Bulletin of Mathematical Biology (2007) 69: 677–698
DOI 10.1007/s11538-006-9147-z

ORIGINAL ARTICLE

An Analytical Model of Gene Evolution with 9 Mutation Parameters: An Application to the Amino Acids Coded by the Common Circular Code

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

Received: 17 January 2006 / Accepted: 31 May 2006 / Published online: 2 September 2006

Abstract We develop here an analytical evolutionary model based on a trinucleotide mutation matrix 64×64 with nine substitution parameters associated with the three types of substitutions in the three trinucleotide sites. It generalizes the previous models based on the nucleotide mutation matrices 4×4 and the trinucleotide mutation matrix 64×64 with three and six parameters. It determines at some time t the exact occurrence probabilities of trinucleotides mutating randomly according to these nine substitution parameters. An application of this model allows an evolutionary study of the common circular code $\mathcal{C}$ of eukaryotes and prokaryotes and its 12 coded amino acids. The main property of this code $\mathcal{C}$ is the retrieval of the reading frames in genes, both locally, i.e. anywhere in genes and in particular without a start codon, and automatically with a window of a few nucleotides. However, since its identification in 1996, amino acid information coded by $\mathcal{C}$ has never been studied. Very unexpectedly, this evolutionary model demonstrates that random substitutions in this code $\mathcal{C}$ and with particular values for the nine substitution parameters retrieve after a certain time of evolution a frequency distribution of these 12 amino acids very close to the one coded by the actual genes.

Keywords Analytical model · Parameter · Evolution · Mutation · Circular code · Gene · Amino acid

1. Introduction

1.1. Presentation of the approach

Each genome has its own trinucleotide distribution (Grantham et al., 1980). Indeed, the synonymous codons (codons coding for the same amino acids) do

E-mail address: michel@dpt-info.u-strasbg.fr.

not occur with the same frequencies in genes. This synonymous codon usage is biased: a restricted subset of codons is preferred in genes. Codon usage is generally correlated with gene expressivity (Grantham et al., 1981; Ikemura, 1985; Sharp and Matassi, 1994) even if its strength varies among bacterial species (Sharp et al., 2005). A proposed explanation is that codon usage reflects the variation in the concentration of tRNAs. Major codons encoded by more abundant tRNAs should increase translational efficacy (Bulmer, 1991; Akashi and Eyre-Walker, 1998). Nevertheless, tRNA abundance could also have evolved for matching codon pattern in a genome (Fedorov et al., 2002) and then would rather be a consequence of the synonymous codon bias.

Several other processes may influence codon usage (Jukes and Bhushan, 1986; Campbell et al., 1999; Llopart and Aguade, 2000; Smith and Eyre-Walker, 2001; Konu and Li, 2002; Krakauer and Jansen, 2002; Rogozin et al., 2005) (see also the review (Ermolaeva, 2001)). In particular, codon choice may depend on its context, i.e. the surrounding nucleotides (Yarus and Folley, 1984; Shpaer, 1986; Berg and Silva, 1997). These pressures might be frame independent (Antezana and Kreitman, 1999). In this line of research, we have studied the occurrences of the 64 trinucleotides $\mathcal{T} = \{AAA, \ldots, TTT\}$ in the three frames of genes by computing their frequencies. This approach has led to the identification of a particular code in genes called circular code.

By convention, the reading frame established by a start codon belonging to $\mathcal{T}_{\text{start}} = \{ATG, GTG, TTG\}$, is the frame 0, and the frames 1 and 2 are the reading frame shifted by 1 and 2 nucleotides in the 5′–3′ direction, respectively. After excluding the trinucleotides with identical nucleotides $\mathcal{T}_{\text{id}} = \{AAA, CCC, GGG, TTT\}$ and by assigning each trinucleotide to a preferential frame, three subsets of 20 trinucleotides per frame have been identified statistically in the gene populations of both eukaryotes and prokaryotes (Arquès and Michel, 1996). These three trinucleotide sets $\mathcal{C}_0$, $\mathcal{C}_1$ and $\mathcal{C}_2$ associated with the frames 0, 1 and 2, respectively, have several strong properties, in particular the property of circular code. The circular code concept will be briefly pointed out without mathematical notations after a short historical presentation of an another class of code which has been searched but not found in genes (over the alphabet $\{A, C, G, T\}$).

A code in genes has been proposed by Crick et al. (1957) in order to explain how the reading of a series of nucleotides could code for the amino acids constituting the proteins. The two problems stressed were: why are there more trinucleotides than amino acids and how to choose the reading frames? Crick et al. (1957) have then proposed that only 20 among 64 trinucleotides code for the 20 amino acids. Furthermore, such a bijective code implies that the coding trinucleotides are found only in one frame. Such a particular code is called a comma-free code or a code without commas. However, the determination of a set of 20 trinucleotides forming a comma-free code has several constraints:

(i) A trinucleotide $\mathcal{T}_{\text{id}}$ must be excluded from such a code. Indeed, the concatenation of AAA with itself, for example, does not allow the reading (original) frame to be retrieved as there are three possible decompositions: $\ldots AAA, AAA, AAA, \ldots$, $\ldots A, AAA, AAA, AA\ldots$ and $\ldots AA, AAA, AAA, A\ldots$, the commas showing the way of construction (decomposition).

(ii) Two trinucleotides related to circular permutation, for example AAC and ACA, must be also excluded from such a code. Indeed, the concatenation of AAC with itself, for example, also does not allow the reading frame to be retrieved as there are two possible decompositions: $\ldots AAC, AAC, AAC, \ldots$ and $\ldots A, ACA, ACA, AC \ldots$

Therefore, by excluding the four trinucleotides $\mathcal{T}_{\mathrm{id}}$ and by gathering the 60 remaining trinucleotides in 20 classes of three trinucleotides such that, in each class, three trinucleotides are deduced from each other by circular permutations, e.g. AAC, ACA and CAA, a comma-free code has only one trinucleotide per class and therefore contains at most 20 trinucleotides. This trinucleotide number is identical to the amino acid one, thus leading to a comma-free code assigning one trinucleotide per amino acid without ambiguity.

The determination of comma-free codes and their properties are unrealizable without computer as there are $3^{20} \approx 3.5$ billions potential codes. A comma-free code search algorithm demonstrates in particular that there are only 408 comma-free codes of 20 trinucleotides. None of them is self-complementary (see also the property (iii) in Section 1.2.2) as the maximal complementary comma-free codes contain only 16 trinucleotides (results not shown). Furthermore, in the late fifties, the two discoveries that the trinucleotide TTT, an excluded trinucleotide in a comma-free code, codes for phenylalanine (Nirenberg and Matthaei, 1961) and that genes are placed in reading frames with a start trinucleotide $\mathcal{T}_{\mathrm{start}}$, have led to give up the concept of comma-free code over the alphabet $\{A, C, G, T\}$. For several biological reasons, in particular the interaction between mRNA and tRNA, this concept is taken again later over the purine/pyrimidine alphabet $\{R, Y\}$ ($R = \{A,G\}$, $Y = \{C,T\}$) with two comma-free codes for primitive genes: RRY (Crick et al., 1976) and RNY ($N = \{R, Y\}$) (Eigen and Schuster, 1978).

A circular code also allows the reading frames of genes to be retrieved but with weaker conditions compared to a comma-free code. It is a set of words over an alphabet such that any word written on a circle (the next letter after the last letter of the word being the first letter) has at most one decomposition (factorization) into words of the circular code. As an example, let the set X be composed of the six following words: $X = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ and the word w, be a series of the nine following letters: $w = ATGGCCCTA$. The word w, written on a circle, can be factorized into words of X according to two different ways: ATG, GCC, CTA and AAT, GGC, CCT. Therefore, X is not a circular code. In contrast, if the set Y obtained by replacing the word GGC of X by GTC is considered, i.e. $Y = \{AAT, ATG, CCT, CTA, GCC, GTC\}$, then there never exists an ambiguous word with Y, in particular w is not ambiguous, and Y is a circular code. The construction frame of a word generated by any concatenation of words of a circular code can be retrieved after the reading, anywhere in the generated word, of a certain number of nucleotides depending on the code. This series of nucleotides is called the window W of the circular code. Therefore, a circular code has the ability to retrieve the reading frames in genes, both locally, i.e. anywhere in genes and in particular without a start codon, and automatically with a window of a few nucleotides. In genes, the circular code information for retrieving the reading frames is added to the classical genetic code for coding the amino acids. Such an

Table 1 The 12 amino acids $\mathcal{A}$ coded by the trinucleotides $\mathcal{C}$ of the common circular code and by the trinucleotides $\mathcal{T}_{\mathcal{A}}\backslash\mathcal{C}$ which do not belong to the common circular code.

	Trinucleotides $\mathcal{C}$		Trinucleotides $\mathcal{T}_{\mathcal{A}}\backslash\mathcal{C}$	
	Number	Type	Number	Type
Ala	1	*GCC*	3	*GCA*, *GCG*, *GCT*
Asn	2	*AAC*, *AAT*	0	
Asp	2	*GAC*, *GAT*	0	
Gln	1	*CAG*	1	*CAA*
Glu	2	*GAA*, *GAG*	0	
Gly	2	*GGC*, *GGT*	2	*GGA*, *GGG*
Ile	2	*ATC*, *ATT*	1	*ATA*
Leu	2	*CTC*, *CTG*	4	*CTA*, *CTT*, *TTA*, *TTG*
Phe	1	*TTC*	1	*TTT*
Thr	1	*ACC*	3	*ACA*, *ACG*, *ACT*
Tyr	1	*TAC*	1	*TAT*
Val	3	*GTA*, *GTC*, *GTT*	1	*GTG*

important property might be involved in the transcription and translation apparatus of primitive genes (Arquès and Michel, 1996).

A comma-free code has conditions stronger than a circular code. Indeed, the 20 trinucleotides of a comma-free code are found only in one frame, i.e. in the reading frame, while some trinucleotides of a circular code can be found in the two shifted frames 1 and 2 (property (vib) in Section 1.2.2). On the other hand, the lengths of the windows W for retrieving the reading frames of a comma-free code and a circular code are less than or equal to 4 and 13 nucleotides, respectively.

