Available online at www.sciencedirect.com

ScienceDirect

Journal of Theoretical Biology 249 (2007) 271–277

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

# Evolution probabilities and phylogenetic distance of dinucleotides

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 18 February 2007; received in revised form 18 July 2007; accepted 20 July 2007
Available online 11 August 2007

**Abstract**

We develop here an analytical evolution model based on a dinucleotide mutation matrix 16 × 16 with six substitution parameters associated with the three types of substitutions in the two dinucleotide sites. It generalizes the previous models based on the nucleotide mutation matrices 4 × 4. It determines at some time *t* the exact occurrence probabilities of dinucleotides mutating randomly according to these six substitution parameters. Furthermore, several properties and two applications of this model allow to derive 16 evolutionary analytical solutions of dinucleotides and also a dinucleotide phylogenetic distance. Finally, based on this mathematical model, the SED (Stochastic Evolution of Dinucleotides) web server has been developed for deriving evolutionary analytical solutions of dinucleotides.
© 2007 Elsevier Ltd. All rights reserved.

*Keywords:* Stochastic evolution model; Analytical solution; Evolutionary distance

## 1. Introduction

Models of gene evolution were initially developed on the basis of nucleotide information, the first model was proposed by Jukes and Cantor (1969). DNA sequencing has revealed that the structure of the different genome regions are based on a variety of motifs of different sizes: dinucleotides, trinucleotides, oligonucleotides, either on a 2-letter alphabet, e.g. the purine/pyrimidine alphabet, or on the classical 4-letter alphabet. After the creation of the first gene databases (EMBL, Genbank), these biological results have been widely studied by different signal processing methods, such as the correlation functions, the power spectrum, the Fourier transforms, etc., in order to identify the statistical properties of these motifs, e.g. periodicities and subperiodicities, global and local maxima, etc. For example, a periodicity modulo 2 associated with dinucleotides has been identified in introns (Arquès and Michel, 1987; Konopka et al., 1987) and in the 5′ and 3′ regions surrounding the eukaryotic genes (Arquès and Michel, 1990), a periodicity modulo 3 associated with trinucleotides has been observed in genes (Shepherd, 1981; Michel, 1986), etc. On the other hand, a great number of mutation studies have shown that the neighboring bases influence the types and rates of mutation at a given sequence position, e.g. Hess et al. (1994). After an analysis phase of the statistical properties of motifs, their evolutionary properties have been studied by extending the nucleotide evolution models to the motif evolution models, in particular to the trinucleotide (Arquès and Michel, 1993) and dinucleotide (Arquès and Michel, 1995) ones. The variety and the complexity of these models increase regularly and the dinucleotide evolution model developed here is based on six substitution parameters.

A new stochastic evolution model will determine at some time *t* the occurrence probabilities of dinucleotides mutating randomly according to several types of substitutions in the dinucleotide sites. It is based on a dinucleotide mutation matrix 16 × 16 with six substitution parameters and with non-zero elements on the main diagonal. Occurrence probabilities of dinucleotide sets can obviously be deduced from this approach. This model with six substitution parameters associated with the three types of substitutions in the two dinucleotide sites generalizes several previous models based on the nucleotide mutation matrices 4 × 4, in particular with one substitution parameter (Jukes and Cantor, 1969), two parameters

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

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.07.032

(transitions and transversions) (Kimura, 1980), three parameters (Kimura, 1981), four parameters (Takahata and Kimura, 1981) and six parameters (Kimura, 1981), and the dinucleotide model with one substitution parameter (Arquès and Michel, 1995). Other approaches have been developed later for studying evolution of dinucleotides. For example, a computer simulation approach (construction of simulated genes and then applying random mutations) has been proposed by Fryxell and Zuckerkandl (2000), a discrete version with time steps $\Delta t/L$ where $\Delta t$ is the time increment and $L$, the length of the sequence, has been developed in Arndt et al. (2002).

Two types of results are presented in this paper:

(i) A mathematical model of gene evolution with six substitution parameters is developed: $a$ and $d$ are the rates of transitions $A \longleftrightarrow G$ (a substitution from one purine $\{A, G\}$ to the other) and $C \longleftrightarrow T$ (a substitution from one pyrimidine $\{C, T\}$ to the other) in the two sites, respectively, $b$ and $e$ are the rates of transversions (a substitution from a purine to a pyrimidine, or reciprocally) $A \longleftrightarrow T$ and $C \longleftrightarrow G$ in the two sites, respectively, and $c$ and $f$ are the rates of transversions $A \longleftrightarrow C$ and $G \longleftrightarrow T$ in the two sites, respectively.

