Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry 31 (2007) 36–43
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

Codon phylogenetic distance

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 19 October 2006; received in revised form 22 October 2006; accepted 24 November 2006

Abstract

We develop here an analytical evolution 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 and with non-zero elements on its main diagonal. It generalizes the previous models based on the nucleotide mutation matrices 4×4 and the trinucleotide mutation matrices 64×64 with zero elements on its main diagonal. It determines at some time t the exact occurrence probabilities of trinucleotides mutating randomly according to these nine substitution parameters. Furthermore, applications of this model allow to generalize an evolutionary analytical solution of the common circular code of eukaryotes and prokaryotes and also to derive a codon phylogenetic distance.

Keywords: Evolution model; Stochastic model; Analytical solution; Mutation matrix; Gene; Trinucleotide; Codon; Circular code; Phylogenetic distance

1. Introduction

A new stochastic evolution model will determine at some time t the occurrence probabilities of trinucleotides mutating randomly according to several types of substitutions in the trinucleotide sites. Occurrence probabilities of trinucleotide sets can obviously be deduced from this approach. This model with nine substitution parameters associated with the three types of substitutions in the three trinucleotide sites and with non-zero elements on the main diagonal of the mutation matrix generalizes the previous models both based on the nucleotide mutation matrices 4×4, in particular with one substitution parameter (Jukes and Cantor, 1969), two parameters (transitions and transversions) (Kimura, 1980), three parameters (Kimura, 1981), four parameters (Takahata and Kimura, 1981) and six parameters (Kimura, 1981), and based on the trinucleotide mutation matrices 64×64 with three, six and nine substitution parameters and with zero elements on the main diagonal (Arquès et al., 1998; Frey and Michel, 2006; Michel, 2007a).

Two types of results are presented in this paper:

(i) A mathematical model of gene evolution with nine substitution parameters is developed: a, d and g are the rates of transitions $A \leftrightarrow G$ (a substitution from one purine $\{A, G\}$ to the other) and $C \leftrightarrow T$ (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) $A \leftrightarrow T$ and $C \leftrightarrow G$ in the three sites, respectively, and c, f and k are the rates of transversions $A \leftrightarrow C$ and $G \leftrightarrow T$ in the three sites, respectively.

(ii) The applications of this model proposed here allow to generalize a previous evolutionary analytical solution of the common circular code and to derive a codon phylogenetic distance.

2. Mathematical model

The mathematical model will determine at an evolutionary time t the occurrence probabilities $P(t)$ of the 64 trinucleotides mutating according to nine substitution parameters a, b, c, d, e, f, g, h and k: a, d and g are the transition rates $A \leftrightarrow G$ and $C \leftrightarrow T$ in the three sites, respectively, b, e and h are the transversion rates $A \leftrightarrow T$ and $C \leftrightarrow G$ in the three sites, respectively, and c, f and k are the transversion rates $A \leftrightarrow C$ and $G \leftrightarrow T$ in the three sites, respectively.

By convention, the indexes $i, j \in \{1, \ldots, 64\}$ represent the 64 trinucleotides $\mathcal{T} = \{AAA, \ldots, TTT\}$ 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,

* Corresponding author. Tel.: +33 3 90 24 44 62.
E-mail address: michel@dpt-info.u-strasbg.fr.

1476-9271/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2006.11.001

i.e., if j and i differ by more than one nucleotide as the time interval T is assumed to be small enough 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 $j \notin \{AAC, AAG, AAT, ACA, AGA, ATA, CAA, GAA, TAA\}$. Compared to the previous models, the substitution probability $P(i \to i)$ of a trinucleotide i into itself is introduced in this stochastic approach with a value greater than 0 (see below (2.1)).

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. By rescaling time, we can assume that $\alpha = 1$, i.e., there is one substitution per trinucleotide per time interval. Then,

$$\begin{aligned} &P_i(t+T) - P_i(t) \\ &= T \sum_{j=1}^{64} P(j \to i) P_j(t) - T P_i(t) \\ &= T \sum_{j=1 j \neq i}^{64} P(j \to i) P_j(t) + T P(i \to i) P_i(t) - T P_i(t) \\ &= T \sum_{j=1 j \neq i}^{64} P(j \to i) P_j(t) \\ &\quad + T \left(1 - \sum_{j=1 j \neq i}^{64} P(j \to i) \right) P_i(t) - T P_i(t). \end{aligned} \tag{2.1}$$

The formula (2.1) leads to

$$\lim_{T \to 0} \frac{P_i(t+T) - P_i(t)}{T} = P_i'(t) = \sum_{j=1}^{64} P(j \to i) P_j(t) - P_i(t). \tag{2.2}$$

when $T \to 0$ and with non-zero elements on the main diagonal.

