Bulletin of Mathematical Biology (2004) **66**, 763–778
doi:10.1016/j.bulm.2003.10.004

A Stochastic Gene Evolution Model with Time Dependent Mutations

JACQUES M. BAHI

LIFC - FRE CNRS 2661,
IUT de Belfort,
Université de Franche-Comté,
BP 527,
90016 Belfort Cédex,
France
E-mail: bahi@iut-bm.univ-fcomte.fr

CHRISTIAN J. MICHEL*

Equipe de Bioinformatique Théorique,
LSIIT (UMR CNRS-ULP 7005),
Université Louis Pasteur de Strasbourg,
Pôle API,
Boulevard Sébastien Brant,
67400 Illkirch,
France
E-mail: michel@dpt-info.u-strasbg.fr

We develop here a new class of gene evolution models in which the nucleotide mutations are time dependent. These models allow to study nonlinear gene evolution by accelerating or decelerating the mutation rates at different evolutionary times. They generalize the previous ones which are based on constant mutation rates. The stochastic model developed in this class determines at some time t the occurrence probabilities of trinucleotides mutating according to 3 time dependent substitution parameters associated with the 3 trinucleotide sites. Therefore, it allows to simulate the evolution of the circular code recently observed in genes. By varying the class of function for the substitution parameters, 1 among 12 models retrieves after mutation the statistical properties of the observed circular code in the 3 frames of actual genes. In this model, the mutation rate in the 3rd trinucleotide site increases during gene evolution while the mutation rates in the 1st and 2nd sites decrease. This property agrees with the actual degeneracy of the genetic code. This approach can easily be generalized to study evolution of motifs of various lengths, e.g., dicodons, etc., with time dependent mutations.

*Author to whom correspondence should be addressed.

0092-8240/04/040763 + 16 $30.00/0

1. INTRODUCTION

1.1. ***Presentation of the approach.*** The trinucleotide distribution in (protein coding) genes is not random. Indeed, a few trinucleotides occur with higher frequencies in the reading frame of genes (Grantham *et al*., 1980). This trinucleotide usage preference has been related to several biological factors, including translational selection (Shpaer, 1986; Akashi and Eyre-Walker, 1998), GC composition (Jukes and Bhushan, 1986; Konu and Li, 2002), strand-specific mutational bias (Sharp and Matassi, 1994; Berg and Silva, 1997; Campbell *et al*., 1999), transcriptional selection, RNA stability and tRNA content (Ikemura, 1985) [see also the review Ermolaeva (2001)]. In this line of research, we have recently identified a preferential subset of 20 trinucleotides in the reading frame of genes (Arquès and Michel, 1996). By convention, the reading frame established by the codon ATG is the frame 0, and the frames 1 and 2 are the reading frame 0 shifted in the $5' - 3'$ direction by 1 and 2 nucleotides respectively. The occurrence frequencies of the 64 trinucleotides $AAA, \ldots, TTT$ are computed in the 3 frames of genes. Then, a preferential frame for the 64 trinucleotides can be deduced by assigning to each trinucleotide the frame associated with its highest occurrence frequency. Totally unexpected, by excluding the identical trinucleotides (AAA, CCC, GGG and TTT) and with a few exceptions, this approach identifies 3 subsets X_0, X_1 and X_2 of 20 trinucleotides in the frames 0, 1 and 2 respectively of genes. Furthermore, the same 3 subsets X_0, X_1 and X_2 are found in 2 large and different gene populations of eukaryotes (26 757 sequences, 11 397 678 trinucleotides) and prokaryotes (13 686 sequences, 4 708 758 trinucleotides) (Arquès and Michel, 1996). The subset X_0 of 20 trinucleotides in frame 0 is $X_0 = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$ (Arquès and Michel, 1996). Unexpectedly, the start codon ATG does not belong to X_0. However, the scanning mechanism for initiation of translation in eukaryotes is based on the consensus sequence $GCCRCCATG(R = \{A, G\})$ [review in Kozak (2002)]. Surprisingly, the 2 trinucleotides preceding ATG belong to $X_0(ACC, GCC \in X_0)$. Therefore, this motif of 6 base length could have been the translation initiation signal in primitive genes. The 2 subsets X_1 and X_2 of 20 trinucleotides in the frames 1 and 2 respectively of genes can be deduced from the subset X_0 by the permutation $\mathcal{P}$ (Arquès and Michel, 1996): $\mathcal{P}(X_0) = X_1$ and $\mathcal{P}(\mathcal{P}(X_0)) = \mathcal{P}(X_1) = X_2$ knowing that the (left circular) permutation $\mathcal{P}$ of a trinucleotide $w_0 = l_0 l_1 l_2, l_0, l_1, l_2 \in \mathbb{A} = \{A, C, G, T\}$, is the permuted trinucleotide $\mathcal{P}(w_0) = w_1 = l_1 l_2 l_0$, e.g., $AAC \in X_0$ implies $\mathcal{P}(AAC) = ACA \in X_1$ and $\mathcal{P}(\mathcal{P}(AAC)) = CAA \in X_2$. The same 3 subsets X_0, X_1 and X_2 found in both eukaryotic and prokaryotic genes, have interesting properties, in particular X_0, X_1 and X_2 are circular codes (Arquès and Michel, 1996).