(ii) Several properties and two applications of this model allow to derive 16 evolutionary analytical solutions of dinucleotides and also a dinucleotide phylogenetic distance.

## 2. Mathematical model

The mathematical model will determine at an evolutionary time $t$ the occurrence probabilities $P(t)$ of the 16 dinucleotides mutating according to six real substitution parameters $a$, $b$, $c$, $d$, $e$ and $f$: $a$ and $d$ are the transition rates $A \longleftrightarrow G$ and $C \longleftrightarrow T$ in the two sites, respectively, $b$ and $e$ are the transversion rates $A \longleftrightarrow T$ and $C \longleftrightarrow G$ in the two sites, respectively, and $c$ and $f$ are the transversion rates $A \longleftrightarrow C$ and $G \longleftrightarrow T$ in the two sites, respectively.

By convention, the indexes $i, j \in \{1, \ldots, 16\}$ represent the 16 dinucleotides $\mathscr{D} = \{AA, \ldots, TT\}$ in alphabetical order. Let $P(j \to i)$ be the substitution probability of a dinucleotide $j$ into a dinucleotide $i$. 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 $\boldsymbol{T}$ is assumed to be enough small that a dinucleotide cannot mutate successively two times during $\boldsymbol{T}$. Otherwise, it is given as a function of the six substitution rates $a$, $b$, $c$, $d$, $e$ and $f$. For example with the dinucleotide $AA$ associated with $i = 1$, $P(CA \to AA) = c$, $P(GA \to AA) = a$, $P(TA \to AA) = b$, $P(AC \to AA) = f$, $P(AG \to AA) = d$, $P(AT \to AA) = e$ and $P(j \to AA) = 0$ with $j \notin \{AC, AG, AT, CA, GA, TA\}$.

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

$$P_i(t + \boldsymbol{T}) - P_i(t) = \alpha \boldsymbol{T} \sum_{j=1}^{16} P(j \to i) P_j(t) - \alpha \boldsymbol{T} P_i(t),$$

where $\alpha$ is the probability that a dinucleotide is subjected to one substitution during $\boldsymbol{T}$. By rescaling time, we can assume that $\alpha = 1$, i.e. there is one substitution per dinucleotide per time interval. Then,

$$\begin{aligned} &P_i(t + \boldsymbol{T}) - P_i(t) \\ &\quad = \boldsymbol{T} \sum_{j=1}^{16} P(j \to i) P_j(t) - \boldsymbol{T} P_i(t) \\ &\quad = \boldsymbol{T} \sum_{\substack{j=1 \\ j \neq i}}^{16} P(j \to i) P_j(t) + \boldsymbol{T} P(i \to i) P_i(t) - \boldsymbol{T} P_i(t) \\ &\quad = \boldsymbol{T} \sum_{\substack{j=1 \\ j \neq i}}^{16} P(j \to i) P_j(t) + \boldsymbol{T} \left( 1 - \sum_{\substack{j=1 \\ j \neq i}}^{16} P(j \to i) \right) P_i(t) \\ &\quad - \boldsymbol{T} P_i(t). \end{aligned} \tag{2.1}$$

The formula (2.1) leads to

$$\lim_{\boldsymbol{T} \to 0} \frac{P_i(t + \boldsymbol{T}) - P_i(t)}{\boldsymbol{T}} = P_i'(t) = \sum_{j=1}^{16} P(j \to i) P_j(t) - P_i(t) \tag{2.2}$$

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

By considering the column vector $P(t) = [P_i(t)]_{1 \leqslant i \leqslant 16}$ made of the 16 $P_i(t)$ and the mutation matrix $A$ (16, 16) of the 256 dinucleotide substitution probabilities $P(j \to i)$, the differential equation (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 $\cdot$, the matrix product.

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

$$A = \left( \begin{array}{ccccc} & 1 \cdots 4 & 5 \cdots 8 & 9 \cdots 12 & 13 \cdots 16 \\ 1 \cdots 4 & B & cI & aI & bI \\ 5 \cdots 8 & cI & B & bI & aI \\ 9 \cdots 12 & aI & bI & B & cI \\ 13 \cdots 16 & bI & aI & cI & B \end{array} \right).$$

The index ranges $\{1, \ldots, 4\}$, $\{5, \ldots, 8\}$, $\{9, \ldots, 12\}$ and $\{13, \ldots, 16\}$ are associated with the dinucleotides $\{AA, \ldots, AT\}$, $\{CA, \ldots, CT\}$, $\{GA, \ldots, GT\}$ and $\{TA, \ldots, TT\}$, respectively. The square submatrix $B$ (4,4) is equal to

$$B = \begin{pmatrix} n & f & d & e \\ f & n & e & d \\ d & e & n & f \\ e & d & f & n \end{pmatrix}$$

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

The mutation matrix $A$ is a doubly stochastic and positive matrix.

