Genome evolution by transformation, expansion and contraction (GETEC)

Emmanuel Benard¹, Sophie Lèbre, Christian J. Michel*

Theoretical Bioinformatics, ICube, University of Strasbourg, CNRS, 300 Boulevard Sébastien Brant, 67400 Illkirch, France

A R T I C L E   I N F O
Article history:
Received 16 December 2014
Received in revised form 4 May 2015
Accepted 21 May 2015
Available online 29 June 2015

Keywords:
Model of gene evolution
Substitution
Insertion
Deletion
Differential equation

A B S T R A C T
We propose here the GETEC (Genome Evolution by Transformation, Expansion and Contraction) model of gene evolution based on substitution, insertion and deletion of genetic motifs. The GETEC model unifies two classes of evolution models: models of substitution, insertion and deletion of nucleotides as function of time (Lèbre and Michel, 2010) and sequence length (Lèbre and Michel, 2012), and models of symmetric substitution of genetic motifs as function of time (Benard and Michel, 2011). Evolution of genetic motifs based on substitution, insertion and deletion is modeled by a differential equation whose analytical solutions give an expression of the genetic motif occurrence probabilities as a function of time or sequence length, as well as in direct time direction (past–present) or inverse time direction (present–past). Evolution models with “substitution only”, i.e. without insertion and deletion, and with “insertion and deletion only”, i.e. without substitution, are particular cases of the GETEC model. We have also developed a research software for computing the analytical solutions of the GETEC model. It is freely accessible at http://icube-bioinfo.u-strasbg.fr/webMathematica/GETEC or via the web site http://dpt-info.u-strasbg.fr/~michel/.

1. Introduction
Substitution, insertion and deletion of nucleotides are important molecular evolution processes. A major challenge for understanding genome and gene evolution is the mathematical analysis of these three processes.

1.1. Substitution models
Stochastic models of evolution were initially developed to study the substitution rates of nucleotides (adenine A, cytosine C, guanine G, thymine T). Typically, the substitution process is described by a differential equation defined by a constant rate substitution matrix (of size (4,4)) and whose analytical solutions give an expression of the nucleotide occurrence probabilities as function of time. The first substitution models were based on symmetric substitution matrices with one formal parameter for all nucleotide substitution types (Jukes and Cantor, 1969), two formal parameters for nucleotide transitions and transversions (Kimura, 1980) and three formal parameters for transitions and the two types of transversions (Kimura, 1981). These substitution models were later generalized to asymmetric substitution matrices (Felsenstein, 1981; Takahata and Kimura, 1981; Hasegawa et al., 1985; Tavaré, 1986; Tamura and Nei, 1993; Yang, 1994; Felsenstein and Churchill, 1996) with an equilibrium distribution different from 1/4 for all nucleotides.

During the last 25 years, parallel to the growth in complexity of nucleotide substitution matrices, we introduced substitution matrices for genetic motifs, i.e. matrices of sizes (16,16) for dinucleotides, (64,64) for trinucleotides, etc., and obtained the analytical solutions of the associated differential equations in various cases. They were expressed as a mean number of random substitutions per base site or as function of time with several formal parameters: trinucleotide matrix on the alphabet \( \{R,Y\} \) \( R = \{A,G\} \) and \( Y = \{C,T\} \) (Arquès and Michel, 1993), dinucleotide matrix on the alphabet \( \{A,C,G,T\} \) (Arquès and Michel, 1995) and with \( 2 \times 3 = 6 \) formal parameters (Michel, 2007c), trinucleotide matrix on the alphabet \( \{A,C,G,T\} \) with \( 3 \times 1 = 3 \) formal parameters (Arquès et al., 1998, 1999), \( 3 \times 2 = 6 \) formal parameters (Frey and Michel, 2006) and \( 3 \times 3 = 9 \) formal parameters (Michel, 2007a; Benard and Michel, 2009). The development of these models, i.e. the determination of analytical solutions of the differential equations, required several years, mainly due to two facts: (i) the analytical expression of eigenvalues and eigenvectors of motif substitution matrices cannot be computed in a straightforward way and required some
tedious algebra manipulation, in particular linear combination of the determinant and block-matrix factorization (Tian and Styan, 2001); and (ii) the limited power of formal calculus, e.g. Mathematica, at that time. This problem has recently been solved by using Kronecker operators (sum and product) which allowed us to generalize the 3-parameter symmetric substitution matrix (Kimura, 1981) to substitution matrices for motifs of any (finite) size (Benard and Michel, 2011). Motif substitution matrices with constant formal parameters were also generalized to time dependent parameters (Bahi and Michel, 2004), then later to chaotic constant parameters (Bahi and Michel, 2008) and finally to chaotic time dependent parameters (Bahi and Michel, 2009). Notably, in the particular case of 3-letter motifs, these approaches can be used for codon substitution models (see Anisimova and Kosiol, 2009, for a review).