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 Eq. (2.2) can be represented by the following matrix equation

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

where I represents the identity matrix and the symbol '$\cdot$' represents 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

$$A = \begin{pmatrix} & 1 \ldots 16 & 17 \ldots 32 & 33 \ldots 48 & 49 \ldots 64 \\ 1 \ldots 16 & B & cI & aI & bI \\ 17 \ldots 32 & cI & B & bI & aI \\ 33 \ldots 48 & aI & bI & B & cI \\ 49 \ldots 64 & bI & aI & cI & B \end{pmatrix}.$$

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

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

with $n = 1 - (a + b + c + d + e + f + g + h + k)$.

Remark 1. The mutation matrix A is a doubly stochastic and positive matrix.

The differential Eq. (2.3) 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 exist 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 equation has the classical solution (Lange, 2005)

$$P(t) = Q \cdot e^{Dt} \cdot Q^{-1} \cdot P(0) \tag{2.4}$$

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

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

$$\begin{aligned}\det(A - \mu I) = {} & \det(B - (a + b - c + \mu)I) \\ & \times \det(B - (a - b + c + \mu)I) \\ & \times \det(B - (-a + b + c + \mu)I) \\ & \times \det(B - (-a - b - c + \mu)I). \end{aligned} \tag{2.5}$$

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 $\det(A - \mu I)$

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

Therefore, by substituting in (2.5) $\nu = a + b - c + \mu$, $\nu = a - b + c + \mu$, $\nu = -a + b + c + \mu$ or $\nu = -a - b - c + \mu$, the determinant $\det(A - \mu I)$ becomes

$$\begin{aligned}\det(A - \mu I) = {} & \det(C - (a + b - c + d + e - f + \mu)I) \\ & \times \det(C - (a + b - c + d - e + f + \mu)I) \\ & \times \det(C - (a + b - c - d + e + f + \mu)I) \\ & \times \det(C - (a + b - c - d - e - f + \mu)I) \\ & \times \det(C - (a - b + c + d + e - f + \mu)I) \\ & \times \det(C - (a - b + c + d - e + f + \mu)I) \\ & \times \det(C - (a - b + c - d + e + f + \mu)I) \\ & \times \det(C - (a - b + c - d - e - f + \mu)I) \\ & \times \det(C - (-a + b + c + d + e - f + \mu)I) \\ & \times \det(C - (-a + b + c + d - e + f + \mu)I) \\ & \times \det(C - (-a + b + c - d + e + f + \mu)I) \\ & \times \det(C - (-a + b + c - d - e - f + \mu)I) \\ & \times \det(C - (-a - b - c + d + e - f + \mu)I) \\ & \times \det(C - (-a - b - c + d - e + f + \mu)I) \\ & \times \det(C - (-a - b - c - d + e + f + \mu)I) \\ & \times \det(C - (-a - b - c - d - e - f + \mu)I) \\ = {} & \prod_{i=1}^{16} \det(C - T_i(a, b, c, d, e, f, \mu)I)\end{aligned} \tag{2.6}$$

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

$$\begin{aligned}\det(C - \xi I) = {} & (1 - a - b - c - d - e - f - \xi) \\ & \times (1 - a - b - c - d - e - f - 2g - 2h - \xi) \\ & \times (1 - a - b - c - d - e - f - 2g - 2k - \xi) \\ & \times (1 - a - b - c - d - e - f - 2h - 2k - \xi).\end{aligned}$$

Therefore, by substituting in (2.6) 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 (Appendix A): nine eigenvalues depend on two parameters, 27 eigenvalues, on four parameters and 27 eigenvalues, on six parameters.

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 (results not shown).

The formula (2.4) 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}, determines 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. The matrix $R = Q \cdot e^{Dt} \cdot Q^{-1}$ is given in Appendix B for the reader who wants to develop different evolutionary applications by varying the choice of $P(0)$. Two applications, one with the common circular code and the other with the codon phylogenetic distance, are given in the Section 3.

3. Results

3.1. Time inversion

The formula $P(t) = Q \cdot e^{Dt} \cdot Q^{-1} \cdot P(0)$ (2.4) gives the trinucleotide probabilities at the evolutionary time t from their past ones $P(0)$. By expressing $P(0)$ as a function of $P(t)$ in (2.4), then $P(0) = Q \cdot e^{-Dt} \cdot Q^{-1} \cdot P(t)$. Therefore, the formula $\widetilde{P}(t) = Q \cdot e^{-Dt} \cdot Q^{-1} \cdot \widetilde{P}(0)$, by replacing t by $-t$ in (2.4), gives the past trinucleotide probabilities from their actual ones $\widetilde{P}(0)$, i.e., by inverting the direction of the evolutionary time.

