.
\]

All observed frequencies cannot exactly be simulated, e.g. the simulated probability \(P(X_1, X_2) = 0.470\) is equal to the observed probability \(P(X_1, P_{PRMV_0}) \approx 0.435\) (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 \(X_0, X_1, X_2\) 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 (Groupe d’Intérêt Public, Groupe de Recherches et d’Études sur les Génomes), INSERM grant (Contrat de Recherche Externe No 930101) and Mr Jean-Marc Vassards (Director of the Society RVH, Mulhouse).

REFERENCES