1.2. Substitution, insertion, deletion models

In addition, some molecular evolution models were extended to the insertion and deletion of residues (nucleotides, amino acids) as well as residue substitution. These substitution–insertion–deletion (SID) models can be divided into three classes. A pioneering paper by Thorne et al. (1991) proposed a time-reversible Markov model for insertions and deletions (termed the TKF91 model). This SID model represents sequence evolution in two steps. First, the sequence is subjected to an insertion–deletion process which is homogeneous over all sites in the sequence. Second, conditional on the result of the insertion–deletion process, a substitution process is applied to the two sequences. The total process is time-reversible whenever the substitution process is. Some drawbacks of the preliminary TKF91 model were improved by the same authors with the TKF92 version of the model (Thorne et al., 1992). Later, the original SID models were refined in many ways, for instance by Metzler (2003) and Miklós et al. (2004) (see e.g. Miklós et al., 2009, for a review). A second class of SID models was introduced by McGuire et al. (2001) who defined a Markov model by extending the F84 substitution matrix (Felsenstein and Churchill, 1996) of size four comprising the four nucleotides to a substitution matrix of size five with one additional line and one additional column for the gap character involved in the alignment. Then, an insertion is described by the substitution of a gap by a nucleotide whereas a deletion amounts to the substitution of a nucleotide by a gap. The insertion rate is proportional to the F84 substitution matrix equilibrium distribution. A third class of SID models was introduced by Rivas (2005) with a non-reversible evolution model which extends the model of McGuire et al. (2001) for the evolution of sequences of residues in any alphabet of size K, i.e. for any substitution matrix. The insertion rates are defined by explicit parameters and the deletion rate is uniform for all residues. In the particular case where the insertion rate is proportional to the substitution matrix equilibrium distribution, an analytical expression of the substitution probabilities P(i,j) of residue j by residue i over time t can be derived (Rivas and Eddy, 2008). However, even if the insertion process is independent of the substitution process, the substitution and deletion processes are not independent (detailed in Section 1 in Lèbre and Michel, 2012).

More recently, we have developed a dynamic evolution model (called IDIS model) inspired by a concept in population dynamics (Malthus, 1798) where the three processes of substitution, insertion and deletion of nucleotides are independent of each other (Lèbre and Michel, 2010, 2012). This model is defined by a differential equation whose analytical solution gives an expression of the sequence content vector P(t) at evolution time t (Lèbre and Michel, 2010) or P(l) at sequence length l (Lèbre and Michel, 2012) for any diagonalizable substitution matrix M of nucleotides.

1.3. GETEC model

The molecular evolution models we have developed over the last 25 years, i.e. substitution models of motifs as well as substitution–insertion–deletion models of nucleotides (summarized in Fig. 1), have several interesting mathematical properties compared to some other evolution models in this research field: (i) they rely on a real physical process of sequence evolution, in other words, the analytical expressions of the sequence content at time t are identical (by numerical approximations) to the values obtained by simulating sequence evolution under substitution, insertion and deletion; thus, they allow a realistic interpretation of the model parameters (evolution time t, sequence length l and rates of substitution, insertion and deletion); (ii) they enable the mathematical analysis of the sequence content curves along time with local/global maxima or minima, increasing or decreasing curves, crossing curves, asymptotic behavior, etc.; (iii) they provide a description of sequence content evolution and in particular the evolution of motif content inside the sequence, unlike the phylogenetic approaches for tree reconstruction; and (iv) they allow to introduce models of “primitive” genes or “primitive” motifs of nucleotides or amino acids, to study substitution rates, to analyze the residue occurrence probabilities in the natural evolution time direction (from past to present or from present to future) or the inverse direction (from present to past).