The common circular code $\mathcal{C} = \mathcal{C}_0$ of 20 trinucleotides identified in the reading frames (frames 0) of genes belonging to two large and different populations of eukaryotes (26757 genes, 11397678 trinucleotides) and prokaryotes (13686 genes, 4708758 trinucleotides) is $\mathcal{C} = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$ (Arquès and Michel, 1996). It codes for the 12 amino acids $\mathcal{A}$={*Ala*, *Asn*, *Asp*, *Gln*, *Glu*, *Gly*, *Ile*, *Leu*, *Phe*, *Thr*, *Tyr*, *Val*} according to the (standard) genetic code (Table 1).

Five amino acids (AA) *Ala*, *Gln*, *Phe*, *Thr* and *Tyr* are coded by one trinucleotide of the code $\mathcal{C}$, six AA *Asn*, *Asp*, *Glu*, *Gly*, *Ile* and *Leu*, by two trinucleotides of $\mathcal{C}$, and one AA *Val*, by three trinucleotides of $\mathcal{C}$ (Table 1). Some biological properties of these 12 AA associated with the code $\mathcal{C}$ have been given in Arquès and Michel (1996); Koch and Lehmann (1997). In actual genes, these 12 AA $\mathcal{A}$ are coded by a set $\mathcal{T}_{\mathcal{A}}$ of 37 trinucleotides. Let $\mathcal{T}_{\mathcal{A}}\backslash\mathcal{C}$ be the set of trinucleotides which belong to $\mathcal{T}_{\mathcal{A}}$ but not to $\mathcal{C}$. The set $\mathcal{T}_{\mathcal{A}}\backslash\mathcal{C}$ has 17 trinucleotides: $\mathcal{T}_{\mathcal{A}}\backslash\mathcal{C} = \{ACA, ACG, ACT, ATA, CAA, CTA, CTT, GCA, GCG, GCT, GGA, GGG, GTG, TAT, TTA, TTG, TTT\}$. It codes for nine among 12 AA: five AA *Gln*, *Ile*, *Phe*, *Tyr* and *Val* are coded by one trinucleotide of $\mathcal{T}_{\mathcal{A}}\backslash\mathcal{C}$, one AA *Gly*, by two trinucleotides of $\mathcal{T}_{\mathcal{A}}\backslash\mathcal{C}$, two AA *Ala* and *Thr*, by three trinucleotides of $\mathcal{T}_{\mathcal{A}}\backslash\mathcal{C}$, and one AA *Leu*, by four trinucleotides of $\mathcal{T}_{\mathcal{A}}\backslash\mathcal{C}$ (Table 1).

The observation of a preferential trinucleotide set $\mathcal{C}$ in various genes from the two largest domains, the eukaryotes and the prokaryotes, is the basis of our

development of an evolutionary model. Indeed, if such a "universal" set occurs with a frequency higher than the random one in actual genes after (mainly) random mutations, then a realistic hypothesis consists in asserting that this set had a frequency in past higher than in actual time. In other words, the trinucleotides of $\mathcal{C}$ are the basic words of "primitive" genes (genes before evolution). As the "primitive" genes will be constructed by trinucleotides of $\mathcal{C}$ and as the amino acids are coded by trinucleotides, the mathematical model will be based on a trinucleotide mutation matrix 64×64. The evolutionary model proposed will be based on two processes: a construction process with a random mixing of 20 trinucleotides of $\mathcal{C}$ with equiprobability (1/20) followed by an evolutionary process with random substitutions which are modelled by nine parameters a, b, c, d, e, f, g, h and k associated with the three types of substitutions in the three trinucleotide sites: a, d and g are the rates of transitions $\mathrm{A} \longleftrightarrow \mathrm{G}$ (a substitution from one purine $\{A, G\}$ to the other) and $\mathrm{C} \longleftrightarrow \mathrm{T}$ (a substitution from one pyrimidine $\{C, T\}$ to the other) in the three sites, respectively; b, e and h are the rates of transversions (a substitution from a purine to a pyrimidine, or reciprocally) $\mathrm{A} \longleftrightarrow \mathrm{T}$ and $\mathrm{C} \longleftrightarrow \mathrm{G}$ in the three sites, respectively; and c, f and k are the rates of transversions $\mathrm{A} \longleftrightarrow \mathrm{C}$ and $\mathrm{G} \longleftrightarrow \mathrm{T}$ in the three sites, respectively.

Two types of results are presented in this paper: a development of a mathematical model and its application to an amino acid evolution. The stochastic evolutionary model will determine at some time t the occurrence probabilities of trinucleotides mutating randomly according to these nine substitution parameters in order to derive the evolutionary analytical solutions of the common circular code $\mathcal{C}$ and the 12 amino acids $\mathcal{A}$. Therefore, it will generalize the previous models, in particular the nucleotide mutation matrices 4×4 at one substitution parameter (Jukes and Cantor, 1969), two parameters (transitions and transversions) (Kimura, 1980) and the trinucleotide mutation matrix 64×64 with three and six substitution parameters (Arquès et al., 1998; Frey and Michel, 2006). Since the identification of the common circular code $\mathcal{C}$ in 1996, evolution of its 12 coded amino acids $\mathcal{A}$ has never been investigated. An application of this evolutionary model will show that random substitutions in the trinucleotides of the code $\mathcal{C}$ and with particular values for the nine substitution parameters, will generate after a certain time of evolution other trinucleotides in the reading frames of genes in an unbalanced way, extend the capacity of coding the 12 amino acids $\mathcal{A}$, from 20 trinucleotides ($\mathcal{C}$) to 37 trinucleotides ($\mathcal{T}_{\mathcal{A}}$), and retrieve an amino acid frequency distribution very close to the one coded by the actual genes.

In the next two sections 1.2 and 1.3, the two stages of our approach are briefly detailed: the observation of a common circular code in eukaryotic and prokaryotic genes and the two processes of the evolutionary model.

1.2. A common circular code in eukaryotic and prokaryotic genes

1.2.1. Definition

Notation. $\mathbb{A}$ being a finite alphabet, $\mathbb{A}^*$ denotes the words over $\mathbb{A}$ of finite length including the empty word of length 0 and $\mathbb{A}^+$, the words over $\mathbb{A}$ of finite length greater or equal to 1. Let $w_1 w_2$ be the concatenation of the two 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 $sx_2 \ldots x_n r = y_1 y_2 \ldots y_m$ and $x_1 = rs$ imply $n = m$, $r = \epsilon$ (empty word) and $x_i = y_i$, $1 \leq i \leq n$ (Berstel and Perrin, 1985; Béal, 1993).

1.2.2. Properties of the common circular code $\mathcal{C}$

Definition 2. The (left circular) permutation $\mathcal{P}$ of a trinucleotide $w_0 = l_0 l_1 l_2$, $l_0 l_1 l_2 \in \mathcal{T}$, is the permuted trinucleotide $\mathcal{P}(w_0) = w_1 = l_1 l_2 l_0$, e.g. $\mathcal{P}(AAC) = ACA$, and $\mathcal{P}(\mathcal{P}(w_0)) = \mathcal{P}(w_1) = w_2 = l_2 l_0 l_1$, e.g. $\mathcal{P}(\mathcal{P}(AAC)) = CAA$. This definition is naturally extended to the trinucleotide set permutation: The permutation $\mathcal{P}$ of a set of trinucleotides is the permuted trinucleotide set obtained by the permutation $\mathcal{P}$ of all its trinucleotides.

The properties of the common circular code $\mathcal{C} = \mathcal{C}_0$ identified in frames 0 of eukaryotic and prokaryotic genes are briefly recalled (details can be found in Arquès and Michel (1996); Lacan and Michel (2001)):

(i) Maximality: $\mathcal{C}$ is a maximal circular code, i.e. with 20 trinucleotides, as it is not contained in a larger circular code, i.e. in a circular code with more words. For words of length 3 over a 4-letter alphabet, a circular code has at most 20 words. Then, any 20-long circular code is maximal.

(ii) Permutation: $\mathcal{C}$ generates $\mathcal{C}_1$ by one permutation and $\mathcal{C}_2$ by another permutation, i.e. $\mathcal{P}(\mathcal{C}) = \mathcal{C}_1$ and $\mathcal{P}(\mathcal{P}(\mathcal{C})) = \mathcal{C}_2$.

(iii) Complementarity: $\mathcal{C}$ is self-complementary (10 trinucleotides of $\mathcal{C}$ are complementary to 10 other trinucleotides of $\mathcal{C}$) and, $\mathcal{C}_1$ and $\mathcal{C}_2$ are complementary to each other (the 20 trinucleotides of $\mathcal{C}_1$ are complementary to the 20 trinucleotides of $\mathcal{C}_2$).

(iv) $\mathcal{C}^3$ code: $\mathcal{C}_1$ and $\mathcal{C}_2$ obtained by permutation of $\mathcal{C}$ (property ii) are maximal circular codes. Therefore, if $\mathcal{C}$, $\mathcal{C}_1$ and $\mathcal{C}_2$ are circular codes, then $\mathcal{C}$, $\mathcal{C}_1$ and $\mathcal{C}_2$ are $\mathcal{C}^3$ codes. As the circular code $\mathcal{C}$ is associated with the reading frame (frame 0) in genes, i.e. the most important frame with a biological function, it is considered as the main $\mathcal{C}^3$ code. It is important to stress that a circular code X_0 does not necessarily imply that X_1 and X_2 obtained by its permutations, are also circular codes, i.e. a circular code is not necessarily a $\mathcal{C}^3$ code.

(v) Rarity: the occurrence probability of the $\mathcal{C}^3$ code $\mathcal{C}$ is equal to $216/3^{20} \approx 6 \times 10^{-8}$, i.e. the computed number of complementary $\mathcal{C}^3$ codes (216) divided by the number of potential codes ($3^{20} = 3486784401$).

(vi) Flexibility:

(via) The lengths of the minimal windows of $\mathcal{C}$, $\mathcal{C}_1$ and $\mathcal{C}_2$ for retrieving automatically the frames 0, 1 and 2, respectively, are all equal to 13 nucleotides and represent the largest window length among the 216 $\mathcal{C}^3$ codes.