The observation of a preferential subset of trinucleotides in the reading frame of genes is the basis of our development of an evolutionary model. Indeed, if a trinucleotide preferential subset occurs with frequencies higher than the random

one in the reading frame of actual genes after (mainly) random evolution, then a realistic hypothesis of an evolutionary model consists in asserting that this subset had higher frequencies in the past compared to the actual time, i.e., in the reading frame of 'primitive' genes (genes before evolution). As X_0 is a trinucleotide preferential subset in the reading frame of actual genes, we take the hypothesis that the primitive genes are constructed by trinucleotides of X_0. Therefore, the evolutionary model proposed is based on 2 processes: a construction process with a random mixing of the 20 trinucleotides of X_0 with equiprobability $(1/20)$ followed by an evolutionary process with random substitutions in the 3 trinucleotide sites. These random substitutions are modelled by 3 time dependent parameters. Therefore, the stochastic model developed here is based on a time dependent mutation matrix. It generalizes the previous models with constant mutation matrices, in particular the matrices 4×4 for the 4 nucleotides at 1 substitution parameter (Jukes and Cantor, 1969), 2 parameters (transition and transversion) (Kimura, 1980), 3 and 6 parameters (Kimura, 1981), and the matrix 64×64 for the 64 trinucleotides at 3 constant substitution parameters (Arquès *et al*., 1998). Otherwise, the probabilistic model based on the nucleotide frequencies with a hypothesis of absence of correlation between successive bases on a DNA strand (Koch and Lehmann, 1997) cannot generate X_0 (Lacan and Michel, 2001). The evolutionary model proposed here allows to retrieve the circular code X_0 in actual genes and to identify evolutionary properties of X_0.

In the next 2 sections, the 2 steps of our approach are briefly detailed: the observation of a circular code in genes and the evolutionary model. In particular, some quantitative results used in the model are given.

1.2. *A circular code in genes*

1.2.1. *Definition and recall of a few basic properties [detailed in Arquès and Michel (1996) in particular]. Recall of a few notations.* $\mathbb{A}$ being a finite alphabet, $\mathbb{A}^*$ denotes the words on $\mathbb{A}$ of finite length including the empty word of length 0 and $\mathbb{A}^+$, the words on $\mathbb{A}$ of finite length greater or equal to 1. Let $w_1 w_2$ be the concatenation of the 2 words w_1 and w_2.

DEFINITION 1. A subset X of $\mathbb{A}^+$ is a circular code if $\forall n, m \geq 1$ and $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m \in X$, and $r \in \mathbb{A}^*, s \in \mathbb{A}^+$, the equalities $s x_2 \ldots x_n r = y_1 y_2 \ldots y_m$ and $x_1 = rs$ imply $n = m, r = 1$ and $x_i = y_i, 1 \leq i \leq n$ (Berstel and Perrin, 1985; Béal, 1993).

A circular code is a set of words on an alphabet such as any word written on a circle (the next letter after the last letter of the word being the first letter) has a unique decomposition into words of the circular code. For example, let X be the 6 word set $X = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ and consider as the word w, the sequence of 9 letters $w = ATGGCCCTA$. The word w can be factorized

circularly in 2 different ways, ATG, GCC, CTA and AAT, GGC, CCT. Therefore, the set X is not a circular code. But the set $\tilde{X}$ obtained by replacing the last word GGC of X by GTC, $\tilde{X} = \{AAT, ATG, CCT, CTA, GCC, GTC\}$, is a circular code as the factorization of w, and more generally of any word generated by $\tilde{X}$, is unique.

An important property of a circular code is the automatic retrieval of the construction frame of a word. Indeed, the construction frame of a word generated by a concatenation of the words of a circular code can be retrieved after the reading, anywhere in the word, of a certain number of nucleotides depending on the code. This series of nucleotides is called the window of the circular code. The main biological consequence of this property is the ability to retrieve automatically the reading frame in genes (i.e., without start codon) which might be involved in the transcription and the translation apparatus (Arquès and Michel, 1996).

Property 1: X_0, X_1 and X_2 are maximal (20 trinucleotides) circular codes [proof in Arquès and Michel (1996)].

Property 2: X_0 is a C^3 code.

As X_0, X_1 and X_2 are circular codes (Property 1) and related to each other by permutation such as $\mathcal{P}(X_0) = X_1$ and $\mathcal{P}(X_1) = X_2$ (see Section 1.1), each circular code can be deduced from the permutation of another circular code. The code X_1 associated with the frame defined as the (left) permutation of the reading frame, i.e., the frame 1, can be obtained by the permutation of the code X_0 in frame 0. Similarly, the code X_2 in frame 2 can be deduced from the permutation of the code X_1 and the code X_0, from the code X_2. As the code X_0 is coding for the reading frame in genes, it is considered as the main code and then called C^3 code (maximal circular code with 2 permuted maximal circular codes).

Property 3: X_0, X_1 and X_2 are also related to each other by the complementarity $\mathcal{C}$: $\mathcal{C}(X_0) = X_0$ (X_0 is self-complementary) and $\mathcal{C}(X_1) = X_2$ (X_1 and X_2 are complementary to each other) knowing that the complementarity $\mathcal{C}$ of a trinucleotide $w_0 = l_0 l_1 l_2, l_0, l_1, l_2 \in \mathbb{A} = \{A, C, G, T\}$, is the complementary trinucleotide $\mathcal{C}(w_0) = \mathcal{C}(l_2)\mathcal{C}(l_1)\mathcal{C}(l_0)$ with $\mathcal{C}(A) = T$, $\mathcal{C}(C) = G$, $\mathcal{C}(G) = C$, $\mathcal{C}(T) = A$, e.g., $AAC \in X_0$ implies $\mathcal{C}(AAC) = GTT \in X_0$ and $ACA \in X_1$ implies $\mathcal{C}(ACA) = TGT \in X_2$.

Other properties, such as the rarity and the flexibility, as well as the different biological consequences, in particular on the 2-letter genetic alphabets, the genetic code and the amino acid frequencies in proteins, are given and detailed in Arquès and Michel (1996).