We propose here to generalize the evolution model for motif substitution (Benard and Michel, 2011) to an evolution model for motif substitution–insertion–deletion using the IDIS model (Lèbre and Michel, 2010, 2012). The generalized model, called GETEC (Genome Evolution by Transformation, Expansion and Contraction), is based on substitution, insertion and deletion of genetic motifs of any (finite) size. The three evolution processes are independent of each other (following the IDIS model assumptions) and the motif substitution matrix extends the classical 3-parameter symmetric substitution matrix (Kimura, 1981). The GETEC model yields an analytical expression of the vector P(t) of motif content in the sequence at evolution time t or the vector P(l) of motif content at sequence length l as function of the substitution parameters (three parameters (a, b, c) per site s ranging from 1 to the motif length), a vector R of the motif insertion rates, the total insertion rate r, a deletion rate d and the vector P(l0) of initial motif content in the sequence at evolution time t0 or the vector P(l0) of initial motif content at sequence length l0. We have also developed a research software for computing online the analytical solutions of the GETEC model associated with a chosen set of parameters. It is freely accessible at http://icube-bioinfo.u-strasbg.fr/webMathematica/GETEC/ or via the web site http://dpt-info.u-strasbg.fr/~michel/. It allows biologists and bioinformaticians to develop their own gene evolution models by studying evolution of genetic motifs, both in the direct evolution time direction (past–present) and the inverse evolution time direction (present–past). To our knowledge, the model GETEC and its computational software have no equivalent in this evolutionary field.

This paper is organized as follows. Section 2 presents first a new and comprehensive formulation of the substitution–insertion–deletion models for nucleotides initiated in Lèbre and Michel (2010, 2012), and second the substitution model for motifs (Benard and Michel, 2011) with here a new and simplified proof for the recursive construction of a motif substitution matrix. Section 3 describes the construction of the GETEC model which allows the substitution, insertion and deletion of genetic motifs. The analytical solutions are given for the GETEC model at time t and sequence length l, and for particular cases of the GETEC model: “substitution only” model at time t and “insertion–deletion only” model at time t and sequence length l. The relationship between time t and sequence length l in the GETEC
model is presented. Section 4 describes the research software GETEC which computes the analytical solutions of the GETEC model online. Section 5 provides the detailed procedures of the GETEC software to retrieve the classical formulas of nucleotide substitution matrices with one formal parameter (Kimura, 1980) and two formal parameters (Kimura, 1980). Section 6 provides an example of biological application to an evolution study of the amino acid glycine in bacterial genes.

2. Two classes of evolution models


We propose here a new and global formulation of the substitution–insertion–deletion models for nucleotides as function of time and sequence length, called IDIS (Insertion and Deletion Independent of Substitution) models initiated in Lèbre and Michel (2010, 2012).

The IDIS model is defined by explicit parameters for the insertion rate $r_i$ of each residue $i$ and the deletion rate $d$. The insertion rates and the deletion rate are independent of each other and also independent of the substitution parameters. Let us consider an alphabet of $K$ residues. For example, $K = 4$ for the set of nucleotides {A, C, G, T}, $K = 20$ for the set of amino-acids, $K = 2$ for the set of purine and pyrimidine {R, Y}. For all $1 \leq i \leq K$, let $p_i(t)$ be the occurrence probability of residue $i$ at time $t > 0$ per “residue site” in the sequence and $P(t) = [p_i(t)]_{1 \leq i \leq K}$ the column vector of size $K$ made of the probabilities $p_i(t)$ for all $1 \leq i \leq K$.

The IDIS model superimposes a substitution process and an insertion–deletion process. By assuming that the substitution and the insertion–deletion processes are independent, i.e. a substitution event does not alter the probability of an insertion–deletion event and reciprocally, the derivative $P'(t)$ of the residue occurrence probability at time $t$ is the result of the instantaneous variation due to substitution and insertion–deletion,

$$P'(t) = (M - I) \cdot P(t) + (-rP(t) + R)$$

(2.1)

where $A = M - (1 + r)I$, $M = [Pr(j \rightarrow i)]_{1 \leq i,j \leq K}$ is the substitution probability matrix, $R = [r_i]_{1 \leq i \leq K}$ is the vector of the residue insertion rates per site and $r = \sum_{1 \leq i \leq K} r_i > 0$ is the total insertion rate, $\forall 1 \leq i \leq K$, $r_i \geq 0$. Explanation of Eq. (2.1) is briefly recalled below (see detail in Lèbre and Michel, 2010).