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

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

with

$$M = A - I.$$

As the six 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 $\lambda_k$ 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)

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

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

The eigenvalues $\lambda_k$ of $M$ are deduced from the eigenvalues $\mu_k$ of $A$ such that $\lambda_k = \mu_k - 1$. The eigenvalues $\mu_k$ 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

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

After linear combinations, the determinant $\det(B - \nu I)$ is equal to

$$\begin{aligned} \det(B - \nu I) = {} & (1 - a - b - c - \nu) \\ & \times (1 - a - b - c - 2d - 2e - \nu) \\ & \times (1 - a - b - c - 2d - 2f - \nu) \\ & \times (1 - a - b - c - 2e - 2f - \nu). \end{aligned}$$

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 $\det(A - \mu I)$ is obtained and then, the eigenvalues $\lambda_k$ of $M$ are deduced. There are 16 eigenvalues $\lambda_k$ of $M$ of algebraic multiplicity 1: six eigenvalues depend on two parameters and nine eigenvalues, on four parameters

$$\begin{aligned} & \lambda_1 = 0, \\ & \lambda_2 = -2(a + b), \lambda_3 = -2(a + c), \lambda_4 = -2(b + c), \\ & \lambda_5 = -2(d + e), \lambda_6 = -2(d + f), \lambda_7 = -2(e + f), \\ & \lambda_8 = -2(a + b + d + e), \lambda_9 = -2(a + b + d + f), \\ & \lambda_{10} = -2(a + b + e + f), \lambda_{11} = -2(a + c + d + e), \\ & \lambda_{12} = -2(a + c + d + f), \lambda_{13} = -2(a + c + e + f), \\ & \lambda_{14} = -2(b + c + d + e), \lambda_{15} = -2(b + c + d + f), \\ & \lambda_{16} = -2(b + c + e + f). \end{aligned} \tag{2.6}$$

The 16 eigenvectors of $M$ associated with these 16 eigenvalues $\lambda_k$ computed by formal calculus can be put in a form independent of $a$, $b$, $c$, $d$, $e$ and $f$ (results not shown).

The formula (2.4) with the initial probability vector $P(0)$ before the substitution process ($t = 0$), the diagonal matrix $\mathrm{e}^{Dt}$ of exponential eigenvalues $\mathrm{e}^{\lambda_k t}$ of $M$, its eigenvector matrix $Q$ and its inverse $Q^{-1}$, determine the 16 dinucleotide probabilities $P_i(t)$ after $t$ substitutions as a function of the six parameters $a$, $b$, $c$, $d$, $e$ and $f$. The matrix $R = Q \cdot \mathrm{e}^{Dt} \cdot Q^{-1}$ is given in Appendix for the reader who wants to develop different evolutionary applications by varying the choice of $P(0)$. Several properties and two applications of this model with the 16 evolutionary analytical solutions of dinucleotides and a dinucleotide phylogenetic distance are given in Section 3.

## 3. Results

### 3.1. Time inversion

The formula $P(t) = Q \cdot \mathrm{e}^{Dt} \cdot Q^{-1} \cdot P(0)$ (2.4) gives the dinucleotide 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 $P(0) = Q \cdot \mathrm{e}^{-Dt} \cdot Q^{-1} \cdot P(t)$. Therefore, the formula

$$\widetilde{P}(t) = Q \cdot \mathrm{e}^{-Dt} \cdot Q^{-1} \cdot \widetilde{P}(0) \tag{3.1}$$

by replacing $t$ by $-t$ in (2.4), gives the past dinucleotide 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 dinucleotide 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 $P(t_2) = Q \cdot \mathrm{e}^{D(t_2 - t_1)} \cdot Q^{-1} \cdot P(t_1)$.

### 3.3. Properties

If $n = 0$, i.e. $a + b + c + d + e + f = 1$, then the substitution probability $P(i \to i)$ of a dinucleotide $i$ into itself is impossible: $P(i \to i) = 0$ is associated with zero elements on the main diagonal of the dinucleotide mutation matrix $A$.

The dinucleotide probability sum $\sum_{i=1}^{16} P_i(t) = 1$ whatever $a$, $b$, $c$, $d$, $e$ and $f$ in the range $[0, 1]$ and whatever the evolutionary time $t$.

The given initial probabilities $P(0)$ of the 16 dinucleotides at the time $t = 0$ can (obviously) be obtained from their analytical solutions $P(t)$ with $t = 0$.