3.2. Time steps

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

3.3. Analytical solution of the common circular code

3.3.1. Identification

In 1996, a simple statistical analysis of the trinucleotide occurrence in the three frames of genes has identified the same subset $\mathcal{C}$ of 20 trinucleotides in the reading frames of two large and different gene populations of eukaryotes (26757 sequences, 11397678 trinucleotides) and prokaryotes (13686 sequences, 4708758 trinucleotides) (Arquès and Michel, 1996). This common trinucleotide subset $\mathcal{C} = \{AAC, AAT, ACC,$

$ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$ presents several strong biomathematical properties, in particular the property of circular code. Due to the law of large numbers, this subset $\mathcal{C}$ is (obviously) retrieved in these two gene populations with the actual statistical studies (results not shown). We briefly point out the property of circular code.

Notation 1. *$\mathbb{A}$ being a finite alphabet, $\mathbb{A}^*$ denotes the words over $\mathbb{A}$ of finite length including the empty word ϵ of length 0 and $\mathbb{A}^+$, the words over $\mathbb{A}$ of finite length greater or equal to 1. Let w_1w_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 $n = m$, $r = \epsilon$ and $x_i = y_i$, $1 \leq i \leq n$ (Lassez, 1976; Berstel and Perrin, 1985).

A circular code allows the reading frames of genes to be retrieved. 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 a unique 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, the commas showing the way of decomposition. 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 of the circular code. Then, the minimal window length is the size of the longest ambiguous word which can be read in at least two frames, more one letter. 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.

The main properties of the common circular code $\mathcal{C}$ are reviewed in (Michel, 2007b): maximality, permutation, complementarity, $\mathcal{C}^3$ code, rarity, largest window length, higher frequency of "misplaced" trinucleotides, flexibility, evolutionary properties and common occurrence in both eukaryotic and prokaryotic genes. In genes, the circular code information for retrieving the reading frames is added to the classical genetic code for coding the amino acids (Michel, 2007b).

3.3.2. Evolution model

The observation of a common trinucleotide set $\mathcal{C}$ in the reading frames of various genes from the two largest domains, the eukaryotes and the prokaryotes, is the basis of our development of an evolution 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 these primitive genes will be constructed by trinucleotides of $\mathcal{C}$, the mathematical model will be based on a trinucleotide mutation matrix 64×64. The evolution model proposed will be based on two processes. A construction process ($t = 0$) will generate 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 ($t > 0$) will transform these primitive genes into simulated actual ones. Random substitutions with different rates in the three sites of the 20 trinucleotides of $\mathcal{C}$ modelled by nine substitution parameters will generate other trinucleotides and distribute them according to an unbalanced way in the hope of retrieving the statistical distribution of $\mathcal{C}$ in the actual genes.

This stochastic approach with exact solutions in the past-present evolutionary sense relies on a gene evolution physical model by constructing simulated sequences and then by applying random substitutions to them. Note that in such a physical model, a population of large sequences must be simulated in the statistical analysis, which is time consuming for obtaining computer results with a good approximation.

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

The 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

$$P(X, t) = \sum_{i \in X} P_i(t) \tag{3.1}$$

with $P_i(t)$ defined by (2.4). As the code $\mathcal{C}$ cannot contain a trinucleotide $\mathcal{T}_{id} = \{AAA, CCC, GGG, TTT\}$ by definition (explained in Michel, 2007b), 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 Appendix A

$$\begin{aligned} P(\mathcal{C}, t) = \frac{\sum_{i \in \mathcal{C}} P_i(t)}{\sum_{i \in \mathcal{T} - \mathcal{T}_{id}} P_i(t)} = \frac{1}{2D}(100 &+ 4e^{\lambda_2 t} + 9e^{\lambda_3 t} + 25e^{\lambda_4 t} \\ &+ 36e^{\lambda_6 t} + 4e^{\lambda_8 t} + 9e^{\lambda_9 t} + 25e^{\lambda_{10} t} + 4e^{\lambda_{13} t} + 4e^{\lambda_{14} t} \\ &+ e^{\lambda_{15} t} + e^{\lambda_{16} t} + e^{\lambda_{17} t} + e^{\lambda_{18} t} + e^{\lambda_{19} t} + e^{\lambda_{20} t} + 4e^{\lambda_{22} t} \\ &+ e^{\lambda_{23} t} + e^{\lambda_{24} t} + e^{\lambda_{25} t} + e^{\lambda_{26} t} + 4e^{\lambda_{27} t} + 16e^{\lambda_{28} t} \\ &+ e^{\lambda_{30} t} + e^{\lambda_{31} t} + e^{\lambda_{33} t} + e^{\lambda_{34} t} + 4e^{\lambda_{35} t} + e^{\lambda_{36} t} \\ &+ e^{\lambda_{37} t} + e^{\lambda_{39} t} + e^{\lambda_{40} t} + 4e^{\lambda_{41} t} + e^{\lambda_{42} t} + e^{\lambda_{43} t} + e^{\lambda_{45} t} \end{aligned}$$