1.2.2. *Actual probabilities of the circular code in genes.* Let $\mathcal{P}(X_j, f, t_{\text{actual}})$ be the occurrence probability of a circular code X_j, $j \in \{0, 1, 2\}$, in the frame f, $f \in \{0, 1, 2\}$, of genes at the actual time t_{actual}. The computation of the 9 occurrence probabilities of X_0, X_1 and X_2 in the 3 frames in a eukaryotic gene

population (34 144 genes), leads to the following actual values: $\mathcal{P}(X_0, 0, t_{\text{actual}}) = 0.485$, $\mathcal{P}(X_1, 0, t_{\text{actual}}) = 0.29$, $\mathcal{P}(X_2, 0, t_{\text{actual}}) = 0.225$, $\mathcal{P}(X_0, 1, t_{\text{actual}}) = 0.255$, $\mathcal{P}(X_1, 1, t_{\text{actual}}) = 0.435$, $\mathcal{P}(X_2, 1, t_{\text{actual}}) = 0.31$, $\mathcal{P}(X_0, 2, t_{\text{actual}}) = 0.31$, $\mathcal{P}(X_1, 2, t_{\text{actual}}) = 0.225$ and $\mathcal{P}(X_2, 2, t_{\text{actual}}) = 0.465$. According to the law of large numbers, all these probabilities are stable and significant. They are retrieved with other gene populations (data not shown).

Therefore, the following actual probability inequalities at the actual time t_{actual} can be deduced

$$\begin{cases} \mathcal{P}(X_0, 0, t_{\text{actual}}) > \mathcal{P}(X_1, 0, t_{\text{actual}}) > \mathcal{P}(X_2, 0, t_{\text{actual}}) \text{ in frame } 0 \\ \mathcal{P}(X_1, 1, t_{\text{actual}}) > \mathcal{P}(X_2, 1, t_{\text{actual}}) > \mathcal{P}(X_0, 1, t_{\text{actual}}) \text{ in frame } 1 \\ \mathcal{P}(X_2, 2, t_{\text{actual}}) > \mathcal{P}(X_0, 2, t_{\text{actual}}) > \mathcal{P}(X_1, 2, t_{\text{actual}}) \text{ in frame } 2. \end{cases} \tag{1.1}$$

1.3. *An evolutionary model of a circular code*

1.3.1. *The construction process.* The construction process generates 'primitive' genes according to a random mixing of the 20 trinucleotides of X_0 with equiprobability $(1/20)$. The occurrence probability $\mathcal{P}(X_j, f, t_0)$ of a circular code X_j, $j \in \{0, 1, 2\}$, in the frame f, $f \in \{0, 1, 2\}$, of genes at the initial past time t_0 can be easily determined. Obviously, in frame 0, $\mathcal{P}(X_0, 0, t_0) = 1$ and $\mathcal{P}(X_1, 0, t_0) = \mathcal{P}(X_2, 0, t_0) = 0$ (absence of X_1 and X_2 in frame 0). By considering the 400 pairs of trinucleotides of X_0 and by computing exactly the number of trinucleotides in their frames 1 and 2 belonging to X_0, X_1 or X_2, then the following probabilities are obtained: $\mathcal{P}(X_0, 1, t_0) = 0.119$, $\mathcal{P}(X_1, 1, t_0) = 0.754$, $\mathcal{P}(X_2, 1, t_0) = 0.127$, $\mathcal{P}(X_0, 2, t_0) = 0.119$, $\mathcal{P}(X_1, 2, t_0) = 0.127$ and $\mathcal{P}(X_2, 2, t_0) = 0.754$. A few trinucleotides of X_0 and X_2 (resp. X_0 and X_1) occur in frame 1 (resp. 2) (not detailed). The different symmetries observed with these probabilities are the consequence of the complementarity property of the C^3 code X_0.

REMARK 1. The comma free code of Crick *et al.* (1957) has stronger conditions compared to the circular code. Indeed, the 20 trinucleotides of the comma free code which code for the 20 amino acids, are found in 1 frame only.

Therefore, the following past probability inequalities at the initial past time t_0 can be deduced

$$\begin{cases} \mathcal{P}(X_0, 0, t_0) > \mathcal{P}(X_1, 0, t_0) = \mathcal{P}(X_2, 0, t_0) \text{ in frame } 0 \\ \mathcal{P}(X_1, 1, t_0) > \mathcal{P}(X_2, 1, t_0) > \mathcal{P}(X_0, 1, t_0) \text{ in frame } 1 \\ \mathcal{P}(X_2, 2, t_0) > \mathcal{P}(X_1, 2, t_0) > \mathcal{P}(X_0, 2, t_0) \text{ in frame } 2. \end{cases} \tag{1.2}$$

These past probability inequalities are not observed in the actual genes. Indeed, the actual probability inequalities have unexpected asymmetries in contradiction with the complementarity property of the C^3 code X_0 [see the probability inequalities (1.1)]:

(i) the frequency of the code X_1 is higher than the code X_2 in frame 0 of actual genes, i.e., $\mathcal{P}(X_1, 0, t_{\text{actual}}) > \mathcal{P}(X_2, 0, t_{\text{actual}})$, while these 2 codes do not exist in frame 0 of primitive genes, i.e., $\mathcal{P}(X_1, 0, t_0) = \mathcal{P}(X_2, 0, t_0) = 0$.
(ii) the frequency of the code X_0 is higher than the code X_1 in frame 2 of actual genes, i.e., $\mathcal{P}(X_0, 2, t_{\text{actual}}) > \mathcal{P}(X_1, 2, t_{\text{actual}})$, while an inverse situation exists in frame 2 of primitive genes, i.e., $\mathcal{P}(X_1, 2, t_0) > \mathcal{P}(X_0, 2, t_0)$.

Therefore, an evolutionary process is added to the construction one in order to retrieve the same probability inequalities with the actual and modelled genes.

1.3.2. *The evolutionary process.* The evolutionary process is based on random substitutions according to 3 time dependent parameters associated with the 3 trinucleotide sites. It transforms the primitive genes into simulated actual ones. Substitutions with different time dependent functions in the sites of the trinucleotides of X_0 will allow to generate the trinucleotides of X_1 and X_2 according to a nonbalanced way and then, to retrieve the asymmetrical probability inequalities (1.1) of actual genes. The aim of the mathematical model proposed consists in determining the occurrence probabilities of the circular codes X_0, X_1 and X_2 in the 3 frames during evolution (Section 2).