The probabilities $P(t)$ of the 16 dinucleotides at the limit time $t \to \infty$ can (obviously) be obtained from their limit study or also by a simple probability calculus. Indeed, whatever $a$, $b$, $c$, $d$, $e$ and $f$ in the range $]0, 1[$, $\lim_{t\to\infty} P_i(t) = \frac{1}{16}$ as the 16 dinucleotides $i$ occur with the same probability when the evolutionary time $t \to \infty$. When one (or more) substitution has a rate equal to 0, some dinucleotides $i$ may be either not generated or generated without equiprobability and $\lim_{t\to\infty} P_i(t) \neq \frac{1}{16}$ (not detailed, an example of probability calculus can be found with the property 3 in Michel, 2007).

The analytical formulas $P_1(t)$ of the 16 dinucleotides as a function of the two substitution rates $p$ and $q$ associated with the two dinucleotide sites, respectively, are particular cases of $P(t)$ (2.4) with $a = b = c = p/3$ and $d = e = f = q/3$.

The analytical formulas $P_2(t)$ of the 16 dinucleotides as a function of the four substitution rates $u$, $v$, $w$ and $x$ such that $u$ and $v$ ($w$ and $x$ resp.) are the transition and the transversion rates in the 1st (2nd resp.) dinucleotide sites, respectively, are particular cases of $P(t)$ (2.4) with $a = u$, $b = c = v/2$, $d = w$ and $e = f = x/2$.

The analytical formulas $P(t)$ (2.4) of the 16 dinucleotides obviously allows to deduce the occurrence probability $P(X, t)$ of a dinucleotide set $X$ by summing the probabilities of the dinucleotides $i$ belonging to the set $X$, i.e. $P(X, t) = \sum_{i \in X} P_i(t)$.

The stochastic model leads to exact solutions. In contrast, a gene evolution physical model constructing and transforming simulated sequences, leads to approximate solutions. It also requires a computer time consuming calculation. Indeed, for approximating analytical solutions by computer simulation (construction of simulated genes and then applying random mutations according to the substitution rates $a$, $b$, $c$, $d$, $e$ and $f$) several hours or even a few days can be necessary with a PC, in particular when some substitutions rates are closed to their limits (0 and 1) and/or close to each other and/or when a high decimal precision for the formula $P(t)$, e.g. a probability with 3 decimals, is required.

Two applications of this stochastic model are given by choosing an initial probability vector $P(0)$ containing only one dinucleotide, e.g. $AA$, i.e.

$$P(0) = \begin{cases} P_1(0) = 1, \\ P_i(0) = 0, \quad \forall i \in \{2, \ldots, 16\}. \end{cases} \tag{3.2}$$

### 3.4. Evolutionary analytical solutions of dinucleotides

The formula $P(t)$ (2.4) with the initial probability vector $P(0)$ (3.2) allows several evolutionary analytical solutions of dinucleotides to be deduced. They can be expressed by the following general equation:

$$\begin{aligned} P(BB \to B'B'') = {} & \tfrac{1}{16} e^{-2(a+b+c+d+e+f)t} \\ & \times (\delta_a e^{2at} + \delta_b e^{2bt} + \delta_c e^{2ct} + \delta_p e^{2(a+b+c)t}) \\ & \times (\delta_d e^{2dt} + \delta_e e^{2et} + \delta_f e^{2ft} + \delta_q e^{2(d+e+f)t}) \end{aligned}$$

with $\delta_a, \delta_b, \delta_c, \delta_d, \delta_e, \delta_f, \delta_p, \delta_q \in \{-1, 1\}$. The different solutions are