(vib) The frequencies of "misplaced" trinucleotides in the shifted frames 1 and 2 are both equal to 24.6%. If the trinucleotides of $\mathcal{C}$ are randomly concatenated, for example as follows:
$\ldots GAA, GAG, GTA, GTA, ACC, AAT, GTA, CTC, TAC, TTC, ACC,$

$ATC\ldots$ then, the trinucleotides in frame 1:
$\ldots G, AAG, AGG, TAG, TAA, CCA, ATG, TAC, TCT, ACT, TCA, CCA,$
$TC\ldots$ and the trinucleotides in frame 2:
$GA, AGA, GGT, AGT, AAC, CAA, TGT, ACT, CTA, CTT, CAC, CAT,$
$C\ldots$ mainly belong to $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. A few trinucleotides are misplaced in the shifted frames. With this example, in frame 1, nine trinucleotides belong to $\mathcal{C}_1$, one trinucleotide (TAC) to $\mathcal{C}$ and one trinucleotide ($TAA = \mathcal{P}(\mathcal{P}(AAT))$) to $\mathcal{C}_2$. In frame 2, eight trinucleotides belong to $\mathcal{C}_2$, two trinucleotides (GGT, AAC) to $\mathcal{C}$ and one trinucleotide ($ACT = \mathcal{P}(TAC)$) to $\mathcal{C}_1$. By computing exactly, the frequencies of misplaced trinucleotides in frame 1 are 11.9% for $\mathcal{C}$ and 12.7% for $\mathcal{C}_2$. In frame 2, the frequencies of misplaced trinucleotides are 11.9% for $\mathcal{C}$ and 12.7% for $\mathcal{C}_1$. The complementarity property (iii) explains on the one hand, the identical frequencies of $\mathcal{C}$ in frames 1 and 2, and on the other hand, the identical frequencies of $\mathcal{C}_2$ in frame 1 and $\mathcal{C}_1$ in frame 2. Then, the frequency sum of misplaced trinucleotides in frame 1 ($\mathcal{C}$ and $\mathcal{C}_2$) is equal to the one of misplaced trinucleotides in frame 2 ($\mathcal{C}$ and $\mathcal{C}_1$) and is equal to 24.6%. This value is close to the highest frequency (27.9%) of misplaced trinucleotides among the 216 $\mathcal{C}^3$ codes. Note that misplaced trinucleotides are impossible with a comma-free code (Section 1.1).

(vic) The four types of nucleotides occur in the three trinucleotide sites of $\mathcal{C}$, and also obviously by the permutation property (ii) in those of $\mathcal{C}_1$ and $\mathcal{C}_2$. It is important to stress that $\mathcal{C}^3$ codes can have missing nucleotides in their trinucleotide sites.

(vii) Common: $\mathcal{C}$ has a "universal" distribution in eukaryotic and prokaryotic genes.

1.2.3. Mean occurrence probabilities of the 12 *amino acids* $\mathcal{A}$ *in prokaryotic genomes*

The mean occurrence probabilities of the 12 amino acids $\mathcal{A}$ coded by the common circular code $\mathcal{C}$ and the trinucleotides $\mathcal{T}_{\mathcal{A}}\backslash\mathcal{C}$, are computed in all (valid) genes of 175 complete prokaryotic genomes representing 487863 genes of 453749 kb (Table 2).

This amino acid distribution is (obviously) correlated with the number of codons coding for these amino acids. It is very similar to the one published in 1996 with only 9510 genes of 9132 kb (Table 5 in Arquès and Michel (1996)). Table 2 shows in particular that *Leu* occurs with the highest frequency, then it appears a group of three amino acids *Ala*, *Gly* and *Val*, and *Tyr* has the lowest frequency. These statistical features will constitute the principal constraints in the application of the stochastic model for studying evolution of the amino acids $\mathcal{A}$.

Table 2 Mean occurrence probabilities (in %) of the 12 amino acids $\mathcal{A}$ computed in all genes of 175 complete prokaryotic genomes.

Ala	*Asn*	*Asp*	*Gln*	*Glu*	*Gly*	*Ile*	*Leu*	*Phe*	*Thr*	*Tyr*	*Val*
9.4	3.9	5.3	3.9	6.1	7.4	6.3	10.3	4.1	5.4	3.0	7.1

1.3. An amino acid evolutionary model

Founded on the principle described in Introduction, the model is based on a construction process ($t = 0$) which generates "primitive" genes according to a random mixing of the 20 trinucleotides of the common circular code $\mathcal{C}$ with equiprobability (1/20). Then, an evolutionary process ($t > 0$) transforms these primitive genes into simulated actual ones. Random substitutions with different rates in the three sites of the 20 trinucleotides of the code $\mathcal{C}$ will generate other trinucleotides and distribute them according to an unbalanced way in the hope of retrieving the distribution of the 12 amino acids $\mathcal{A}$ coded by the actual genes. This problem is a priori hard as there are two effects: the initial amino acid probabilities which strongly differ from the actual ones (Tables 1 and 2) and the amino acid coding which is carried out only by the code $\mathcal{C}$ at the initial condition ($t = 0$) and by the genetic code ($\mathcal{C}$ and $\mathcal{T}_{\mathcal{A}}\backslash\mathcal{C}$) during the evolutionary process ($t > 0$). For example, *Val* coded by $\mathcal{C}$ at $t = 0$ (primitive genes) has the highest probability $3/20 = 15\%$ (Table 1) while *Leu* coded by the genetic code in actual time occurs with the highest frequency 10.3% (Table 2). In the same line, *Ala*, *Gln*, *Phe*, *Thr* and *Tyr* at $t = 0$ have the same lowest probability $1/20 = 5\%$ (Table 1) while *Ala* and *Tyr*, for example, have a completely different distribution in actual time, precisely the second highest frequency and the lowest one, respectively (Table 2).

The aim of this mathematical model consists in determining the analytical solutions of the occurrence probabilities of the common circular code $\mathcal{C}$ and the 12 amino acids $\mathcal{A}$ as a function of the evolutionary time t and the nine substitution parameters a, b, c, d, e, f, g, h and k (Section 2). It should be stressed that this stochastic approach with exact solutions relies on a gene evolutionary physical model by applying random substitutions in simulated sequences. However, in order to get computer results with a good approximation in such a physical model, a population of large sequences must be simulated in the statistical analysis, which is time consuming.

This evolutionary model will demonstrate here that the actual distribution of the 12 amino acids $\mathcal{A}$ can be simulated after a certain evolutionary time t of random substitutions in the common circular code $\mathcal{C}$ and with particular values for the nine substitution parameters.

2. Mathematical model

The mathematical model will determine at an evolutionary time t the occurrence probability $P(X, t)$ of a trinucleotide set X whose trinucleotides mutate according to nine real substitution parameters a, b, c, d, e, f, g, h and k: a, d and g are the transition rates A$\longleftrightarrow$G and C$\longleftrightarrow$T in the three sites, respectively; b, e and h are the transversion rates A$\longleftrightarrow$T and C$\longleftrightarrow$G in the three sites, respectively; and c, f and k are the transversion rates A$\longleftrightarrow$C and G$\longleftrightarrow$T in the three sites, respectively. The trinucleotide sets X studied are the common circular code $\mathcal{C}$ and the 12 trinucleotide sets coding the 12 amino acids $\mathcal{A}$. This model generalizes the previous mathematical models based on the nucleotide mutation matrices 4×4 (Jukes and Cantor, 1969; Kimura, 1980) and the trinucleotide mutation matrix 64×64 with three and six parameters (Arquès et al., 1998; Frey and Michel, 2006).

By convention, the indexes $i,\ j \in \{1, \ldots, 64\}$ represent the 64 trinucleotides $\mathcal{T}$ in alphabetical order. Let $P(j \to i)$ be the substitution probability of a trinucleotide j, $j \neq i$, into a trinucleotide i. The probability $P(j \to i)$ is equal to 0 if the substitution is impossible, i.e. if j and i differ more than one nucleotide as the time interval T is assumed to be enough small that a trinucleotide cannot mutate successively two times during T. Otherwise, it is given as a function of the nine substitution rates a, b, c, d, e, f, g, h and k. For example with the trinucleotide AAA associated with $i = 1$, $P(CAA \to AAA) = c$, $P(GAA \to AAA) = a$, $P(TAA \to AAA) = b$, $P(ACA \to AAA) = f$, $P(AGA \to AAA) = d$, $P(ATA \to AAA) = e$, $P(AAC \to AAA) = k$, $P(AAG \to AAA) = g$, $P(AAT \to AAA) = h$ and $P(j \to AAA) = 0$ with $j \notin \{AAC, AAG, AAT, ACA, AGA, ATA, CAA, GAA, TAA\}$. The substitution probability $P(i \to i)$ of a trinucleotide i into itself is equal to $P(i \to i) = (1 - \sum_{j=1,\ j\neq i}^{64} P(j \to i))$ in a stochastic approach.

Let $P_i(t)$ be the occurrence probability of a trinucleotide i at the time t. At time $t + T$, the occurrence probability of the trinucleotide i is $P_i(t + T)$ so that $P_i(t + T) - P_i(t)$ represents the probabilities of trinucleotides i which appear and disappear during the time interval T

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

where α is the probability that a trinucleotide is subjected to one substitution during T. With a suitable time interval, the probability α is equal to 1, i.e. there is one substitution per trinucleotide per time interval. Then,

$$\begin{aligned} P_i(t + T) - P_i(t) &= T \sum_{j=1}^{64} P(j \to i)\, P_j(t) - T P_i(t) \\ &= T \sum_{\substack{j=1 \\ j\neq i}}^{64} P(j \to i)\, P_j(t) + T P(i \to i)\, P_i(t) - T P_i(t) \\ &= T \sum_{\substack{j=1 \\ j\neq i}}^{64} P(j \to i)\, P_j(t) + T \left(1 - \sum_{\substack{j=1 \\ j\neq i}}^{64} P(j \to i)\right) P_i(t) \\ &\quad - T P_i(t). \end{aligned} \tag{1}$$

As the sum of the substitution probabilities $P(j \to i)$ of trinucleotides j into a trinucleotide i is equal to 1 in this stochastic approach, i.e. $a + b + c + d + e + f + g + h + k = 1$, then $P(i \to i) = 0$, i.e. the substitution probability of a trinucleotide i into itself is impossible, and the formula (1) leads to

$$\lim_{T\to 0} \frac{P_i(t + T) - P_i(t)}{T} = P_i'(t) = \sum_{j=1}^{64} P(j \to i)\, P_j(t) - P_i(t) \tag{2}$$

when $T \to 0$ and with $P(j \to i) = 0$ if $j = i$.