$$+ e^{\lambda_{46}t} + e^{\lambda_{47}t} + 4e^{\lambda_{49}t} + e^{\lambda_{50}t} + 16e^{\lambda_{52}t} + e^{\lambda_{53}t} + e^{\lambda_{56}t} + 4e^{\lambda_{57}t} + e^{\lambda_{59}t} + 16e^{\lambda_{60}t} + e^{\lambda_{62}t}) \quad (3.2)$$

with the denominator D

$$D = 150 + 2e^{\lambda_{14}t} + e^{\lambda_{18}t} - e^{\lambda_{25}t} + 4e^{\lambda_{28}t} + e^{\lambda_{33}t} - e^{\lambda_{37}t} + e^{\lambda_{43}t} - e^{\lambda_{45}t} + 2e^{\lambda_{49}t} - e^{\lambda_{53}t} + 2e^{\lambda_{57}t} + e^{\lambda_{59}t}.$$

Furthermore, if $a+b+c+d+e+f+g+h+k=1$, then $P(\mathcal{C},t)$ is equal to the formula (8) in Michel (2007a).

Property 1. *The initial probability $P(\mathcal{C},0)$ of the code $\mathcal{C}$ at the time $t=0$ can (obviously) be obtained from the analytical solution $P(\mathcal{C},t)$ with $t=0$ (3.2) or also by a simple probability calculus.*

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

Property 2. *The probability $P(\mathcal{C},t)$ of the code $\mathcal{C}$ at the limit time $t\to\infty$ can (obviously) be obtained from their limit study (3.2) or also by a simple probability calculus. Whatever $a,b,c,d,e,f,g,h,k\in]0,1[$, $\lim_{t\to\infty}P(\mathcal{C},t)=1/3$. 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 $\mathcal{T}_{id}$ being not considered).*

Property 3. *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 $P(\mathcal{C},t)$ (3.2) with $a=b=c=p/3$, $d=e=f=q/3$ and $g=h=k=r/3$*

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

with the denominator D_1

$$D_1 = 75 + 3e^{-(4/3)(p+r)t} + 2e^{-(4/3)(p+q+r)t}.$$

Furthermore, if $p+q+r=1$, then $P_1(\mathcal{C},t)$ is equal to the formula of Property 4 in Michel (2007a).

Property 4. *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, respectively) are the transition and the transversion rates in the 1st (2nd and 3rd, respectively) trinucleotide sites, respectively, is a particular case of $P(\mathcal{C},t)$ (3.2) with $a=u$, $b=c=v/2$, $d=w$, $e=f=x/2$, $g=y$ and $h=k=z/2$*

$$P_2(\mathcal{C},t) = \frac{1}{2D_2}(100 + 25e^{-2vt} + 13e^{-(2u+v)t} + e^{-2(v+x)t} + 36e^{-(2w+x)t} + 2e^{-(2u+v+2w+x)t} + 2e^{-(2v+2w+x)t} + 5e^{-(2u+v+2x)t} + e^{-(2v+2x+2y+z)t} + 25e^{-2zt} + 16e^{-2(v+z)t} + e^{-2(x+z)t} + e^{-(2u+v+2x+2z)t} + 13e^{-(2y+z)t} + 6e^{-(2u+v+2y+z)t} + 2e^{-(2w+x+2y+z)t} + 8e^{-(2u+v+2w+x+2y+z)t} + 22e^{-(2v+2w+x+2y+z)t} + 5e^{-(2v+2y+z)t} + 5e^{-(2x+2y+z)t} + 2e^{-(2u+v+2x+2y+z)t} + 5e^{-(2u+v+2z)t} + 2e^{-(2w+x+2z)t} + 22e^{-(2u+v+2w+x+2z)t})$$

with the denominator D_2

$$D_2 = 150 - e^{-2(v+x)t} + e^{-(2u+v+2w+x)t} + 4e^{-2(v+z)t} - e^{-2(x+z)t} + 2e^{-(2u+v+2y+z)t} + e^{-(2w+x+2y+z)t} + 3e^{-(2v+2w+x+2y+z)t} - 2e^{-(2u+v+2x+2y+z)t} + 3e^{-(2u+v+2w+x+2z)t}.$$

Furthermore, if $u+v+w+x+y+z=1$, then $P_2(\mathcal{C},t)$ is equal to the formula of Property 5 in Michel (2007a).