Twelve models are analysed by varying the class of function for the 3 time dependent parameters. Only one model retrieves the circular code observed in actual genes after a certain evolutionary time. Furthermore, this model identifies a property with the mutation rates in the trinucleotide sites.

2. MATHEMATICAL MODEL

The mathematical model will allow to determine at an evolutionary time t the occurrence probability $\mathcal{P}(X_j, f, t)$ of a circular code X_j in the frame f of genes whose trinucleotides mutate according to 3 time dependent substitution parameters $p(t)$, $q(t)$ and $r(t)$ associated with the 3 trinucleotide sites respectively.

By convention, the indexes $i, j, k \in \{0, \ldots, 63\}$ represent the trinucleotides $AAA, \ldots, TTT$ in the alphabetical order. The occurrence probability $P_i(t + dt)$ of a trinucleotide i at a time $t+dt$ is equal to the occurrence probability $P_i(t)$ of this trinucleotide i at the time t minus the substitution probability of this trinucleotide i during $[t, t + dt]$ and plus the substitution probabilities of the trinucleotides j, $j \neq i$, into the trinucleotide i during $[t, t + dt]$

$$P_i(t + dt) = P_i(t) - \alpha dt P_i(t) + \alpha dt \sum_{j=0}^{63} P(j \to i) P_j(t) \tag{2.1}$$

where α is the probability that a trinucleotide is subjected to 1 substitution during a unit interval of time and where $P(j \to i)$ is the substitution probability of a trinucleotide j into a trinucleotide i. The probability $P(j \to i)$ is equal to 0 if the

substitution is impossible (j and i differ more than one nucleotide as dt is assumed to be small enough that a codon cannot mutate successively twice during $[t, t+dt]$) otherwise it is given in the function of the 3 substitution rates $p(t)$, $q(t)$ and $r(t)$. For example with the trinucleotide AAA associated with $i = 0$, $P(CAA \to AAA) = P(GAA \to AAA) = P(TAA \to AAA) = p(t)/3$, $P(ACA \to AAA) = P(AGA \to AAA) = P(ATA \to AAA) = q(t)/3$, $P(AAC \to AAA) = P(AAG \to AAA) = P(AAT \to AAA) = r(t)/3$, and $P(j \to AAA) = 0$ with $j \notin \{AAC, AAG, AAT, ACA, AGA, ATA, CAA, GAA, TAA\}$.

With an appropriate unit of time, the probability α is equal to 1, i.e., there is 1 substitution per codon per unit of time. Then, the formula (2.1) becomes

$$\frac{P_i(t+dt) - P_i(t)}{dt} \approx P_i'(t) = -P_i(t) + \sum_{j=0}^{63} P(j \to i) P_j(t). \tag{2.2}$$

By considering the column vector $P(t) = (P_i(t))_{0 \le i \le 63}$ made of the 64 $P_i(t)$ and the mutation matrix $A(t)$ (64, 64) of the 4096 trinucleotide substitution probabilities $P(j \to i)$, the differential equation (2.2) can be represented by the following matrix equation

$$P'(t) = -P(t) + A(t) \cdot P(t) = (A(t) - I) \cdot P(t) \tag{2.3}$$

where I represents the identity matrix and the symbol $\cdot$, the matrix product.

REMARK 2. The square matrix $A(t)$ (64, 64) can be defined by a square block matrix (4, 4) whose 4 diagonal elements are formed by 4 identical square submatrices $B(t)$ (16, 16) and whose 12 nondiagonal elements are formed by 12 identical square submatrices $(p(t)/3)I$ (16, 16)

$$A(t) = \left(\begin{array}{c|cccc} & 0\cdots15 & 16\cdots31 & 32\cdots47 & 48\cdots63 \\ \hline 0\cdots15 & B(t) & (p(t)/3)I & (p(t)/3)I & (p(t)/3)I \\ 16\cdots31 & (p(t)/3)I & B(t) & (P(t)/3)I & (p(t)/3)I \\ 32\cdots47 & (p(t)/3)I & (p(t)/3)I & B(t) & (p(t)/3)I \\ 48\cdots63 & (p(t)/3)I & (p(t)/3)I & (p(t)/3)I & B(t) \end{array}\right).$$

The index ranges $\{0, \ldots, 15\}$, $\{16, \ldots, 31\}$, $\{32, \ldots, 47\}$ and $\{48, \ldots, 63\}$ are associated with the trinucleotides $\{AAA, \ldots, ATT\}$, $\{CAA, \ldots, CTT\}$, $\{GAA, \ldots, GTT\}$ and $\{TAA, \ldots, TTT\}$ respectively. The square submatrix $B(t)$ (16,16) can again be defined by a square block matrix (4, 4) whose 4 diagonal elements are formed by 4 identical square submatrices $C(t)$ (4, 4) and whose 12 nondiagonal elements are formed by 12 identical square submatrices $(q(t)/3)I$ (4, 4)

$$B(t) = \begin{pmatrix} C(t) & (q(t)/3)I & (q(t)/3)I & (q(t)/3)I \\ (q(t)/3)I & C(t) & (q(t)/3)I & (q(t)/3)I \\ (q(t)/3)I & (q(t)/3)I & C(t) & (q(t)/3)I \\ (q(t)/3)I & (q(t)/3)I & (q(t)/3)I & C(t) \end{pmatrix}.$$

Finally, the square submatrix $C(t)$ (4, 4) is equal to

$$C(t) = \begin{pmatrix} 0 & r(t)/3 & r(t)/3 & r(t)/3 \\ r(t)/3 & 0 & r(t)/3 & r(t)/3 \\ r(t)/3 & r(t)/3 & 0 & r(t)/3 \\ r(t)/3 & r(t)/3 & r(t)/3 & 0 \end{pmatrix}.$$

The matrix $A(t)$ is stochastic when $p(t) + q(t) + r(t) = 1$.