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

$$P'(t) = A \cdot P(t) - P(t) = (A - I) \cdot P(t) \tag{3}$$

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

The square mutation matrix $A\,(64, 64)$ can be defined by a square block matrix $(4, 4)$ whose four diagonal elements are formed by four identical square submatrices $B\ (16, 16)$ and whose 12 non-diagonal elements are formed by four square submatrices $aI\ (16, 16)$, four square submatrices bI (16,16) and four square submatrices $cI\ (16, 16)$ as follows

$$A = \left(\begin{array}{c|cccc} & 1\cdots16 & 17\cdots32 & 33\cdots48 & 49\cdots64 \\ \hline 1\cdots16 & B & cI & aI & bI \\ 17\cdots32 & cI & B & bI & aI \\ 33\cdots48 & aI & bI & B & cI \\ 49\cdots64 & bI & aI & cI & B \end{array}\right).$$

The index ranges $\{1, \ldots, 16\}$, $\{17, \ldots, 32\}$, $\{33, \ldots, 48\}$ and $\{49, \ldots, 64\}$ are associated with the trinucleotides $\{AAA, \ldots, ATT\}$, $\{CAA, \ldots, CTT\}$, $\{GAA, \ldots, GTT\}$ and $\{TAA, \ldots, TTT\}$, respectively. The square submatrix B $(16, 16)$ can again be defined by a square block matrix $(4, 4)$ whose four diagonal elements are formed by four identical square submatrices $C\ (4, 4)$ and whose 12 non-diagonal elements are formed by four square submatrices dI $(4, 4)$, four square submatrices $eI\ (4, 4)$ and four square submatrices $fI\ (4, 4)$ as follows

$$B = \begin{pmatrix} C & fI & dI & eI \\ fI & C & eI & dI \\ dI & eI & C & fI \\ eI & dI & fI & C \end{pmatrix}.$$

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

$$C = \begin{pmatrix} 0 & k & g & h \\ k & 0 & h & g \\ g & h & 0 & k \\ h & g & k & 0 \end{pmatrix}.$$

The differential equation (3) can then be written in the following form

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

with

$$M = A - I.$$

As the nine substitution parameters are real, the matrix A is real and also symmetrical by construction. Therefore, the matrix M is also real and symmetrical. There exists an eigenvector matrix Q and a diagonal matrix D of eigenvalues λ_k of M ordered in the same way as the eigenvector columns in Q such that $M = Q \cdot D \cdot Q^{-1}$. Then,

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

This backward equation has the classical solution (Lange, 2005)

$$P(t) = Q \cdot \mathrm{e}^{Dt} \cdot Q^{-1} \cdot P(0) \tag{4}$$

where e^{Dt} is the diagonal matrix of exponential eigenvalues $\mathrm{e}^{\lambda_k t}$.

The eigenvalues λ_k of M are deduced from the eigenvalues μ_k of A such that $\lambda_k = \mu_k - 1$. The eigenvalues μ_k of A can be obtained by determining the roots of the characteristic equation $\det(A - \mu I) = 0$ of A using its block matrix properties. Therefore, after linear combinations, the determinant $\det(A - \mu I)$ is equal to

$$\begin{aligned}\det(A - \mu I) = {} & \det\left(B - (a + b - c + \mu)\, I\right) \times \det\left(B - (a - b + c + \mu)\, I\right) \\ & \times \det\left(B - (-a + b + c + \mu)\, I\right) \times \det\left(B - (-a - b - c + \mu)\, I\right).\end{aligned} \tag{5}$$

As the matrix B has a block structure similar to the matrix A, the form of the determinant $\det(B - \nu I)$ can be easily deduced from $\det(A - \mu I)$

$$\begin{aligned}\det(B - \nu I) = {} & \det\left(C - (d + e - f + \nu)\, I\right) \times \det\left(C - (d - e + f + \nu)\, I\right) \\ & \times \det\left(C - (-d + e + f + \nu)\, I\right) \times \det\left(C - (-d - e - f + \nu)\, I\right).\end{aligned}$$

Therefore, by substituting in (5) $\nu = a + b - c + \mu$, $\nu = a - b + c + \mu$, $\nu = -a + b + c + \mu$ or $\nu = -a - b - c + \mu$, the determinant $\det(A - \mu I)$ becomes

$$\begin{aligned}\det(A - \mu I) = {} & \det\left(C - (a + b - c + d + e - f + \mu)\, I\right) \\ & \times \det\left(C - (a + b - c + d - e + f + \mu)\, I\right) \\ & \times \det\left(C - (a + b - c - d + e + f + \mu)\, I\right) \\ & \times \det\left(C - (a + b - c - d - e - f + \mu)\, I\right) \\ & \times \det\left(C - (a - b + c + d + e - f + \mu)\, I\right)\end{aligned}$$

$$\times \det\left(C-(a-b+c+d-e+f+\mu)\,I\right)$$
$$\times \det\left(C-(a-b+c-d+e+f+\mu)\,I\right)$$
$$\times \det\left(C-(a-b+c-d-e-f+\mu)\,I\right)$$
$$\times \det\left(C-(-a+b+c+d+e-f+\mu)\,I\right)$$
$$\times \det\left(C-(-a+b+c+d-e+f+\mu)\,I\right)$$
$$\times \det\left(C-(-a+b+c-d+e+f+\mu)\,I\right)$$
$$\times \det\left(C-(-a+b+c-d-e-f+\mu)\,I\right)$$
$$\times \det\left(C-(-a-b-c+d+e-f+\mu)\,I\right)$$
$$\times \det\left(C-(-a-b-c+d-e+f+\mu)\,I\right)$$
$$\times \det\left(C-(-a-b-c-d+e+f+\mu)\,I\right)$$
$$\times \det\left(C-(-a-b-c-d-e-f+\mu)\,I\right)$$
$$= \prod_{i=1}^{16} \det\left(C-T_i\,(a,b,c,d,e,f,\mu)\,I\right) \tag{6}$$

where T_i is an internal term as a function of a, b, c, d, e, f and μ. After linear combinations, the determinant $\det(C-\xi I)$ is equal to

$$\det(C-\xi I) = (g+h+k-\xi)(g-h-k-\xi)$$
$$\times(-g+h-k-\xi)(-g-h+k-\xi).$$

Therefore, by substituting in (6) ξ with the 16 terms $T_i\,(a,b,c,d,e,f,\mu)$, the determinant $\det(A-\mu I)$ is obtained and then, the eigenvalues λ_k of M are deduced. There are 64 eigenvalues λ_k of M of algebraic multiplicity 1 (Annex 1).

The 64 eigenvectors of M associated with these 64 eigenvalues λ_k computed by formal calculus can be put in a form independent of a, b, c, d, e, f, g, h and k (data not shown).

The independent mixing of the 20 trinucleotides of the code $\mathcal{C}$ with equiprobability (1/20) leads to the following initial vector $P(0) = [0, 1/20, 0, 1/20, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 1/20, 0, 1/20, 0, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/20, 1/20, 0, 1/20, 1/20, 1/20, 1/20, 0, 1/20, 0, 0, 0, 1/20, 0, 1/20, 1/20, 1/20, 0, 1/20, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/20, 0, 0]$.

The formula (4) with the initial probability vector $P(0)$ before the substitution process ($t=0$), the diagonal matrix e^{Dt} of exponential eigenvalues $e^{\lambda_k t}$ of M, its eigenvector matrix Q and its inverse Q^{-1}, determine the 64 trinucleotide probabilities $P_i(t)$ after t substitutions as a function of the nine parameters a, b, c, d, e, f, g, h and k.

Then, the occurrence probability $P(X,t)$ of a trinucleotide set X at the evolutionary time t as a function of the nine substitution parameters a, b, c, d, e, f, g, h and k, is

$$P(X,t) = \sum_{i \in X} P_i(t). \tag{7}$$

This formula $P(X,t)$ allows the evolutionary analytical formulas $P(\mathcal{C},t)$ and $P(\mathbf{A},t)$ of the common circular code $\mathcal{C}$ and its 12 amino acids $\mathbf{A} \in \mathcal{A}$, respectively, to be deduced. As the code $\mathcal{C}$ cannot contain a trinucleotide $\mathcal{T}_{\mathrm{id}}$ (AAA, CCC, GGG, TTT) by definition, its probability $P(\mathcal{C},t)$ is renormalized. Furthermore, it can be expressed as a function of eigenvalues λ_k of M, λ_k being given in Annex 1

$$\begin{aligned} P(\mathcal{C},t) &= \frac{\sum_{i\in\mathcal{C}} P_i(t)}{\sum_{i\in\mathcal{T}-\mathcal{T}_{\mathrm{id}}} P_i(t)} \\ &= \frac{1}{2D}(100 + 25e^{\lambda_2 t} + 9e^{\lambda_3 t} + 4e^{\lambda_4 t} + e^{\lambda_6 t} + e^{\lambda_7 t} + 4e^{\lambda_8 t} + 36e^{\lambda_9 t} \\ &\quad + e^{\lambda_{10} t} + e^{\lambda_{11} t} + e^{\lambda_{14} t} + e^{\lambda_{15} t} + 25e^{\lambda_{17} t} + 16e^{\lambda_{18} t} + 4e^{\lambda_{19} t} + e^{\lambda_{20} t} + e^{\lambda_{21} t} \\ &\quad + e^{\lambda_{24} t} + e^{\lambda_{25} t} + 16e^{\lambda_{27} t} + e^{\lambda_{28} t} + e^{\lambda_{29} t} + 4e^{\lambda_{31} t} + e^{\lambda_{32} t} + 9e^{\lambda_{33} t} + 4e^{\lambda_{34} t} \\ &\quad + e^{\lambda_{36} t} + e^{\lambda_{37} t} + e^{\lambda_{40} t} + e^{\lambda_{41} t} + 16e^{\lambda_{42} t} + e^{\lambda_{44} t} + e^{\lambda_{45} t} + 4e^{\lambda_{46} t} + e^{\lambda_{48} t} \\ &\quad + 4e^{\lambda_{49} t} + e^{\lambda_{50} t} + e^{\lambda_{51} t} + 4e^{\lambda_{52} t} + 4e^{\lambda_{53} t} + e^{\lambda_{54} t} + e^{\lambda_{55} t} + e^{\lambda_{58} t} + e^{\lambda_{59} t} \\ &\quad + 4e^{\lambda_{60} t} + e^{\lambda_{62} t} + e^{\lambda_{63} t}) \end{aligned} \tag{8}$$

with the denominator D

$$\begin{aligned} D = 150 &- e^{\lambda_6 t} + e^{\lambda_{11} t} + 4e^{\lambda_{18} t} - e^{\lambda_{21} t} + e^{\lambda_{28} t} + 2e^{\lambda_{31} t} - e^{\lambda_{40} t} + e^{\lambda_{41} t} + 2e^{\lambda_{46} t} \\ &+ 2e^{\lambda_{52} t} - e^{\lambda_{55} t} + e^{\lambda_{58} t}. \end{aligned}$$

In Annex 2, we give the analytical formulas $P(\mathbf{A},t) = \sum_{i\in\mathcal{T}_{\mathbf{A}}} P_i(t)$ of the 12 amino acids $\mathbf{A}$, $\mathcal{T}_{\mathbf{A}}$ being the set of trinucleotides $\mathcal{T}$ coding the amino acid $\mathbf{A}$, for the reader who wants detailed results (see also Discussion).