3.4. Codon phylogenetic distance

In order to derive a codon phylogenetic distance, we choose an initial vector $P(0)$ containing only one trinucleotide, e.g., *AAA*, i.e.,

$$P(0) = \begin{cases} P_1(0) = 1 \\ P_i(0) = 0 \ \forall i \in \{2,\ldots,64\}. \end{cases}$$

By convention, the index l_s, $l\in\{1,\ldots,4\}$ and $s\in\{1,\ldots,3\}$, represents the four nucleotides $\{A,C,G,T\}$ in alphabetical order in the three codon sites s and let P_{l_s}, be their associated evolutionary probabilities. Then,

$$P_{l_1} = \sum_{i=0}^{15} P_{i+16(l_1-1)+1}(t), \quad (3.3)$$

$$P_{l_2} = \sum_{i=0}^{15} P_{i \bmod 4 + 16\left\lfloor \frac{i}{4} \right\rfloor + 4(l_2-1)+1}(t), \quad (3.4)$$

$$P_{l_3} = \sum_{i=0}^{15} P_{4(i \bmod 4)+16\lfloor \frac{i}{4} \rfloor + l_3}(t). \quad (3.5)$$

The nine substitution parameters *a, b, c, d, e, f, g, h* and *k* are renamed here by considering their codon site: $a_s, s\in\{1,\ldots,3\}$, are the transition rates $A\leftrightarrow G$ and $C\leftrightarrow T$ in the three codon sites s, i.e., $a_1=a$, $a_2=d$ and $a_3=g$, b_s, $s\in\{1,\ldots,3\}$, are the transversion rates $A\leftrightarrow T$ and $C\leftrightarrow G$ in the three sites s, i.e., $b_1=b$, $b_2=e$ and $b_3=h$, and c_s, $s\in\{1,\ldots,3\}$, are the transversion rates $A\leftrightarrow C$ and $G\leftrightarrow T$ in the three sites s, i.e., $c_1=c$, $c_2=f$ and $c_3=k$. Let α_s, β_s and γ_s be the probabilities associated with the nucleotide differences between a codon site s of a 1st gene and the same codon site s of a 2nd gene: α_s, $s\in\{1,\ldots,3\}$, is the probability that the sth codon site of a 1st gene and the same sth codon site of a 2nd gene differ by the transitions $A\leftrightarrow G$ and $C\leftrightarrow T$, β_s, $s\in\{1,\ldots,3\}$, is the probability that the sth codon site of a 1st gene and the same sth codon site of a 2nd gene differ by the transversions $A\leftrightarrow T$

and $C \leftrightarrow G$, and γ_s, $s \in \{1, \ldots, 3\}$, is the probability that the sth codon site of a 1st gene and the same sth codon site of a 2nd gene differ by the transversions $A \leftrightarrow C$ and $G \leftrightarrow T$. Then, these nine probabilities can be expressed as a function of the substitution parameters a_s, b_s and c_s. Indeed,

$$\alpha_s = 2(P_{A_s} \times P_{G_s} + P_{C_s} \times P_{T_s}) = \frac{1}{4}(1 - e^{-4(a_s+b_s)t} - e^{-4(a_s+c_s)t} + e^{-4(b_s+c_s)t}),$$

$$\beta_s = 2(P_{A_s} \times P_{T_s} + P_{C_s} \times P_{G_s}) = \frac{1}{4}(1 - e^{-4(a_s+b_s)t} + e^{-4(a_s+c_s)t} - e^{-4(b_s+c_s)t}),$$

$$\gamma_s = 2(P_{A_s} \times P_{C_s} + P_{G_s} \times P_{T_s}) = \frac{1}{4}(1 + e^{-4(a_s+b_s)t} - e^{-4(a_s+c_s)t} - e^{-4(b_s+c_s)t})$$

with P_{A_s}, P_{C_s}, P_{G_s} and P_{T_s} obtained by the formulas (3.3), (3.4) and (3.5).