| $P(BB \to B'B'')$ with | $\delta_a$ | $\delta_b$ | $\delta_c$ | $\delta_p$ | $\delta_d$ | $\delta_e$ | $\delta_f$ | $\delta_q$ |
|---|---|---|---|---|---|---|---|---|
| $B' = B'' = B$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $P(B \to B') = a$ | 1 | −1 | −1 | 1 | 1 | 1 | 1 | 1 |
| $P(B \to B') = b$ | −1 | 1 | −1 | 1 | 1 | 1 | 1 | 1 |
| $P(B \to B') = c$ | −1 | −1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $P(B \to B'') = d$ | 1 | 1 | 1 | 1 | 1 | −1 | −1 | 1 |
| $P(B \to B'') = e$ | 1 | 1 | 1 | 1 | −1 | 1 | −1 | 1 |
| $P(B \to B'') = f$ | 1 | 1 | 1 | 1 | −1 | −1 | 1 | 1 |
| $P(B \to B') = a$ and<br>$P(B \to B'') = d$ | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 |
| $P(B \to B') = a$ and<br>$P(B \to B'') = e$ | 1 | −1 | −1 | 1 | −1 | 1 | −1 | 1 |
| $P(B \to B') = a$ and<br>$P(B \to B'') = f$ | 1 | −1 | −1 | 1 | −1 | −1 | 1 | 1 |
| $P(B \to B') = b$ and<br>$P(B \to B'') = d$ | −1 | 1 | −1 | 1 | 1 | −1 | −1 | 1 |
| $P(B \to B') = b$ and<br>$P(B \to B'') = e$ | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 |
| $P(B \to B') = b$ and<br>$P(B \to B'') = f$ | 1 | −1 | 1 | −1 | 1 | 1 | −1 | −1 |
| $P(B \to B') = c$ and<br>$P(B \to B'') = d$ | −1 | −1 | 1 | 1 | 1 | −1 | −1 | 1 |
| $P(B \to B') = c$ and<br>$P(B \to B'') = e$ | 1 | 1 | −1 | −1 | 1 | −1 | 1 | −1 |
| $P(B \to B') = c$ and<br>$P(B \to B'') = f$ | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 |

### 3.5. Dinucleotide phylogenetic distance

By convention, the index $l_s$, $l \in \{1, \ldots, 4\}$ and $s \in \{1, 2\}$, represents the four nucleotides $\{A, C, G, T\}$ in alphabetical order in the two dinucleotide sites $s$ and let $P_{l_s}$, be their associated evolutionary probabilities. Then,

$$P_{l_1} = \sum_{i=0}^{3} P_{i+4(l_1-1)+1}(t), \tag{3.3}$$

$$P_{l_2} = \sum_{i=0}^{3} P_{4(i \bmod 4)+l_2}(t). \tag{3.4}$$

The six substitution parameters $a$, $b$, $c$, $d$, $e$ and $f$ are renamed here by considering their dinucleotide site: $a_s$, $s \in \{1,2\}$, are the transition rates $A \longleftrightarrow G$ and $C \longleftrightarrow T$ in the two dinucleotide sites $s$, i.e. $a_1 = a$ and $a_2 = d$, $b_s$, $s \in \{1,2\}$, are the transversion rates $A \longleftrightarrow T$ and $C \longleftrightarrow G$ in the two sites $s$, i.e. $b_1 = b$ and $b_2 = e$, and $c_s$, $s \in \{1,2\}$, are the transversion rates $A \longleftrightarrow C$ and $G \longleftrightarrow T$ in the two sites $s$, i.e. $c_1 = c$ and $c_2 = f$. Let $\alpha_s$, $\beta_s$ and $\gamma_s$ be the probabilities associated with the nucleotide differences between a dinucleotide site $s$ of a 1st gene and the same dinucleotide site $s$ of a 2nd gene: $\alpha_s$, $s \in \{1,2\}$, is the probability that the $s$th dinucleotide site of a 1st gene and the same $s$th dinucleotide site of a 2nd gene differ by the transitions $A \longleftrightarrow G$ and $C \longleftrightarrow T$, $\beta_s$, $s \in \{1,2\}$, is the probability that the $s$th dinucleotide site of a 1st gene and the same $s$th dinucleotide site of a 2nd gene differ by the transversions $A \longleftrightarrow T$ and $C \longleftrightarrow G$, and $\gamma_s$, $s \in \{1,2\}$, is the probability that the $s$th dinucleotide site of a 1st gene and the same $s$th dinucleotide site of a 2nd gene differ by the transversions $A \longleftrightarrow C$ and $G \longleftrightarrow T$. Then, these six probabilities can be expressed as a function of the substitution parameters $a_s$, $b_s$ and $c_s$. Indeed,

$$\begin{aligned}\alpha_s &= 2(P_{A_s} \times P_{G_s} + P_{C_s} \times P_{T_s}) \\ &= \tfrac{1}{4}(1 - \mathrm{e}^{-4(a_s+b_s)t} - \mathrm{e}^{-4(a_s+c_s)t} + \mathrm{e}^{-4(b_s+c_s)t}),\end{aligned}$$

$$\begin{aligned}\beta_s &= 2(P_{A_s} \times P_{T_s} + P_{C_s} \times P_{G_s}) \\ &= \tfrac{1}{4}(1 - \mathrm{e}^{-4(a_s+b_s)t} + \mathrm{e}^{-4(a_s+c_s)t} - \mathrm{e}^{-4(b_s+c_s)t}),\end{aligned}$$

$$\begin{aligned}\gamma_s &= 2(P_{A_s} \times P_{C_s} + P_{G_s} \times P_{T_s}) \\ &= \tfrac{1}{4}(1 + \mathrm{e}^{-4(a_s+b_s)t} - \mathrm{e}^{-4(a_s+c_s)t} - \mathrm{e}^{-4(b_s+c_s)t}).\end{aligned}$$