Property 1. *The initial probability $P(X,0)$ of a trinucleotide set X (the code $\mathcal{C}$ or an amino acid $\mathbf{A}$) at the time $t = 0$ can (obviously) be obtained from the analytical solution $P(X,t)$ with $t = 0$ (8 and Annex 2) or also by a simple probability calculus.*

Indeed, the probability $P(\mathcal{C},0)$ is equal to 1 as the primitive genes in this evolutionary model are generated by the code $\mathcal{C}$ (20 among 20 trinucleotides).

The probability $P(\mathbf{A},0)$ is also equal to the number of trinucleotides of $\mathcal{C}$ coding $\mathbf{A}$ divided by 20 (deduced from Table 1).

Property 2. *The probability $P(X,t)$ of a trinucleotide set X at the limit time $t \to \infty$ can (obviously) be obtained from their limit study (8 and Annex 2) or also by a simple probability calculus.*

Whatever $a, b, c, d, e, f, g, h, k \in]0, 1[$ *such that* $a+b+c+d+e+f+g+h+k=1$, $\lim_{t\to\infty} P(\mathcal{C}, t) = 1/3$. *Indeed, the nine substitutions in the* 20 *trinucleotides of* $\mathcal{C}$ *generate the* 44 *other trinucleotides. When* $t \to \infty$, *the* 64 *trinucleotides* $\mathcal{T}$ *occur with the same probability and therefore, the probability of* $\mathcal{C}$ *is equal to* $20/60 = 1/3$ *(the four trinucleotides* $\mathcal{T}_{\mathrm{id}}$ *being not considered).*

Whatever $a, b, c, d, e, f, g, h, k \in]0, 1[$ *such that* $a+b+c+d+e+f+g+h+k=1$, $\lim_{t\to\infty} P(\mathbf{A}, t) = \lim_{t\to\infty} \sum_{i\in\mathcal{T}_{\mathbf{A}}} P_i(t) = K_{\mathbf{A}}$ where $K_{\mathbf{A}}$ is a constant.

Property 3. *When one (or more) substitution has a rate equal to* 0, *some trinucleotides may be either not generated or generated without equiprobability and* $\lim_{t\to\infty} P(\mathcal{C}, t) \neq 1/3$, *or* $\lim_{t\to\infty} P(\mathbf{A}, t) \neq K_{\mathbf{A}}$. *As an example, we explain by a simple probability calculus why* $\lim_{t\to\infty} P(\mathcal{C}, t) = 5/12$ *when* $b = c = 0$, *i.e. no transversion in the first trinucleotide sites of* $\mathcal{C}$. *The code* $\mathcal{C}$ *has* 20 *trinucleotides with* 15 *trinucleotides beginning with a purine base forming the subset* $\mathcal{C}_R$ *and five trinucleotides beginning with a pyrimidine base forming the subset* $\mathcal{C}_Y$, $\mathcal{C} = \mathcal{C}_R \cup \mathcal{C}_Y$. *Each trinucleotide* $w \in \mathcal{C}$ *occurs with the same probability* $P(w) = 1/20$. *As there are purine and pyrimidine bases in the first trinucleotide sites of* $\mathcal{C}$ *and as the transitions and the transversions are allowed in the second and third sites of* $\mathcal{C}$ $(d, e, f, g, h, k > 0)$, *the* 64 *trinucleotides* $\mathcal{T}$ *are generated during the evolutionary process. Among these* 64 *trinucleotides* $\mathcal{T}$, *let* $\mathcal{T}_R$ *be the subset of* 32 *trinucleotides beginning with a purine base and* $\mathcal{T}_Y$, *the subset of* 32 *trinucleotides beginning with a pyrimidine base,* $\mathcal{T} = \mathcal{T}_R \cup \mathcal{T}_Y$. *As in the first sites of* $\mathcal{C}$ *the transitions are allowed* $(a > 0)$ *but not the transversions* $(b = c = 0)$, *the trinucleotide set* $\mathcal{T}_R$ *can only be generated from* $\mathcal{C}_R$. *When* $t \to \infty$, *the trinucleotides* w *of* $\mathcal{C}_R$ *and* $\mathcal{T}_R$ *occur with the same probability* $P(w, t) = (15/20)/32 = 3/128$. *Similarly, when* $t \to \infty$, *the trinucleotides* w *of* $\mathcal{C}_Y$ *and* $\mathcal{T}_Y$ *occur with the same probability* $P(w, t) = (5/20)/32 = 1/128$. *The trinucleotides AAA and GGG (CCC and TTT resp.) belong to* $\mathcal{T}_{id_R}$ *(*$\mathcal{T}_{id_Y}$ *resp.). Therefore, when* $t \to \infty$, *the trinucleotides* w *of* $\mathcal{T}_{\mathrm{id}}$ *occur with the same probability* $P(w, t) = (6+2)/128 = 1/16$. *Finally,* $\lim_{t\to\infty} P(\mathcal{C}, t)$ *is equal to*

$$\lim_{t\to\infty} P(\mathcal{C}, t) = \frac{\sum_{w\in\mathcal{C}_R\cup\mathcal{C}_Y} \lim_{t\to\infty} P(w, t)}{1 - \lim_{\substack{t\to\infty \\ w\in\mathcal{T}_{\mathrm{id}}}} P(w, t)} = \frac{\frac{45+5}{128}}{1 - \frac{1}{16}} = \frac{5}{12}.$$

Property 4. *The evolutionary analytical formula* $P_1(\mathcal{C}, t)$ *of the common circular code* $\mathcal{C}$ *as a function of the three substitution rates* p, q *and* r *associated with the three trinucleotide sites respectively, is a particular case of* $P(\mathcal{C}, t)$ *(8) with* $a = b = c = p/3$, $d = e = f = q/3$ *and* $g = h = k = r/3$

$$P_1(\mathcal{C}, t) = \frac{1}{2D_1}\Big(50 + 28\mathrm{e}^{-\frac{4}{3}t} + 19\mathrm{e}^{-\frac{4}{3}pt} + 18\mathrm{e}^{-\frac{4}{3}qt} + 19\mathrm{e}^{-\frac{4}{3}rt} + 5\mathrm{e}^{-\frac{4}{3}(1-p)t} + 16\mathrm{e}^{-\frac{4}{3}(1-q)t} + 5\mathrm{e}^{-\frac{4}{3}(1-r)t}\Big)$$

with the denominator D_1

$$D_1 = 75 + 2\mathrm{e}^{-\frac{4}{3}t} + 3\mathrm{e}^{-\frac{4}{3}(1-q)t}.$$

Property 5. *The evolutionary analytical formula* $P_2(\mathcal{C}, t)$ *of the common circular code* $\mathcal{C}$ *as a function of the six substitution rates u, v, w, x, y and z such that u and v (w and x, y and z resp.) are the transition and the transversion rates in the first (second and third resp.) trinucleotide sites respectively, is a particular case of* $P(\mathcal{C}, t)$ *(8) with* $a = u$, $b = v/2$, $c = v/2$, $d = w$, $e = x/2$, $f = x/2$, $g = y$, $h = z/2$ *and* $k = z/2$ *(Frey and Michel, 2006)*

$$\begin{aligned} P_2(\mathcal{C}, t) = \frac{1}{2D_2}(100 &+ 25\mathrm{e}^{\mu_1 t} + \mathrm{e}^{\mu_2 t} + 25\mathrm{e}^{\mu_3 t} + 16\mathrm{e}^{\mu_4 t} + \mathrm{e}^{\mu_5 t} + 13\mathrm{e}^{\mu_6 t} + 5\mathrm{e}^{\mu_7 t} \\ &+ 36\mathrm{e}^{\mu_8 t} + 2\mathrm{e}^{\mu_9 t} + 5\mathrm{e}^{\mu_{10} t} + \mathrm{e}^{\mu_{11} t} + 2\mathrm{e}^{\mu_{12} t} + 13\mathrm{e}^{\mu_{13} t} + 5\mathrm{e}^{\mu_{14} t} + 5\mathrm{e}^{\mu_{15} t} \\ &+ \mathrm{e}^{\mu_{16} t} + 2\mathrm{e}^{\mu_{17} t} + 22\mathrm{e}^{\mu_{18} t} + 6\mathrm{e}^{\mu_{19} t} + 2\mathrm{e}^{\mu_{20} t} + 2\mathrm{e}^{\mu_{21} t} + 22\mathrm{e}^{\mu_{22} t} + 8\mathrm{e}^{\mu_{23} t}) \end{aligned}$$

with the denominator D_2

$$D_2 = 150 - \mathrm{e}^{\mu_2 t} + 4\mathrm{e}^{\mu_4 t} - \mathrm{e}^{\mu_5 t} + \mathrm{e}^{\mu_{17} t} + 3\mathrm{e}^{\mu_{18} t} + 2\mathrm{e}^{\mu_{19} t} - 2\mathrm{e}^{\mu_{20} t} + \mathrm{e}^{\mu_{21} t} + 3\mathrm{e}^{\mu_{22} t}$$

and with $\mu_1 = -1 + u + v + w + x + y - z$, $\mu_2 = -1 + u + v + w - x + y - z$, $\mu_3 = -1 + u - v + w + x + y + z$, $\mu_4 = -1 + u - v + w + x + y - z$, $\mu_5 = -1 + u - v + w - x + y + z$, $\mu_6 = -1 + u + v + w + x - y$, $\mu_7 = -1 + u + v + w - x - y$, $\mu_8 = -1 + u + v - w + y + z$, $\mu_9 = -1 + u + v - w + y - z$, $\mu_{10} = -1 + u - v + w + x - y$, $\mu_{11} = -1 + u - v + w - x - y$, $\mu_{12} = -1 + u - v - w + y + z$, $\mu_{13} = -1 - u + w + x + y + z$, $\mu_{14} = -1 - u + w + x + y - z$, $\mu_{15} = -1 - u + w - x + y + z$, $\mu_{16} = -1 - u + w - x + y - z$, $\mu_{17} = -1 + u + v - w - y$, $\mu_{18} = -1 + u - v - w - y$, $\mu_{19} = -1 - u + w + x - y$, $\mu_{20} = -1 - u + w - x - y$, $\mu_{21} = -1 - u - w + y + z$, $\mu_{22} = -1 - u - w + y - z$ *and* $\mu_{23} = -1 - u - w - y$.