The phylogenetic distance, classically defined per site, is extended per codon of length $n = 3$. As there are nine substitution parameters per codon per time unit (see the matrices A, B and C) in each branch of the phylogenetic tree, the codon phylogenetic distance D_3 is defined as

$$D_3 = 2t\sum_{s=1}^{3}(a_s + b_s + c_s).$$

By solving a_s, b_s and c_s as a function of α_s, β_s and γ_s, then

$$D_3 = -\frac{1}{4}\sum_{s=1}^{3}(\ln(1 - 2\alpha_s - 2\beta_s) + \ln(1 - 2\alpha_s - 2\gamma_s) + \ln(1 - 2\beta_s - 2\gamma_s)) \quad (3.6)$$

with $\alpha_s + \beta_s < 1/2$, $\alpha_s + \gamma_s < 1/2$ and $\beta_s + \gamma_s < 1/2$.

Property 5. *By using a similar reasoning with mutation matrices of different sizes, the phylogenetic distance D_n associated with a word (sequence) of length n can easily be generalized from the distance D_3 (3.6)*

$$D_n = -\frac{1}{4}\sum_{s=1}^{n}(\ln(1 - 2\alpha_s - 2\beta_s) + \ln(1 - 2\alpha_s - 2\gamma_s) + \ln(1 - 2\beta_s - 2\gamma_s)).$$

Remark 2. The distance D_1 associated with a letter is

$$D_1 = -\frac{1}{4}(\ln(1 - 2\alpha - 2\beta) + \ln(1 - 2\alpha - 2\gamma) + \ln(1 - 2\beta - 2\gamma))$$

and is equal to the site distance formula (6) in Kimura (1981) (p. 455) which extends the site distance formulas with one and two substitution parameters (Jukes and Cantor, 1969; Kimura, 1980).

4. Discussion

A new analytical evolution 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, 1981; Takahata and Kimura, 1981) and based on the trinucleotide mutation matrices 64×64 with three, six and nine substitution parameters and with zero elements on its main diagonal (Arquès et al., 1998; Frey and Michel, 2006; Michel, 2007a). A first application of this model has allowed to generalize an evolutionary analytical solution of the common circular code $\mathcal{C}$ of eukaryotes and prokaryotes. The evolution of this code $\mathcal{C}$ cannot obviously be predicted without modelling as its solution is based on a sum of 46 exponential terms (formula (3.2)), each exponential term being a function of the time t and the nine substitutions parameters. A second application of this model has allowed to derive a codon phylogenetic distance D_3. This distance D_3 extends the classical site phylogenetic distance. According to formula (3.6), more the codons differ more its distance D_3 increases.

Other applications of this model can be applied to various problems. In particular, the eigenvalues given in Appendix A as well as the structure of the matrix $R = Q \cdot e^{Dt} \cdot Q^{-1}$ given in Appendix B can be directly used to develop other evolution 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.

Appendix A. The 64 eigenvalues λ_k of M of algebraic multiplicity 1

$$\lambda_1 = 0, \quad \lambda_2 = -2(a + b),$$
$$\lambda_3 = -2(a + c), \quad \lambda_4 = -2(b + c),$$
$$\lambda_5 = -2(d + e), \quad \lambda_6 = -2(d + f),$$
$$\lambda_7 = -2(e + f), \quad \lambda_8 = -2(g + h),$$
$$\lambda_9 = -2(g + k), \quad \lambda_{10} = -2(h + k),$$
$$\lambda_{11} = -2(a + b + d + e), \quad \lambda_{12} = -2(a + b + d + f),$$
$$\lambda_{13} = -2(a + b + e + f), \quad \lambda_{14} = -2(a + b + g + h),$$
$$\lambda_{15} = -2(a + b + g + k), \quad \lambda_{16} = -2(a + b + h + k),$$
$$\lambda_{17} = -2(a + c + d + e), \quad \lambda_{18} = -2(a + c + d + f),$$
$$\lambda_{19} = -2(a + c + e + f), \quad \lambda_{20} = -2(a + c + g + h),$$
$$\lambda_{21} = -2(a + c + g + k), \quad \lambda_{22} = -2(a + c + h + k),$$
$$\lambda_{23} = -2(b + c + d + e), \quad \lambda_{24} = -2(b + c + d + f),$$
$$\lambda_{25} = -2(b + c + e + f), \quad \lambda_{26} = -2(b + c + g + h),$$
$$\lambda_{27} = -2(b + c + g + k), \quad \lambda_{28} = -2(b + c + h + k),$$
$$\lambda_{29} = -2(d + e + g + h), \quad \lambda_{30} = -2(d + e + g + k),$$