The differential equation (2.3) can then be written in the following form

$$P'(t) = M(t) \cdot P(t)$$

$$M(t) = A(t) - I.$$

Suppose that for a sampling $t_0 < t_1 < \cdots < t_n$, $A(t)$ is a constant matrix on the interval $[t_h, t_{h+1}]$, then denote

$$A(t) = A_h, \qquad \forall t \in [t_h, t_{h+1}].$$

This equation means that although the mutation matrix $A(t)$ is not constant in the entire time interval, there exist (sufficiently small) periods of time in which the mutation factors are constant.

With this realistic hypothesis in mind, the equation (2.3) can be written as follows

$$P'(t) = M_h \cdot P(t), \qquad \forall t \in [t_h, t_{h+1}]$$

where

$$M_h = A_h - I.$$

For $t \in [t_h, t_{h+1}]$, the probability $P(t)$ is then computed by the formulae

$$P(t) = e^{M_h(t-t_h)} \cdot e^{M_{h-1}(t_h-t_{h-1})} \cdots e^{M_1(t_2-t_1)} \cdot e^{M_0(t_1-t_0)} P(0)$$

where $e^{M_h(t-t_h)}$ is the exponential of $M_h(t - t_h)$ and the initial vector $P(0) = \{0, 1/20, 0, 1/20, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 1/20, 0, 1/20, 0, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/20, 1/20, 0, 1/20, 1/20, 1/20, 1/20, 0, 1/20, 0, 0, 0, 1/20, 0, 1/20, 1/20, 1/20, 1/20, 0, 1/20, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/20, 0, 0\}$ (the components of $P(0)$ associated with the trinucleotides of X_0 being equal to $1/20$). Then,

$$P(t) = e^{(A_h-I)(t-t_h)} \cdot e^{(A_{h-1}-I)(t_h-t_{h-1})} \cdots e^{(A_1-I)(t_2-t_1)} \cdot e^{(A_0-I)(t_1-t_0)} P(0).$$

We note $\mathcal{P}(i, f, t)$ the occurrence probability of the trinucleotide i in the frame f, $f \in \{0, 1, 2\}$, of genes at the time t. Therefore, the probability vector $P(t)$

determining the 64 occurrence probabilities $P_i(t)$ of the 64 trinucleotides i which are in frame 0 (reading frame), can be denoted

$$\mathcal{P}(i, 0, t) = P_i(t).$$

The occurrence probability $\mathcal{P}(i, 1, t)$ of the trinucleotide i in the frame 1 of genes at the time t can be obtained from the product of the 2 probabilities $\mathcal{P}(j, 0, t)$ and $\mathcal{P}(k, 0, t)$ associated with the concatenation of the 2 trinucleotides j and k generating the trinucleotide i

$$\mathcal{P}(i, 1, t) = \sum_{j=0}^{3} \mathcal{P}\left(\left\lfloor \frac{i}{4} \right\rfloor + 16j, 0, t\right) \times \sum_{k=0}^{15} \mathcal{P}(16(i \bmod 4) + k, 0, t).$$

Similarly, the occurrence probability $\mathcal{P}(i, 2, t)$ of the trinucleotide i in the frame 2 of genes at the time t is deduced

$$\mathcal{P}(i, 2, t) = \sum_{j=0}^{15} \mathcal{P}\left(\left\lfloor \frac{i}{16} \right\rfloor + 4j, 0, t\right) \times \sum_{k=0}^{3} \mathcal{P}(4(i \bmod 16) + k, 0, t).$$

Then, the occurrence probability $\mathcal{P}(X_j,\ f, t)$ of a circular code X_j, $j \in \{0, 1, 2\}$, in the frame f, $f \in \{0, 1, 2\}$, at the substitution step t, can be obtained

$$\mathcal{P}(X_j,\ f, t) = \frac{\sum_{i \in X_j} \mathcal{P}(i,\ f, t)}{\sum_{i \in X_0 \cup X_1 \cup X_2} \mathcal{P}(i,\ f, t)}. \tag{2.4}$$

Several classes of evolutionary models are analysed with the 3 time dependent substitution parameters $p(t)$, $q(t)$ and $r(t)$. For all these models, we take the realistic hypothesis that the probabilities of substitutions in the 3 trinucleotide sites at the initial past time $t = 0$, are equiprobable, i.e.,

$$p(0) = q(0) = r(0) = \tfrac{1}{3}.$$

The function used here for varying the 3 parameters $p(t)$, $q(t)$ and $r(t)$ such as $p(t) + q(t) + r(t) = 1$ whatever t, is e^{-t} as its convergence to a limit when $t \to \infty$ allows it to remain in a probability space.

The first class of models studied has a parameter constant to $1/3$, a parameter varying according to the function $f(t)$ which exponentially decreases from $1/3$ to 0 and a parameter varying according to the function $g(t)$ which exponentially increases from $1/3$ to $2/3$, i.e.,

$$f(t) = \frac{e^{-t}}{3},$$

$$g(t) = \frac{2}{3} - \frac{e^{-t}}{3}.$$

Therefore, 6 evolutionary models testing all the possible combinations in the 3 trinucleotide sites, can be defined

	$p(t)$	$q(t)$	$r(t)$
Model 1	$f(t)$	$g(t)$	$1/3$
Model 2	$g(t)$	$f(t)$	$1/3$
Model 3	$f(t)$	$1/3$	$g(t)$
Model 4	$g(t)$	$1/3$	$f(t)$
Model 5	$1/3$	$f(t)$	$g(t)$
Model 6	$1/3$	$g(t)$	$f(t)$

REMARK 3. The model $p(t) = q(t) = r(t) = 1/3$ is a particular case of the constant model for which no solution has been found (data not shown).