Property 6. *The formula* $P(t) = Q \cdot \mathrm{e}^{Dt} \cdot Q^{-1} \cdot P(0)$ *(4) gives the trinucleotide occurrence probabilities at the evolutionary time t from their past ones* $P(0)$. *By expressing* $P(0)$ *as a function of* $P(t)$ *in (4), then* $P(0) = Q \cdot \mathrm{e}^{-Dt} \cdot Q^{-1} \cdot P(t)$. *Therefore, the formula* $\widetilde{P}(t) = Q \cdot \mathrm{e}^{-Dt} \cdot Q^{-1} \cdot \widetilde{P}(0)$, *by replacing t by* $-t$ *in (4), gives the past trinucleotide occurrence probabilities from their actual ones* $\widetilde{P}(0)$, *i.e. by inverting the direction of the evolutionary time.*

Property 7. *Let* $t_0 < t_1 < t_2$ *be three evolutionary times. Let* $P(t_1)$ *and* $P(t_2)$ *be the trinucleotide occurrence probabilities at the evolutionary times* t_1 *and* t_2, *respectively, as a function of their past ones* $P(t_0)$, *i.e.* $P(t_1) = Q \cdot \mathrm{e}^{Dt_1} \cdot Q^{-1} \cdot P(t_0)$ *and* $P(t_2) = Q \cdot \mathrm{e}^{Dt_2} \cdot Q^{-1} \cdot P(t_0)$. *Then,* $P(t_2)$ *can be expressed as a function of* $P(t_1)$ *such that* $P(t_2) = Q \cdot \mathrm{e}^{D(t_2 - t_1)} \cdot Q^{-1} \cdot P(t_1)$.

3. Results

The 12 amino acids **A** coded by the common circular code $\mathcal{C}$ have initial probabilities $P(\mathbf{A}, 0)$ ranging from $1/20 = 5\%$ to $3/20 = 15\%$ (Table 1). They can be classified into three groups: *Ala*, *Gln*, *Phe*, *Thr* and *Tyr* with the lowest probability $P(\mathbf{A}, 0) = 5\%$, *Asn*, *Asp*, *Glu*, *Gly*, *Ile* and *Leu* with $P(\mathbf{A}, 0) = 10\%$ and *Val* with the highest probability $P(\mathbf{A}, 0) = 15\%$. This distribution is completely different from the one coded by the actual genes, both from their absolute and relative values (Table 2). Therefore, a random mutation process seems a priori completely unable to retrieve the actual amino acid distribution from the past one coded by the code $\mathcal{C}$.

The stochastic model developed here allows the investigation of such a property by searching the main statistical features of the actual amino acid distribution: *Leu* with the highest frequency, *Tyr* with the lowest one, *Ala*, *Gly* and *Val* with the second highest one, *Asn*, *Gln* and *Phe* with the second lowest one, and *Asp*, *Glu*, *Ile* and *Thr* with an average one (Table 2). Each substitution parameter a, b, c, d, e, f, g, h and k varies in the range $[0, 1]$ with a step of 2% such that their probability sum is equal to 1, and t, in the range $[0, 10]$.

Very unexpectedly, this model retrieves a distribution of the 12 amino acids **A** close to the actual one after an evolutionary time $t \approx 8$ of random substitutions in the common circular code $\mathcal{C}$ and with particular values for the nine substitution parameters (Table 3, Fig. 1). Table 3 gives the barycenter of the solution space (not given) of the nine substitution rates.

This model involves mainly the transition d in the second trinucleotide sites of $\mathcal{C}$, and to a lesser extent, the transversions h and k in its third sites and the transversion b in its first sites, while the transversions c and e in its first and second sites, respectively, have small effects.

The stochastic curves have a complex behaviour (Fig. 1). After an initial decrease up to $t = 1.93$, the curve *Leu* increases, crosses the four curves *Asp*, *Ala*, *Val* and *Gly* successively and becomes the highest one, such as in actual time, after an evolutionary time $t \geq 6.24$. The three curves *Ala*, *Gly* and *Val* (group 1) gather at $t \approx 8$ after different stochastic behaviours: the curve *Val* always decreases from its initial value while the curves *Ala* and *Gly* first increase then decrease. The gathering value of these three curves is lower than the initial values of *Gly* and *Val* but higher than the one of *Ala*. As in actual time, *Ile* and *Thr* (group 2) occur with probabilities lower than those of group 1, *Asn*, *Gln* and *Phe* (group 4) occur with probabilities ranging between those of *Asp* and *Glu* (group 3) and *Tyr*. As in actual time, *Tyr* has the lowest occurrence probability. Certain curves have local

Table 3 Substitution rate barycenter (in %) in the stochastic model leading to a distribution of the 12 amino acids **A** close to the actual one.

Parameters	a	b	c	d	e	f	g	h	k
Barycenter (%)	4.5	9.2	1.7	52.6	2.4	5.9	5.7	8.6	9.4

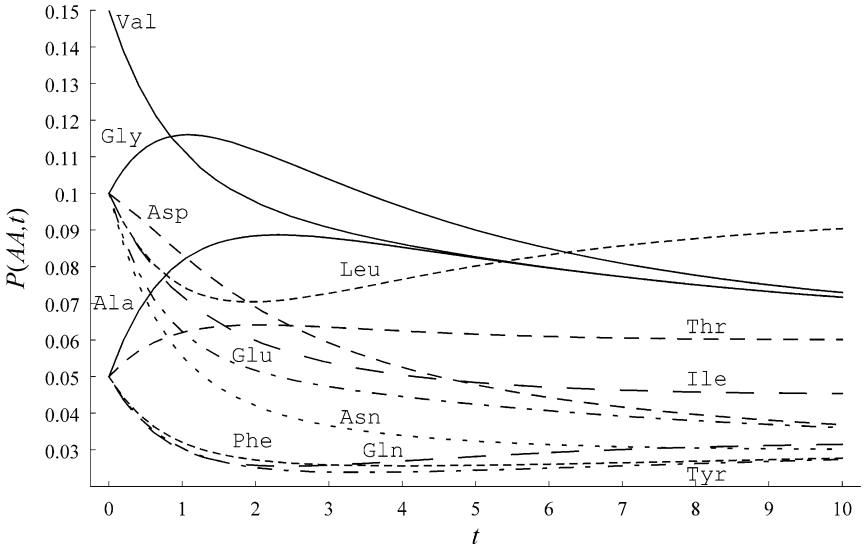

Fig. 1 Evolution of the 12 amino acids **A** from random substitutions in the common circular code $\mathcal{C}$ in the substitution rate barycenter (in %): $a = 4.5$, $b = 9.2$, $c = 1.7$, $d = 52.6$, $e = 2.4$, $f = 5.9$, $g = 5.7$, $h = 8.6$ and $k = 9.4$ (Table 3).

maxima, e.g. *Ala*, *Gly* and *Thr*, or local minima, e.g. *Gln*, *Leu*, *Phe* and *Tyr*, or continuous decreases, e.g. *Asn*, *Asp*, *Glu*, *Ile* and *Val*.

4. Discussion

A new analytical evolutionary model has been developed here in order to generalize several previous models based on the nucleotide mutation matrices 4×4 (Jukes and Cantor, 1969; Kimura, 1980) and the trinucleotide mutation matrix 64×64 with three and six substitution parameters (Arquès et al., 1998; Frey and Michel, 2006). Furthermore, an application of this model allows the evolutionary probabilities of the common circular code $\mathcal{C}$ found in actual genes of eukaryotes and prokaryotes and the 12 amino acids coded by this code $\mathcal{C}$ to be derived as a function of the time t and nine substitution parameters associated with the three types of substitutions in the three trinucleotide sites.

Very unexpectedly, an amino acid distribution very close to the actual one can be derived after an evolutionary time $t \approx 8$ of random substitutions in the common circular code $\mathcal{C}$ and with particular values for the nine substitution parameters (Fig. 1 and Table 3). The main effect is related to the transition d in the second trinucleotide sites of $\mathcal{C}$, in agreement with the chemical properties of nucleotides (one carbon–nitrogen ring for pyrimidines and two carbon–nitrogen rings for purines) and the complementary base pairing showing a universal transition/transversion rate bias in prokaryotic and eukaryotic genomes, e.g. Ochman (2003), Rosenberg

et al. (2003). This model retrieves the main statistical properties of the actual amino acid distribution. However, it cannot explain all the observed features, e.g. the actual probability of *Ala* is higher than those of *Gly* and *Val* (Table 2) while its probability curve cannot cross them in this model (Fig. 1). More general models, e.g. with non-symmetrical mutation matrices, could improve the correlation with reality.

The variations of the stochastic curves $P(\mathbf{A}, t)$ of the 12 amino acids $\mathbf{A}$ cannot obviously be predicted without modelling as their analytical solutions are based on a sum of several exponential terms, e.g. 46 terms for $P(Ile, t)$ (Annex 2), each exponential term being a function of the time t and the nine substitutions parameters. The probability differences existing between the amino acids coded by the "primitive" genes (at $t = 0$) have still some properties after a great number of random substitutions in genes, e.g. at $t = 10$ in Fig. 1. In other words, some primitive traces of amino acid variations can still be observed after a long period of random evolution in spite of the generation of noise. Several properties with these stochastic amino acid curves have been observed with particular values for the nine substitution rates: curves with crossings, curves with local maxima and minima, curves with continuous increases or decreases, curves with a series of fusions and separations, etc. (data not shown). They have not been investigated as they are not directly in the subject of this paper. However, as already mentioned in Section 2, the evolutionary analytical formulas $P(\mathbf{A}, t)$ of the 12 amino acids $\mathbf{A}$ are given in Annex 2 for the reader who wants to deepen the analysis of these stochastic curves with different values for the nine substitution parameters.

The biological meaning of this amino acid evolutionary model would suggest that the primitive genes (at $t = 0$) are constructed by trinucleotides of the common circular code $\mathcal{C}$. 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 set $\mathcal{C}$ which is a maximal self-complementary $\mathcal{C}^3$ code (Section 1.2.2). 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. From a biological point of view, this process can be compared with a mixing of trinucleotides in a primitive soup. A Markov concatenation of trinucleotides (based on a stochastic matrix) would have been too complex at this primitive time.