$$\lambda_{31} = -2(d+e+h+k),\quad \lambda_{32} = -2(d+f+g+h),$$
$$\lambda_{33} = -2(d+f+g+k),\quad \lambda_{34} = -2(d+f+h+k),$$
$$\lambda_{35} = -2(e+f+g+h),\quad \lambda_{36} = -2(e+f+g+k),$$
$$\lambda_{37} = -2(e+f+h+k),$$
$$\lambda_{38} = -2(a+b+d+e+g+h),\quad \lambda_{39} = -2(a+b+d+e+g+k),\quad \lambda_{40} = -2(a+b+d+e+h+k),$$
$$\lambda_{41} = -2(a+b+d+f+g+h),\quad \lambda_{42} = -2(a+b+d+f+g+k),\quad \lambda_{43} = -2(a+b+d+f+h+k),$$
$$\lambda_{44} = -2(a+b+e+f+g+h),\quad \lambda_{45} = -2(a+b+e+f+g+k),\quad \lambda_{46} = -2(a+b+e+f+h+k),$$
$$\lambda_{47} = -2(a+c+d+e+g+h),\quad \lambda_{48} = -2(a+c+d+e+g+k),\quad \lambda_{49} = -2(a+c+d+e+h+k),$$
$$\lambda_{50} = -2(a+c+d+f+g+h),\quad \lambda_{51} = -2(a+c+d+f+g+k),\quad \lambda_{52} = -2(a+c+d+f+h+k),$$
$$\lambda_{53} = -2(a+c+e+f+g+h),\quad \lambda_{54} = -2(a+c+e+f+g+k),\quad \lambda_{55} = -2(a+c+e+f+h+k),$$
$$\lambda_{56} = -2(b+c+d+e+g+h),\quad \lambda_{57} = -2(b+c+d+e+g+k),\quad \lambda_{58} = -2(b+c+d+e+h+k),$$
$$\lambda_{59} = -2(b+c+d+f+g+h),\quad \lambda_{60} = -2(b+c+d+f+g+k),\quad \lambda_{61} = -2(b+c+d+f+h+k),$$
$$\lambda_{62} = -2(b+c+e+f+g+h),\quad \lambda_{63} = -2(b+c+e+f+g+k),\quad \lambda_{64} = -2(b+c+e+f+h+k).$$

Appendix B. The matrix $R = Q \cdot e^{Dt} \cdot Q^{-1}$

The square matrix $R = Q \cdot e^{Dt} \cdot Q^{-1}$ (64,64) can be defined by a square block matrix (4,4) whose four diagonal elements are formed by four identical square submatrices S_1(16,16) and whose 12 non-diagonal elements are formed by four square submatrices S_{17} (16,16), four square submatrices S_{33} (16,16) and four square submatrices S_{49} (16,16) as follows

$$R = Q \cdot e^{Dt} \cdot Q^{-1} = \frac{1}{64}\begin{pmatrix} & 1\ldots16 & 17\ldots32 & 33\ldots48 & 49\ldots64 \\ 1\ldots16 & S_1 & S_{17} & S_{33} & S_{49} \\ 17\ldots32 & S_{17} & S_1 & S_{49} & S_{33} \\ 33\ldots48 & S_{33} & S_{49} & S_1 & S_{17} \\ 49\ldots64 & S_{49} & S_{33} & S_{17} & S_1 \end{pmatrix}.$$

A square submatrix S_i (16,16) can again be defined by a square block matrix (4,4) whose four diagonal elements are formed by four identical square submatrices T_i (4,4) and whose 12 non-diagonal elements are formed by four square submatrices T_{i+4} (4,4), four square submatrices T_{i+8} (4,4) and four square submatrices T_{i+12} (4,4) as follows

$$S_i = \begin{pmatrix} T_i & T_{i+4} & T_{i+8} & T_{i+12} \\ T_{i+4} & T_i & T_{i+12} & T_{i+8} \\ T_{i+8} & T_{i+12} & T_i & T_{i+4} \\ T_{i+12} & T_{i+8} & T_{i+4} & T_i \end{pmatrix}.$$

Finally, the square submatrix T_i (4,4) is defined as follows

$$T_i = \begin{pmatrix} \mathcal{F}_i & \mathcal{F}_{i+1} & \mathcal{F}_{i+2} & \mathcal{F}_{i+3} \\ \mathcal{F}_{i+1} & \mathcal{F}_i & \mathcal{F}_{i+3} & \mathcal{F}_{i+2} \\ \mathcal{F}_{i+2} & \mathcal{F}_{i+3} & \mathcal{F}_i & \mathcal{F}_{i+1} \\ \mathcal{F}_{i+3} & \mathcal{F}_{i+2} & \mathcal{F}_{i+1} & \mathcal{F}_i \end{pmatrix}$$

where the function $\mathcal{F}_i$ associated with the ith line of R is defined as

$$\mathcal{F}_i = \sum_{j=1}^{64} \delta_{ij} e^{\lambda_j t}$$

with the eigenvalues λ_j defined in Appendix A and the constant δ_{ij}, by the following matrix δ (Fig. B.1)