The second class of models has 2 parameters varying according to the same decreasing function $f(t)$ and a parameter varying according to the function $h(t)$ which exponentially increases from $1/3$ to 1

$$h(t) = 1 - \frac{2e^{-t}}{3}.$$

Therefore, there are 3 evolutionary models testing all the possible combinations in the 3 sites

	$p(t)$	$q(t)$	$r(t)$
Model 7	$f(t)$	$f(t)$	$h(t)$
Model 8	$f(t)$	$h(t)$	$f(t)$
Model 9	$h(t)$	$f(t)$	$f(t)$

The third class of models has a parameter varying according to the decreasing function $f(t)$ and 2 parameters varying according to the same function $h(t)/2$ which exponentially increases from $1/3$ to $1/2$. Therefore, 3 evolutionary models can be deduced

	$p(t)$	$q(t)$	$r(t)$
Model 10	$f(t)$	$h(t)/2$	$h(t)/2$
Model 11	$h(t)/2$	$f(t)$	$h(t)/2$
Model 12	$h(t)/2$	$h(t)/2$	$f(t)$

3. RESULTS

The occurrence probabilities $\mathcal{P}(X_j, f, t)$ of the 3 circular codes X_0, X_1 and X_2 in the 3 frames of genes at the time t are computed in the 12 models according to formulae (2.4). Among these 12 models, only model 7 with the evolutionary parameters $p(t) = f(t)$, $q(t) = f(t)$ and $r(t) = h(t)$ represented in Fig. 1, leads to a solution (Figs. 2–4). Indeed, after uncertain evolutionary time, it retrieves the actual probability inequalities (1.1) in the 3 frames simultaneously.

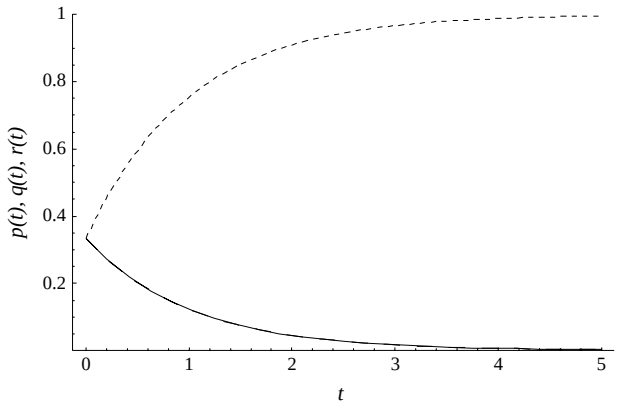

Figure 1. Probabilities of the 3 evolutionary parameters $p(t) = q(t) = e^{-t}/3$ (full line) and $r(t) = 1 - 2e^{-t}/3$ (dash line) under the mutation process t, $0 \leq t \leq 5$, of model 7.

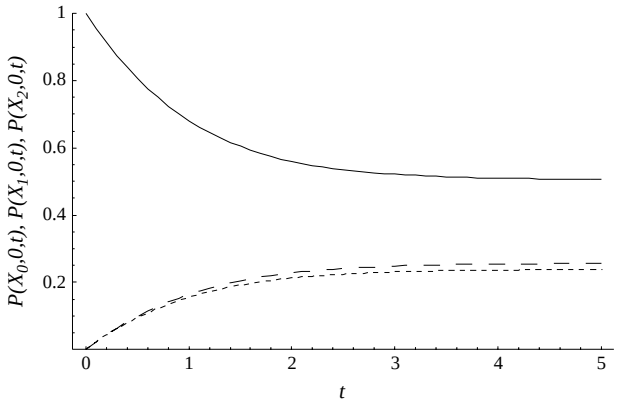

Figure 2. Probabilities $\mathcal{P}(X_j, 0, t)$ of the 3 circular codes X_0 (full line), X_1 (dash line) and X_2 (dot line) in the frame 0 of genes under the mutation process t, $0 \leq t \leq 5$, of model 7.

At the construction process ($t = 0$), model 7 has, as expected, the past probability inequalities (1.2) in the 3 frames of genes (Figs. 2–4) and the actual probability inequalities (1.1) in frame 1 (Fig. 3).

Unexpectedly, the substitution process ($t > 0$) in model 7 allows to retrieve the 2 other actual probability inequalities (1.1) in frames 0 and 2. Indeed, the inequalities (1.1) in frame 0 are verified for $t > 0$ and the difference $\mathcal{P}(X_1, 0, t) - \mathcal{P}(X_2, 0, t)$ increases during evolution (Fig. 2).

The inequalities (1.1) in frame 1 are observed at $t = 0$ and $t > 0$, i.e., for $t \geq 0$ (Fig. 3). Evolution also increases the difference $\mathcal{P}(X_2, 1, t) - \mathcal{P}(X_0, 1, t)$ (Fig. 3).

The variations of the probability curves in frame 2 are totally unexpected. Indeed, the inequalities (1.1) in frame 2 are verified only after a certain evolutionary time as the curve $\mathcal{P}(X_0, 2, t)$ starting with values lower than the curve $\mathcal{P}(X_1, 2, t)$, crosses it at $t \approx 0.6$ and remains higher through evolution (Fig. 4).

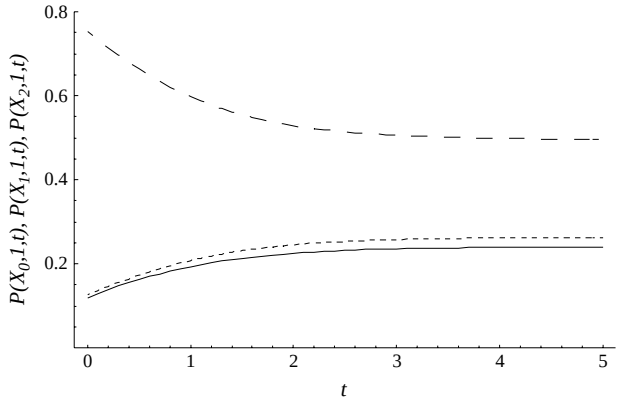

Figure 3. Probabilities $\mathcal{P}(X_j, 1, t)$ of the 3 circular codes X_0 (full line), X_1 (dash line) and X_2 (dot line) in the frame 1 of genes under the mutation process t, $0 \leq t \leq 5$, of model 7.