The mathematical model developed here has demonstrated that a construction process based on the common circular code $\mathcal{C}$ and a random evolutionary process with nine substitutions parameters, retrieves the main properties of the amino acid distribution coded by actual genes. Furthermore, it can be applied to other problems. In particular, the eigenvalues obtained here and given in Annex 1 can be directly used to develop other evolutionary models based on a trinucleotide mutation matrix with nine substitution parameters associated with the three types of substitutions in the three trinucleotide sites. Finally, this approach could also improve some algorithms of phylogenetic tree reconstruction and sequence alignment.

Annex 1: The 64 eigenvalues λ_k of M of algebraic multiplicity 1

$\lambda_1 = -1+a+b+c+d+e+f+g+h+k$, $\lambda_2 = -1+a+b+c+d+e+f+g-h-k$,
$\lambda_3 = -1+a+b+c+d+e+f-g+h-k$, $\lambda_4 = -1+a+b+c+d+e+f-g-h+k$,
$\lambda_5 = -1+a+b+c+d-e-f+g+h+k$, $\lambda_6 = -1+a+b+c+d-e-f+g-h-k$,
$\lambda_7 = -1+a+b+c+d-e-f-g+h-k$, $\lambda_8 = -1+a+b+c+d-e-f-g-h+k$,
$\lambda_9 = -1+a+b+c-d+e-f+g+h+k$, $\lambda_{10} = -1+a+b+c-d+e-f+g-h-k$,
$\lambda_{11} = -1+a+b+c-d+e-f-g+h-k$, $\lambda_{12} = -1+a+b+c-d+e-f-g-h+k$,
$\lambda_{13} = -1+a+b+c-d-e+f+g+h+k$, $\lambda_{14} = -1+a+b+c-d-e+f+g-h-k$,
$\lambda_{15} = -1+a+b+c-d-e+f-g+h-k$, $\lambda_{16} = -1+a+b+c-d-e+f-g-h+k$,
$\lambda_{17} = -1+a-b-c+d+e+f+g+h+k$, $\lambda_{18} = -1+a-b-c+d+e+f+g-h-k$,
$\lambda_{19} = -1+a-b-c+d+e+f-g+h-k$, $\lambda_{20} = -1+a-b-c+d+e+f-g-h+k$,
$\lambda_{21} = -1+a-b-c+d-e-f+g+h+k$, $\lambda_{22} = -1+a-b-c+d-e-f+g-h-k$,
$\lambda_{23} = -1+a-b-c+d-e-f-g+h-k$, $\lambda_{24} = -1+a-b-c+d-e-f-g-h+k$,
$\lambda_{25} = -1+a-b-c-d+e-f+g+h+k$, $\lambda_{26} = -1+a-b-c-d+e-f+g-h-k$,
$\lambda_{27} = -1+a-b-c-d+e-f-g+h-k$, $\lambda_{28} = -1+a-b-c-d+e-f-g-h+k$,
$\lambda_{29} = -1+a-b-c-d-e+f+g+h+k$, $\lambda_{30} = -1+a-b-c-d-e+f+g-h-k$,
$\lambda_{31} = -1+a-b-c-d-e+f-g+h-k$, $\lambda_{32} = -1+a-b-c-d-e+f-g-h+k$,
$\lambda_{33} = -1-a+b-c+d+e+f+g+h+k$, $\lambda_{34} = -1-a+b-c+d+e+f+g-h-k$,
$\lambda_{35} = -1-a+b-c+d+e+f-g+h-k$, $\lambda_{36} = -1-a+b-c+d+e+f-g-h+k$,
$\lambda_{37} = -1-a+b-c+d-e-f+g+h+k$, $\lambda_{38} = -1-a+b-c+d-e-f+g-h-k$,
$\lambda_{39} = -1-a+b-c+d-e-f-g+h-k$, $\lambda_{40} = -1-a+b-c+d-e-f-g-h+k$,
$\lambda_{41} = -1-a+b-c-d+e-f+g+h+k$, $\lambda_{42} = -1-a+b-c-d+e-f+g-h-k$,
$\lambda_{43} = -1-a+b-c-d+e-f-g+h-k$, $\lambda_{44} = -1-a+b-c-d+e-f-g-h+k$,
$\lambda_{45} = -1-a+b-c-d-e+f+g+h+k$, $\lambda_{46} = -1-a+b-c-d-e+f+g-h-k$,
$\lambda_{47} = -1-a+b-c-d-e+f-g+h-k$, $\lambda_{48} = -1-a+b-c-d-e+f-g-h+k$,
$\lambda_{49} = -1-a-b+c+d+e+f+g+h+k$, $\lambda_{50} = -1-a-b+c+d+e+f+g-h-k$,
$\lambda_{51} = -1-a-b+c+d+e+f-g+h-k$, $\lambda_{52} = -1-a-b+c+d+e+f-g-h+k$,
$\lambda_{53} = -1-a-b+c+d-e-f+g+h+k$, $\lambda_{54} = -1-a-b+c+d-e-f+g-h-k$,
$\lambda_{55} = -1-a-b+c+d-e-f-g+h-k$, $\lambda_{56} = -1-a-b+c+d-e-f-g-h+k$,
$\lambda_{57} = -1-a-b+c-d+e-f+g+h+k$, $\lambda_{58} = -1-a-b+c-d+e-f+g-h-k$,
$\lambda_{59} = -1-a-b+c-d+e-f-g+h-k$, $\lambda_{60} = -1-a-b+c-d+e-f-g-h+k$,
$\lambda_{61} = -1-a-b+c-d-e+f+g+h+k$, $\lambda_{62} = -1-a-b+c-d-e+f+g-h-k$,
$\lambda_{63} = -1-a-b+c-d-e+f-g+h-k$, $\lambda_{64} = -1-a-b+c-d-e+f-g-h+k$.

Annex 2: Evolutionary analytical formulas of the 12 amino acids A

With the eigenvalues λ_k of M (Annex 1), the evolutionary analytical formulas $P(\mathbf{A},t) = \sum_{i\in T_{\mathbf{A}}} P_i(t)$ (7) of the 12 amino acids **A** obtained are

$$\begin{aligned}
P(Ala,t) &= P_{GCA}(t) + P_{GCC}(t) + P_{GCG}(t) + P_{GCT}(t)\\
&= \frac{1}{160}\Big(10 - 6e^{\lambda_9 t} + 5e^{\lambda_{17} t} - e^{\lambda_{21} t} - e^{\lambda_{25} t} + e^{\lambda_{29} t} + 3e^{\lambda_{33} t} - e^{\lambda_{37} t}\\
&\quad - e^{\lambda_{41} t} - e^{\lambda_{45} t} + 2e^{\lambda_{49} t} - 2e^{\lambda_{53} t}\Big)\\
P(Asn,t) &= P_{AAC}(t) + P_{AAT}(t)\\
&= \frac{1}{320}\Big(10 + 5e^{\lambda_2 t} - e^{\lambda_6 t} + 6e^{\lambda_9 t} + e^{\lambda_{10} t} - e^{\lambda_{14} t} + 5e^{\lambda_{17} t} + 4e^{\lambda_{18} t}
\end{aligned}$$