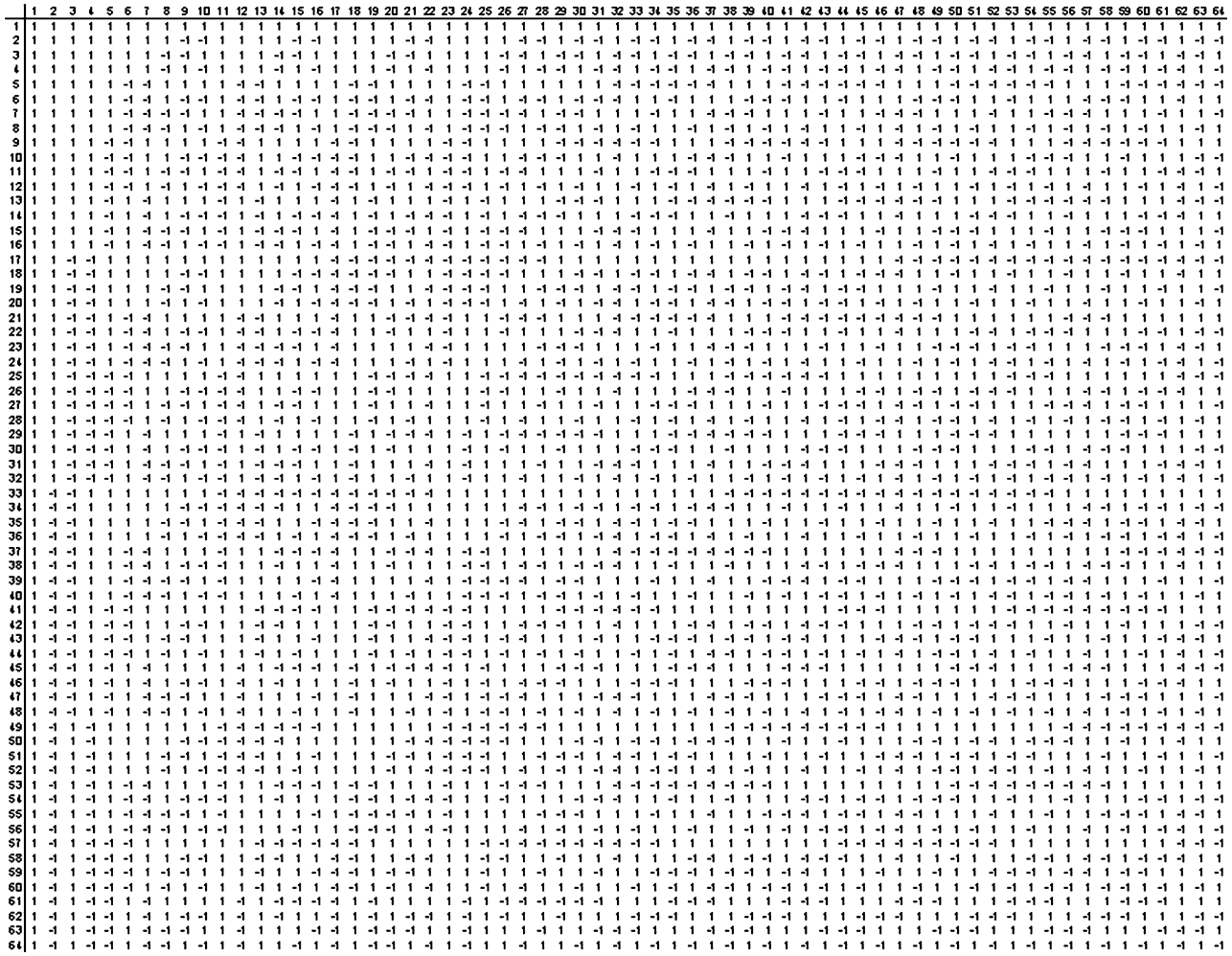

Fig. B.1. The matrix δ.

References

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.

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

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.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, 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. U.S.A. 78, 454–458.

Lange, K., 2005. Applied Probability. Springer-Verlag, New York.

Lassez, J.-L., 1976. Circular codes and synchronization. Int. J. Comput. Syst. Sci. 5, 201–208.

Michel, C.J., 2007a. An analytical model of gene evolution with 9 mutation parameters: an application to the amino acids coded by the common circular code. Bull. Math. Biol.

Michel, C.J., 2007b. A 2006 review of circular codes in genes. Comp. Math. Appl.

Takahata, N., Kimura, M., 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics 98, 641–657.