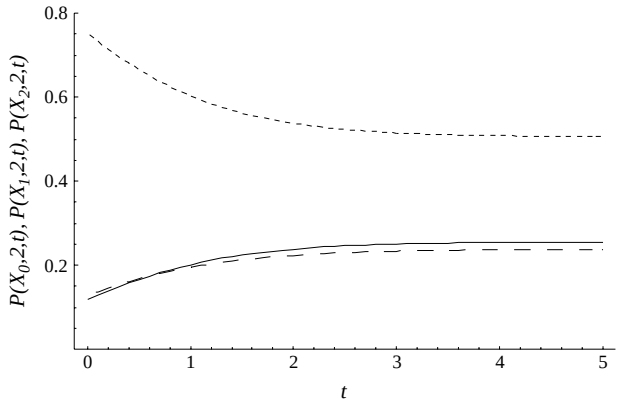

Figure 4. Probabilities $\mathcal{P}(X_j, 2, t)$ of the 3 circular codes X_0 (full line), X_1 (dash line) and X_2 (dot line) in the frame 2 of genes under the mutation process t, $0 \leq t \leq 5$, of model 7. The curve $\mathcal{P}(X_0, 2, t)$ crosses the curve $\mathcal{P}(X_1, 2, t)$ at $t \approx 0.6$.

In summary, the random substitution process generates the 2 inequalities $\mathcal{P}(X_1, 0, t) > \mathcal{P}(X_2, 0, t)$ and $\mathcal{P}(X_0, 2, t) > \mathcal{P}(X_1, 2, t)$ in frames 0 and 2 respectively of genes and increases progressively the amplitude between the 2 lower probability curves until $t \approx 5$.

The 11 other models tested cannot retrieve the actual probability inequalities (1.1) in the 3 frames simultaneously. For each of these 11 models, the properties of the probability curves in contradiction with the circular code observed in actual genes, are briefly given:

- Model 1: $\mathcal{P}(X_1, 0, t) \approx \mathcal{P}(X_2, 0, t)\ \forall t$, $\mathcal{P}(X_0, 1, t) \approx \mathcal{P}(X_2, 1, t)\ \forall t$ and $\mathcal{P}(X_0, 2, t) \approx \mathcal{P}(X_1, 2, t)\ \forall t$.
- Model 2: similar to model 1.

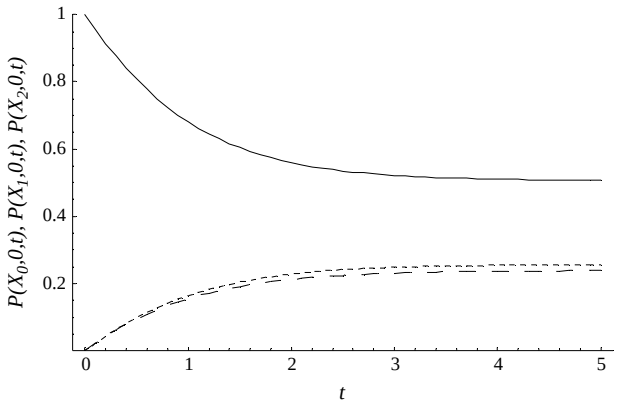

Figure 5. Probabilities $\mathcal{P}(X_j, 0, t)$ of the 3 circular codes X_0 (full line), X_1 (dash line) and X_2 (dot line) in the frame 0 of genes under the mutation process t, $0 \leq t \leq 5$, of model 9. This evolutionary model 9 cannot generate the actual probability inequality (1.1) $\mathcal{P}(X_1, 0, t) > \mathcal{P}(X_2, 0, t)$ associated with the circular code X_0 in genes.

- Model 3: $\mathcal{P}(X_1, 0, t) \gtrapprox \mathcal{P}(X_2, 0, t)\ \forall t$, $\mathcal{P}(X_0, 1, t) \lessapprox \mathcal{P}(X_2, 1, t)\ \forall t$ and $\mathcal{P}(X_0, 2, t) \gtrapprox \mathcal{P}(X_1, 2, t)\ t \to \infty$. Model 3 has the properties of model 7 but with weak amplitudes for the 2 lower curves in the 3 frames.
- Model 4: $\mathcal{P}(X_1, 0, t) \lessapprox \mathcal{P}(X_2, 0, t)\ \forall t$, $\mathcal{P}(X_0, 1, t) \gtrapprox \mathcal{P}(X_2, 1, t)\ t \to \infty$ and $\mathcal{P}(X_0, 2, t) \lessapprox \mathcal{P}(X_1, 2, t)\ \forall t$.
- Model 5: similar to model 1.
- Model 6: similar to model 1.
- Model 8: $\mathcal{P}(X_1, 0, t) = \mathcal{P}(X_2, 0, t)\ \forall t$, $\mathcal{P}(X_0, 1, t) = \mathcal{P}(X_2, 1, t)\ t \to \infty$ and $\mathcal{P}(X_0, 2, t) = \mathcal{P}(X_1, 2, t)\ t \to \infty$.
- Model 9: $\mathcal{P}(X_1, 0, t) < \mathcal{P}(X_2, 0, t)\ \forall t$ (Fig. 5), $\mathcal{P}(X_0, 1, t) > \mathcal{P}(X_2, 1, t)$ $t \to \infty$ and $\mathcal{P}(X_0, 2, t) < \mathcal{P}(X_1, 2, t)\ t \to \infty$.
- Model 10: similar to model 3.
- Model 11: similar to model 8.
- Model 12: similar to model 4.

Fig. 5 gives an example of an evolutionary model, here model 9, generating a probability inequality which is not observed in the actual genes.