(i) Substitution term in Eq. (2.1). The change of the residue occurrence probability due to substitution is governed by the classical matrix differential equation (Michel, 2007a)

$$P(t) = M \cdot P(t) - P(t) = (M - I) \cdot P(t)$$

(2.2)

where $M = [m_{ij}]_{1 \leq i,j \leq K}$ is the substitution probability matrix with element $m_{ij} = Pr(j \rightarrow i)$ in row $i$ and column $j$ referring to the substitution probability of residue $j$ into residue $i$, matrix $I$ is the identity matrix of size $K$ and the symbol $\cdot$ is the matrix product.

Remark 1. The matrix $M = [m_{ij}]_{1 \leq i,j \leq K}$ is the instantaneous substitution probability matrix whose element $m_{ij}$ in row $i$ and column $j$ refers to the substitution probability $m_{ij} = Pr(j \rightarrow i)$ of residue $j$ into residue $i$. Thus, the substitution probability matrix $M$ is stochastic in column. Indeed, for all $1 \leq j \leq K$, the elements of matrix $M$ satisfy $\sum_{1 \leq i \leq K} m_{ij} = \sum_{1 \leq i \leq K} Pr(j \rightarrow i) = 1$. The substitution probability
matrix $M$ is the transpose matrix of the classical substitution matrix $\pi = \{Pr(i \to j)\}_{1 \leq i,j \leq K}$ which is stochastic in line (e.g. Kimura, 1980, 1981), i.e. $\pi_{ij} = Pr(i \to j) = m_{ji}$.

(ii) Insertion–deletion term in Eq. (2.1). The insertion–deletion process is modeled by explicit parameters which are set independently from the substitution parameters: $r_i$ is the insertion rate per site of each residue $i$, $\forall 1 \leq i \leq K$, $r_i \geq 0$, and $d$ is the deletion rate for all residues, $d \geq 0$. Let $n_i(t)$ be the occurrence number of residue $i$ in the biological sequence at time $t$ and $n(t) = \sum_{1 \leq i \leq K} n_i(t)$ be the total number of residues at time $t$. By definition, a sequence has at least one residue, i.e. $n(t) \geq 1$. From a concept in population dynamics (Malthus, 1798), the growth rate $n_i'(t) = \frac{dn_i(t)}{dt}$ of residue $i$ at time $t$ due to insertion is equal to $r_i \times n(t)$. Similarly, the growth rate $n_i'(t)$ of residue $i$ at time $t$ due to deletion is $d \times n_i(t)$. Thus, the growth rate $n_i'(t)$ resulting from the insertion–deletion process is, for all $1 \leq i \leq K$,

$$n_i'(t) = r_i \times n(t) - d \times n_i(t).$$

(2.3)

The derivative $P(t)$ of the occurrence probability of residue $i$ at time $t$ can be written

$$P_i'(t) = \frac{\partial}{\partial t} \left( \frac{n_i(t)}{n(t)} \right) = \frac{1}{n(t)} \left( r_i n(t) - d n_i(t) - n_i(t) \sum_{1 \leq j \leq K} n_j(t) \right).$$

By replacing $n_i(t)$ using Eq. (2.3), we obtain (see detail in Lèbre and Michel, 2010)

$$P_i'(t) = r_i - \left( \sum_{1 \leq j \leq K} r_j \right) P_i(t).$$

With the total insertion rate $r = \sum_{1 \leq i \leq K} r_i > 0$, the change of the residue occurrence probability due to insertion–deletion is explained by the matrix differential equation

$$P'(t) = -rP(t) + R.$$

(2.4)

A general solution of Eq. (2.1) containing substitution Eq. (2.2) and insertion–deletion Eq. (2.4) is derived when the substitution probability matrix $M$ can be diagonalized with real eigenvalues (Proposition 1). It is well known that the substitution matrices of reversible models are diagonalizable with real eigenvalues (Aldous and Fill, 2002) but this is not an exclusive condition as substitution matrices of non-reversible models can also be diagonalized with eigenvalues (e.g. Exercises of Chapter 1 in Kelly, 1979).

