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 $e^{\lambda_k t}$.