$$
\begin{aligned}
&+ e^{\lambda_{21}t} + e^{\lambda_{25}t} + e^{\lambda_{29}t} - 3e^{\lambda_{33}t} + 2e^{\lambda_{34}t} - e^{\lambda_{37}t} - e^{\lambda_{41}t} + 4e^{\lambda_{42}t} \\
&+ e^{\lambda_{45}t} + 2e^{\lambda_{46}t} - 2e^{\lambda_{49}t} - e^{\lambda_{50}t} - 2e^{\lambda_{53}t} - e^{\lambda_{54}t} + e^{\lambda_{58}t} + e^{\lambda_{62}t}\Big) \\
P(Asp,t) &= P_{GAC}(t) + P_{GAT}(t) \\
&= \frac{1}{320}\Big(10 + 5e^{\lambda_{2}t} - e^{\lambda_{6}t} + 6e^{\lambda_{9}t} + e^{\lambda_{10}t} - e^{\lambda_{14}t} + 5e^{\lambda_{17}t} + 4e^{\lambda_{18}t} \\
&+ e^{\lambda_{21}t} + e^{\lambda_{25}t} + e^{\lambda_{29}t} + 3e^{\lambda_{33}t} - 2e^{\lambda_{34}t} + e^{\lambda_{37}t} + e^{\lambda_{41}t} - 4e^{\lambda_{42}t} \\
&- e^{\lambda_{45}t} - 2e^{\lambda_{46}t} + 2e^{\lambda_{49}t} + e^{\lambda_{50}t} + 2e^{\lambda_{53}t} + e^{\lambda_{54}t} - e^{\lambda_{58}t} - e^{\lambda_{62}t}\Big) \\
P(Gln,t) &= P_{CAA}(t) + P_{CAG}(t) \\
&= \frac{1}{320}\Big(10 - 5e^{\lambda_{2}t} + e^{\lambda_{6}t} + 6e^{\lambda_{9}t} - e^{\lambda_{10}t} + e^{\lambda_{14}t} - 5e^{\lambda_{17}t} + 4e^{\lambda_{18}t} \\
&- e^{\lambda_{21}t} - e^{\lambda_{25}t} - e^{\lambda_{29}t} + 3e^{\lambda_{33}t} + 2e^{\lambda_{34}t} + e^{\lambda_{37}t} + e^{\lambda_{41}t} + 4e^{\lambda_{42}t} \\
&- e^{\lambda_{45}t} + 2e^{\lambda_{46}t} - 2e^{\lambda_{49}t} + e^{\lambda_{50}t} - 2e^{\lambda_{53}t} + e^{\lambda_{54}t} - e^{\lambda_{58}t} - e^{\lambda_{62}t}\Big) \\
P(Glu,t) &= P_{GAA}(t) + P_{GAG}(t) \\
&= \frac{1}{320}\Big(10 - 5e^{\lambda_{2}t} + e^{\lambda_{6}t} + 6e^{\lambda_{9}t} - e^{\lambda_{10}t} + e^{\lambda_{14}t} + 5e^{\lambda_{17}t} - 4e^{\lambda_{18}t} \\
&+ e^{\lambda_{21}t} + e^{\lambda_{25}t} + e^{\lambda_{29}t} + 3e^{\lambda_{33}t} + 2e^{\lambda_{34}t} + e^{\lambda_{37}t} + e^{\lambda_{41}t} + 4e^{\lambda_{42}t} \\
&- e^{\lambda_{45}t} + 2e^{\lambda_{46}t} + 2e^{\lambda_{49}t} - e^{\lambda_{50}t} + 2e^{\lambda_{53}t} - e^{\lambda_{54}t} + e^{\lambda_{58}t} + e^{\lambda_{62}t}\Big) \\
P(Gly,t) &= P_{GGA}(t) + P_{GGC}(t) + P_{GGG}(t) + P_{GGT}(t) \\
&= \frac{1}{160}\Big(10 - 6e^{\lambda_{9}t} + 5e^{\lambda_{17}t} + e^{\lambda_{21}t} - e^{\lambda_{25}t} - e^{\lambda_{29}t} + 3e^{\lambda_{33}t} + e^{\lambda_{37}t} \\
&- e^{\lambda_{41}t} + e^{\lambda_{45}t} + 2e^{\lambda_{49}t} + 2e^{\lambda_{53}t}\Big) \\
P(Ile,t) &= P_{ATA}(t) + P_{ATC}(t) + P_{ATT}(t) \\
&= \frac{1}{640}\Big(30 + 5e^{\lambda_{2}t} - 3e^{\lambda_{3}t} + 2e^{\lambda_{4}t} + e^{\lambda_{6}t} - e^{\lambda_{7}t} + 2e^{\lambda_{8}t} + 18e^{\lambda_{9}t} \\
&+ e^{\lambda_{10}t} - e^{\lambda_{11}t} + e^{\lambda_{14}t} + e^{\lambda_{15}t} + 15e^{\lambda_{17}t} + 4e^{\lambda_{18}t} + 2e^{\lambda_{19}t} + e^{\lambda_{20}t} \\
&- 3e^{\lambda_{21}t} + e^{\lambda_{24}t} + 3e^{\lambda_{25}t} + 4e^{\lambda_{27}t} - e^{\lambda_{28}t} - 3e^{\lambda_{29}t} + 2e^{\lambda_{31}t} \\
&- e^{\lambda_{32}t} - 9e^{\lambda_{33}t} + 2e^{\lambda_{34}t} + e^{\lambda_{36}t} + 3e^{\lambda_{37}t} - e^{\lambda_{40}t} - 3e^{\lambda_{41}t} + 4e^{\lambda_{42}t} \\
&+ e^{\lambda_{44}t} - 3e^{\lambda_{45}t} - 2e^{\lambda_{46}t} - e^{\lambda_{48}t} - 6e^{\lambda_{49}t} - e^{\lambda_{50}t} - e^{\lambda_{51}t} - 2e^{\lambda_{52}t} \\
&+ 6e^{\lambda_{53}t} + e^{\lambda_{54}t} - e^{\lambda_{55}t} + e^{\lambda_{58}t} - e^{\lambda_{59}t} - 2e^{\lambda_{60}t} - e^{\lambda_{62}t} - e^{\lambda_{63}t}\Big) \\
P(Leu,t) &= P_{CTA}(t) + P_{CTC}(t) + P_{CTG}(t) + P_{CTT}(t) + P_{TTA}(t) + P_{TTG}(t) \\
&= \frac{1}{320}\Big(30 - 5e^{\lambda_{2}t} - e^{\lambda_{6}t} + 18e^{\lambda_{9}t} - e^{\lambda_{10}t} - e^{\lambda_{14}t} - 15e^{\lambda_{17}t} + 4e^{\lambda_{18}t}
\end{aligned}
$$

$$
\begin{aligned}
&+3e^{\lambda_{21}t}-3e^{\lambda_{25}t}+3e^{\lambda_{29}t}+3e^{\lambda_{33}t}-2e^{\lambda_{34}t}-e^{\lambda_{37}t}+e^{\lambda_{41}t}-4e^{\lambda_{42}t}\\
&+e^{\lambda_{45}t}+2e^{\lambda_{46}t}-2e^{\lambda_{49}t}-e^{\lambda_{50}t}+2e^{\lambda_{53}t}+e^{\lambda_{54}t}+e^{\lambda_{58}t}-e^{\lambda_{62}t}\Big)\\
P(Phe,t) &= P_{TTC}(t)+P_{TTT}(t)\\
&= \frac{1}{320}\Big(10+5e^{\lambda_{2}t}+e^{\lambda_{6}t}+6e^{\lambda_{9}t}+e^{\lambda_{10}t}+e^{\lambda_{14}t}-5e^{\lambda_{17}t}-4e^{\lambda_{18}t}\\
&+e^{\lambda_{21}t}-e^{\lambda_{25}t}+e^{\lambda_{29}t}-3e^{\lambda_{33}t}+2e^{\lambda_{34}t}+e^{\lambda_{37}t}-e^{\lambda_{41}t}+4e^{\lambda_{42}t}\\
&-e^{\lambda_{45}t}-2e^{\lambda_{46}t}+2e^{\lambda_{49}t}+e^{\lambda_{50}t}-2e^{\lambda_{53}t}-e^{\lambda_{54}t}-e^{\lambda_{58}t}+e^{\lambda_{62}t}\Big)\\
P(Thr,t) &= P_{ACA}(t)+P_{ACC}(t)+P_{ACG}(t)+P_{ACT}(t)\\
&= \frac{1}{160}\Big(10-6e^{\lambda_{9}t}+5e^{\lambda_{17}t}-e^{\lambda_{21}t}-e^{\lambda_{25}t}+e^{\lambda_{29}t}-3e^{\lambda_{33}t}+e^{\lambda_{37}t}\\
&+e^{\lambda_{41}t}+e^{\lambda_{45}t}-2e^{\lambda_{49}t}+2e^{\lambda_{53}t}\Big)\\
P(Tyr,t) &= P_{TAC}(t)+P_{TAT}(t)\\
&= \frac{1}{320}\Big(10+5e^{\lambda_{2}t}-e^{\lambda_{6}t}+6e^{\lambda_{9}t}+e^{\lambda_{10}t}-e^{\lambda_{14}t}-5e^{\lambda_{17}t}-4e^{\lambda_{18}t}\\
&-e^{\lambda_{21}t}-e^{\lambda_{25}t}-e^{\lambda_{29}t}-3e^{\lambda_{33}t}+2e^{\lambda_{34}t}-e^{\lambda_{37}t}-e^{\lambda_{41}t}+4e^{\lambda_{42}t}\\
&+e^{\lambda_{45}t}+2e^{\lambda_{46}t}+2e^{\lambda_{49}t}+e^{\lambda_{50}t}+2e^{\lambda_{53}t}+e^{\lambda_{54}t}-e^{\lambda_{58}t}-e^{\lambda_{62}t}\Big)\\
P(Val,t) &= P_{GTA}(t)+P_{GTC}(t)+P_{GTG}(t)+P_{GTT}(t)\\
&= \frac{1}{160}\Big(10+6e^{\lambda_{9}t}+5e^{\lambda_{17}t}-e^{\lambda_{21}t}+e^{\lambda_{25}t}-e^{\lambda_{29}t}+3e^{\lambda_{33}t}-e^{\lambda_{37}t}\\
&+e^{\lambda_{41}t}+e^{\lambda_{45}t}+2e^{\lambda_{49}t}-2e^{\lambda_{53}t}\Big).
\end{aligned}
$$

References

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

Antezana, M.A., Kreitman, M., 1999. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J. Mol. Evol. 49, 36–43.

Arquès, D.G., Fallot, J.-P., Michel, C.J., 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., 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. Masson, Paris.

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

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

Bulmer, M., 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897–907.

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

Crick, F.H.C., Brenner, S., Klug, A., Pieczenik, G., 1976. A speculation on the origin of protein synthesis. Orig. Life 7, 389–397.
Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proc. Natl. Acad. Sci. 43, 416–421.
Fedorov, A., Saxonov, S., Gilbert, W., 2002. Regularities of context-dependent codon bias in eukaryotic genes. Nucleic Acids Res. 30, 1192–1197.
Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pave, A., 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, r49–r62.
Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., Mercier, R., 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9, r43–r74.
Eigen, M., Schuster, P., 1978. The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle. Naturwissenschaften 65, 341–369.
Ermolaeva, M.D., 2001. Synonymous codon usage in bacteria. Curr. Issues Mol. Biol. Oct. 3, 91–97.
Frey, G., Michel, C.J., 2006. An analytical model of gene evolution with 6 mutation parameters: An application to archaeal circular codes. J. Comput. Biol. Chem. 30, 1–11.
Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 12–34.
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.
Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, 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.
Koch, A.J., Lehmann, J., 1997. About a symmetry of the genetic code. J. Theor. Biol. 189, 171–174.
Konu, O., Li, M.D., 2002. Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents. J. Mol. Evol. 54, 35–41.
Krakauer, D.C., Jansen, A.A., 2002. Red Queen Dynamics of protein Translation. J. Theor. Biol. 218, 97–109.
Lacan, J., Michel, C.J., 2001. Analysis of a circular code model. J. Theor. Biol. 213, 159–170.
Lange, K., 2005. Applied Probability, Springer-Verlag, New York.
Llopart, A., Aguade, M., 2000. Nucleotide polymorphism at the RpII215 gene in *Drosophila subobscura*: Weak selection on synonymous mutations. Genetics 155, 1245–1252.
Nirenberg, M.W., Matthaei, J.H., 1961. The dependance of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc. Natl. Acad. Sci. 47, 1588–1602.
Ochman, H., 2003. Neutral mutations and neutral substitutions in bacterial genomes. Mol. Biol. Evol. 20, 2091–2096.
Rogozin, I.B., Malyarchuk, B.A., Pavlov, Y.I., Milanesi, L., 2005. From context-dependance of mutations to molecular mechanisms of mutagenesis. Pac Symp. Biocomput., 409–420.
Rosenberg, M.S., Subramanian, S., Kumar, S., 2003. Patterns of transitional mutation biases within and among mammalian genomes. Mol. Biol. Evol. 20, 988–993.
Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F., Sockett, R.E., 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 33, 1141–1153.
Sharp, P.M., Matassi, G., 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.
Smith, N.G.C., Eyre-Walker, A., 2001. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol. Biol. Evol. 18, 982–986.
Yarus, M., Folley, L.S., 1984. Sense codons are found in specific contexts. J. Mol. Biol. 182, 529–540.