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

The phylogenetic distance, classically defined per site, is extended per dinucleotide of length $n = 2$. As there are six substitution parameters per dinucleotide per time unit (see the matrices $A$ and $B$) in each branch of the phylogenetic tree, the dinucleotide phylogenetic distance $D_2$ is defined as

$$D_2 = 2t\sum_{s=1}^{2}(a_s + b_s + c_s).$$

By solving $a_s$, $b_s$ and $c_s$ as a function of $\alpha_s$, $\beta_s$ and $\gamma_s$, then

$$\begin{aligned}D_2 = -\frac{1}{4}\sum_{s=1}^{2}[&\ln(1 - 2\alpha_s - 2\beta_s) \\ &+ \ln(1 - 2\alpha_s - 2\gamma_s) + \ln(1 - 2\beta_s - 2\gamma_s)] \qquad (3.5)\end{aligned}$$

with $\alpha_s + \beta_s < \frac{1}{2}$, $\alpha_s + \gamma_s < \frac{1}{2}$ and $\beta_s + \gamma_s < \frac{1}{2}$.

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 $D_2$

$$\begin{aligned}D_n = -\frac{1}{4}\sum_{s=1}^{n}[&\ln(1 - 2\alpha_s - 2\beta_s) \\ &+ \ln(1 - 2\alpha_s - 2\gamma_s) + \ln(1 - 2\beta_s - 2\gamma_s)].\end{aligned}$$

The distance $D_1$ associated with a letter is

$$\begin{aligned}D_1 = -\frac{1}{4}[&\ln(1 - 2\alpha - 2\beta) + \ln(1 - 2\alpha - 2\gamma) \\ &+ \ln(1 - 2\beta - 2\gamma)]\end{aligned}$$

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).

### 3.6. Remarks

Let the dinucleotide $i$ be composed of the two letters $j$ and $k$ such that $i = jk$. For some particular dinucleotide initial probability vectors $P(0)$, then

$$P_i(t) = Q_j(t)Q_k(t), \qquad (3.6)$$

where $Q_j(t)$ and $Q_k(t)$ are the occurrence probabilities of the nucleotides $j$ and $k$ in the 1st and 2nd dinucleotide sites, respectively, at the time $t$ with nucleotide initial probability vectors $Q_j(0)$ and $Q_k(0)$ deduced from $P_i(0)$. In the general case, the relation (3.6) cannot be applied and the formula (2.4) must be used.

The phylogenetic distance $D_2$ (resp. its generalization $D_n$) is a sum of 2 (resp. $n$) site distances. This result, which appears obvious a posteriori by using a dinucleotide (resp. $n$-nucleotide) matrix, would have been tedious to prove by using 2 (resp. $n$) nucleotide mutation matrices $(4,4)$ with different substitution parameters and such that one substitution per time interval occurs.

## 4. Development of the SED (Stochastic Evolution of Dinucleotides) web server

The SED (Stochastic Evolution of Dinucleotides) web server is a web application for deriving evolutionary analytical solutions of dinucleotides based on the mathematical model developed here. It will be freely available at http://dpt-info.u-strasbg.fr/~michel/.

The SED server has been implemented with the formal calculus software Mathematica (version 5.2) and web-Mathematica (version 2) for adding interactive computations (calculation and visualization) to the web.

The SED server takes as input several evolution model options which can be specified by the user. The first option is the choice of the evolutionary time sense of the model which can be either direct (past-present) and based on the formula (2.4) or inverse (present-past) and based on the formula (3.1). The second option is the number of substitution parameters, either 6 ($a$, $b$, $c$, $d$, $e$ and $f$) or 4 ($u$, $v$, $w$ and $x$, see Section 3.3) or 2 ($p$ and $q$, see

Section 3.3). The third option is the choice of the initial probability vector $P(0)$ which can be given in rational form (e.g. "$\frac{1}{3}$") or in real form (e.g. "0.3333"). A function returns the sum of the probability vector. This sum is given either in rational form (e.g. "1") if $P(0)$ contains all values in rational form, or in real form (e.g. "1.") if $P(0)$ contains a value in real form. It must be equal to 1 for starting the computation of analytical solutions. This function is also useful for completing the probability vector $P(0)$ such that its sum is equal to 1.