4. DISCUSSION

The evolutionary model 7 retrieves the circular codes X_0, X_1 and X_2 observed in the 3 frames of genes and their main statistical properties. Its biological meaning would suggest that the primitive genes, i.e., the genes before substitutions ($t = 0$), are constructed by trinucleotides. Only 20 among 64 trinucleotides would have been necessary. The 20 types of trinucleotides as well as the type of their concatenation are determined in this model. Indeed, the 20 trinucleotides are defined

by the subset X_0 which is a C^3 code (Section 1.2). Furthermore, the independent concatenation of these 20 trinucleotides with equiprobability is the simplest type of concatenation and therefore, compatible with a primitive stage of gene evolution. A Markov concatenation of trinucleotides (based on a matrix of probabilities) would have been too complex at this primitive time. Model 7 also demonstrates that a substitution process ($t > 0$) must follow the construction process for obtaining a correlation with the actual genes. This substitution process allows the generation of new and totally unexpected probability inequalities, in particular $\mathcal{P}(X_1, 0, t) > \mathcal{P}(X_2, 0, t)$ and $\mathcal{P}(X_0, 2, t) > \mathcal{P}(X_1, 2, t)$ in the frames 0 and 2 respectively of genes. Furthermore, by decreasing the initial probabilities $\mathcal{P}(X_0, 0, t)$, $\mathcal{P}(X_1, 1, t)$ and $\mathcal{P}(X_2, 2, t)$ of the 3 circular codes X_0, X_1 and X_2 in the 3 frames 0, 1 and 2 respectively, it retrieves the frequency orders of X_0, X_1 and X_2 in each of the 3 frames of actual genes.

The negative results with models 1 and 2 suggest that the mutation rate $r(t)$ in the 3rd (trinucleotide) site cannot be constant $(1/3)$ during gene evolution, whatever the variations of the mutation rates $p(t)$ and $q(t)$ in the 1st and 2nd sites respectively [$p(t)$ increasing and $q(t)$ decreasing, or $p(t)$ decreasing and $q(t)$ increasing]. A similar conclusion can be deduced from the negative results with models 5 and 6 which would indicate that $p(t)$ cannot be constant $(1/3)$ during gene evolution, whatever the variations of $q(t)$ and $r(t)$. The partial positive results with models 3 and 10 and the negative results with models 4 and 11 would propose that $r(t)$ has an opposite variation compared to $p(t)$, and precisely, that $r(t)$ increases during gene evolution while $p(t)$ decreases (see also the negative results of model 12). The unique solution obtained with model 7 in the class of models 7, 8 and 9 confirms the previous results. Model 7 which leads to results more significant than those of model 3, allows to deduce an interesting property with the variations of the mutation rates during gene evolution:

In model 7, from an initial equiprobable mutation rate, the mutation rate $r(t)$ in the 3rd site increases during gene evolution while the mutation rates $p(t)$ and $q(t)$ in the 1st and 2nd sites decrease. This property agrees with the actual degeneracy of the genetic code with the highest mutation rate in the 3rd site [see e.g., Ermolaeva (2001)].

The complex behaviour of the curves giving the trinucleotide probabilities after time dependent mutations, is totally unexpected and implies 2 remarks. It is impossible to predict the probability variations of the trinucleotides after random substitutions without modelling. Therefore, the identification of rules explaining why only model 7 leads to a solution, is very difficult. On the other hand, the traces of the primitive probability differences between the trinucleotides, are conserved in the modelled genes even after a great number of substitutions, e.g., at $t = 5$ in the Figs. 2–4.

We are currently investigating exponential functions of the form e^{-kt} in order to increase the correlation between model 7 and the gene reality. Other classes of functions, such as sinusoidal, will also be tested. Furthermore, this approach can

easily be generalized to study evolution of motifs of various lengths, e.g., dicodons, etc., with time dependent mutations. Therefore, it could also be applied to the phylogenetic tree reconstruction and the sequence alignment.

ACKNOWLEDGEMENT

We thank the referee for his advice.

REFERENCES

Akashi, H. and A. Eyre-Walker (1998). Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8**, 688–693.

Arquès, D. G., J.-P. Fallot and C. J. Michel (1998). An evolutionary analytical model of a complementary circular code simulating the protein coding genes, the $5'$ and $3'$ regions. *Bull. Math. Biol.* **60**, 163–194.

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

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

Berg, O. G. and P. J. N. Silva (1997). Codon bias in Escherichia coli: the influence of codon context on mutation and selection. *Nucleic Acids Res.* **25**, 1397–1404.

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

Campbell, A., J. Mrázek and S. Karlin (1999). Genomic signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **96**, 9184–9189.

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

Ermolaeva, M. D. (2001). Synonymous codon usage in bacteria. *Curr. Issues Mol. Biol.* **3**, 91–97.

Grantham, R., C. Gautier, M. Gouy, R. Mercier and A. Pave (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8**, r49–r62.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 12–34.

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

Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules, in *Mammalian Protein Metabolism*, H. N. Munro (Ed.), New York: Academic Press, pp. 21–132.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* **78**, 454–458.

Koch, A. J. and J. Lehmann (1997). About a symmetry of the genetic code. *J. Theor. Biol.* **189**, 171–174.

Konu, O. and M. D. Li (2002). Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents. *J. Mol. Evol.* **54**, 35–41.

Kozak, M. (2002). Pushing the limits of the scanning mechanism for initiation of translation. *Gene* **299**, 1–34.

Lacan, J. and C. J. Michel (2001). Analysis of a circular code model. *J. Theor. Biol.* **213**, 159–170.

Sharp, P. M. and G. Matassi (1994). Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* **4**, 851–860.

Shpaer, E. G. (1986). Constraints on codon context in Escherichia coli genes. Their possible role in modulating the efficiency of translation. *J. Mol. Biol.* **188**, 555–564.

Received 5 June 2003 and accepted 20 October 2003