The output of the SED server gives the analytical solutions of the 16 dinucleotides as a function of the chosen number of parameters and in rational form if $P(0)$ is rational or in real form otherwise. The simplest expressions of analytical solutions are returned, explaining that the computation may take a few seconds on a PC.

## 5. Discussion

A new analytical model of gene evolution has been developed here. It is based on a dinucleotide mutation matrix $16 \times 16$ with six substitution parameters associated with the three types of substitutions in the two dinucleotide sites. It generalizes several previous models based on the nucleotide mutation matrices $4 \times 4$ (Jukes and Cantor, 1969; Kimura, 1980, 1981; Takahata and Kimura, 1981), and the dinucleotide model with one substitution parameter (Arquès and Michel, 1995). Several properties and two applications of this model have allowed to derive evolutionary analytical solutions of dinucleotides and also a dinucleotide phylogenetic distance $D_2$. This distance $D_2$ extends the classical site phylogenetic distance. According to formula (3.5), more the dinucleotides differ more its distance $D_2$ increases.

Other applications of this model can be applied to various problems. In particular, the eigenvalues given in (2.6) as well as the structure of the matrix $R = Q \cdot e^{Dt} \cdot Q^{-1}$ given in Appendix can be directly used to develop other evolution models based on a dinucleotide mutation matrix with six substitution parameters. For example, such a model can be applied to analyze the *CpG* frequencies in/near genes or other nonrepetitive DNA in genomes for identifying some evolutionary properties associated with this dinucleotide or for explaining its statistical distribution in actual genomes or chromosomes (e.g. Sved and Bird, 1990; Karlin et al., 1994; Das et al., 2006; Saxonov et al., 2006). Otherwise, the phylogenetic distance associated with a sequence could also improve some algorithms of phylogenetic tree reconstruction. Finally, the SED (Stochastic Evolution of Dinucleotides) web server has been developed for deriving evolutionary analytical solutions of dinucleotides, allowing the bioinformatics and biologists community to develop their own models of evolution.

## Acknowledgment

I thank a referee for his advice.

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

The square matrix $R = Q \cdot e^{Dt} \cdot Q^{-1}$ (16, 16) can be defined by a square block matrix (4, 4) whose four diagonal elements are formed by four identical square submatrices $S_1$ (4, 4) and whose 12 non-diagonal elements are formed by four square submatrices $S_5$ (4, 4), four square submatrices $S_9$ (4, 4) and four square submatrices $S_{13}$ (4, 4) as follows

$$R = Q \cdot e^{Dt} \cdot Q^{-1}$$

$$= \frac{1}{16}\left(\begin{array}{ccccc} & 1\cdots 4 & 5\cdots 8 & 9\cdots 12 & 13\cdots 16 \\ 1\cdots 4 & S_1 & S_5 & S_9 & S_{13} \\ 5\cdots 8 & S_5 & S_1 & S_{13} & S_9 \\ 9\cdots 12 & S_9 & S_{13} & S_1 & S_5 \\ 13\cdots 16 & S_{13} & S_9 & S_5 & S_1 \end{array}\right).$$

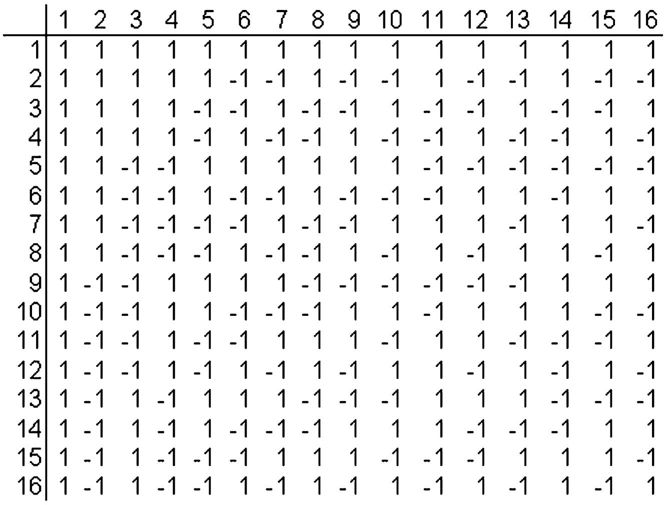
The square submatrix $S_i$ (4, 4) is defined as follows:

$$S_i = \begin{pmatrix} \mathscr{F}_i & \mathscr{F}_{i+1} & \mathscr{F}_{i+2} & \mathscr{F}_{i+3} \\ \mathscr{F}_{i+1} & \mathscr{F}_i & \mathscr{F}_{i+3} & \mathscr{F}_{i+2} \\ \mathscr{F}_{i+2} & \mathscr{F}_{i+3} & \mathscr{F}_i & \mathscr{F}_{i+1} \\ \mathscr{F}_{i+3} & \mathscr{F}_{i+2} & \mathscr{F}_{i+1} & \mathscr{F}_i \end{pmatrix},$$

where the function $\mathscr{F}_i$ associated with the $i$th line of $R$ is defined as

$$\mathscr{F}_i = \sum_{j=1}^{16} \delta_{ij} e^{\lambda_j t}$$

with the eigenvalues $\lambda_j$ defined in (2.6) and the constant $\delta_{ij}$, by the following matrix $\delta$ (Fig. 1).



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | -1 | 1 | -1 | -1 | 1 | -1 | -1 |
| 3 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | -1 | 1 | -1 | -1 | 1 | -1 | -1 | 1 |
| 4 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 | -1 | -1 | 1 | -1 | -1 | 1 | -1 |
| 5 | 1 | 1 | -1 | -1 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | -1 | -1 |
| 6 | 1 | 1 | -1 | -1 | 1 | -1 | -1 | 1 | -1 | -1 | -1 | 1 | 1 | -1 | 1 | 1 |
| 7 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | -1 | -1 | 1 | 1 | 1 | -1 | 1 | 1 | -1 |
| 8 | 1 | 1 | -1 | -1 | -1 | 1 | -1 | -1 | 1 | -1 | 1 | -1 | 1 | 1 | -1 | 1 |
| 9 | 1 | -1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | -1 | -1 | 1 | 1 | 1 |
| 10 | 1 | -1 | -1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | -1 | 1 | 1 | 1 | -1 | -1 |
| 11 | 1 | -1 | -1 | 1 | -1 | -1 | 1 | 1 | 1 | -1 | 1 | 1 | -1 | -1 | -1 | 1 |
| 12 | 1 | -1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 | 1 | -1 | 1 | -1 | 1 | -1 |
| 13 | 1 | -1 | 1 | -1 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 | -1 | -1 |
| 14 | 1 | -1 | 1 | -1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 |
| 15 | 1 | -1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 |
| 16 | 1 | -1 | 1 | -1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 |

Fig. 1. The matrix $\delta$.

## References

Arndt, P.F., Burge, C.B., Hwa, T., 2002. DNA sequence evolution with neighbor-dependent mutation. RECOMB'02, Proceedings of the Sixth Annual International Conference on Computational Biology, pp. 32–38.

Arquès, D.G., Michel, C.J., 1987. Periodicities in introns. Nucleic Acids Res. 15, 7581–7592.

Arquès, D.G., Michel, C.J., 1990. Periodicities in coding and noncoding regions of the genes. J. Theor. Biol. 143, 307–318.

Arquès, D.G., Michel, C.J., 1993. Analytical expression of the purine/pyrimidine codon probability after and before random mutations. Bull. Math. Biol. 55, 1025–1038.

Arquès, D.G., Michel, C.J., 1995. Analytical solutions of the dinucleotide probability after and before random mutations. J. Theor. Biol. 175, 533–544.

Das, R., Dimitrova, N., Xuan, Z., Rollins, R.A., Haghighi, F., Edwards, J.R., Ju, J., Bestor, T.H., Zhang, M.Q., 2006. Computational prediction of methylation status in human genomic sequences. Proc. Natl Acad. Sci. USA 103, 10713–10716.

Fryxell, K.J., Zuckerkandl, E., 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol. Biol. Evol. 17, 1371–1383.

Hess, S.T., Blake, J.D., Blake, R.D., 1994. Wide variations in neighbor-dependent substitution rates. J. Mol. Biol. 236, 1022–1033.

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.

Karlin, S., Ladunga, I., Blaisdell, B.E., 1994. Heterogeneity of genomes: measures and values. Proc. Natl Acad. Sci. USA 91, 12837–12841.

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

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

Konopka, A.K., Smythers, G.W., Owens, J., Maizel, J.V., 1987. Distance analysis helps to establish characteristic motifs in intron sequences. Gene Anal. Tech. 4, 63–74.

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

Michel, C.J., 1986. New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its evaluation. J. Theor. Biol. 120, 223–236.

Michel, C.J., 2007. 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. 69, 677–698.

Saxonov, S., Berg, P., Brutlag, D.L., 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl Acad. Sci. USA 103, 1412–1417.

Shepherd, J.C.W., 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl Acad. Sci. USA 78, 1596–1600.

Sved, J., Bird, A., 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc. Natl Acad. Sci. USA 87, 4692–4696.

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.