**Proposition 1.** When the substitution probability matrix $M$ can be diagonalized with real eigenvalues $(\lambda_k)_{1 \leq k \leq K}$, an analytical solution of the IDS model defined by Eq. (2.1) is derived. Let $Q$ be an associated eigenvector matrix of $M$, the $k$th column of $Q$ being an eigenvector for the IDIS model defined by Eq. (2.1). When the substitution probability matrix $M$ can be diagonalized with real eigenvalues $(\lambda_k)_{1 \leq k \leq K}$, for any (non-zero) residue insertion rate vector $R = [r_i]_{1 \leq i \leq K}$, $\forall 1 \leq i \leq K$, $r_i \geq 0$, and the total insertion rate $r = \sum_{1 \leq i \leq K} r_i > 0$, let $Q(t)$ be the total number of residues at time $t$, the residue occurrence probability $P(x,t)$ as a function of variable $x$ representing time $t$ or sequence length $x = t$ by introducing a function $h(x, x_0)$ which is equal to $h(x, x_0) = e^{-(t-t_0)}$ for evolution time $t$ and $h(x, x_0) = \left( \frac{l}{t_0} \right)^\frac{x}{-r}$ for sequence length $l$.

**Proposition 2.** When the substitution probability matrix $M$ can be diagonalized with real eigenvalues $(\lambda_k)_{1 \leq k \leq K}$, for any (non-zero) residue insertion rate vector $R = [r_i]_{1 \leq i \leq K}$, $\forall 1 \leq i \leq K$, $r_i \geq 0$, and the total insertion rate $r = \sum_{1 \leq i \leq K} r_i > 0$, the residue occurrence probability $P(x,t)$ as a function of a variable $x$ representing time $t$ or sequence length $x = t$ with the following convention $(x, x_0, h(x, x_0)) = (t, t_0, e^{-(t-t_0)})$ for time expression and $(x, x_0, h(x, x_0)) = (l, t_0, \left( \frac{l}{t_0} \right)^\frac{x}{-r})$ for sequence length expression is

$$P(x) = \left( \sum_{k=1}^K \frac{1}{t+1-\lambda_k} O_k \right) \cdot R$$

$$+ \sum_{k=1}^K O_k \cdot \left( P(x_0) - \frac{1}{t+1-\lambda_k} R \right) h(x, x_0)^{t+1-\lambda_k}$$

(2.6)

where the matrices $(O_k)_{1 \leq k \leq K}$ of size $K \times K$ are defined from the eigenvector matrix $Q$ of matrix $M$ by

$$O_k = Q \cdot (\delta_k) \cdot (\epsilon_k)^T \cdot Q^{-1}$$

with $\delta_k = (\delta_{i,k}) = (0,0,\ldots,0,1,0,\ldots,0)^T$, a vector having 1 in $k$th row and 0 otherwise, and $(\epsilon_k)^T$, its transpose vector.

**Proof.**

(i) Case $x = t$. $P(t)$ is obtained after some algebraic manipulation of Eq. (2.5) (see also Eq. (2.13) in Lèbre and Michel, 2010, for the particular case $t_0 = 0$).

(ii) Case $x = l$. $P(l)$ is obtained by deriving from Eq. (2.3)

$$n(t) = \sum_{1 \leq i \leq K} n_i(t) = (r-d)n(t)$$

which leads to $e^{-(t-t_0)} = \left( \frac{n(t)}{M(t)} \right)^\frac{x}{-r} = \left( \frac{l}{t_0} \right)^\frac{x}{-r}$ (see also Eq. (10) in Lèbre and Michel, 2012, for the particular case $t_0 = 0$).

□

The general formula (2.6) allows to derive the residue occurrence probability $P(t)$ at time $t$ and $P(l)$ at sequence length $l$ both in the direct $(t > t_0, l > t_0)$ or inverse $(t < t_0, l < t_0)$ direction of evolution. In the direct evolution direction, $P(t)$ and $P(l)$ converge to the residue equilibrium distribution when $r$ and $l$ increase (Eqs. (4.7) and (4.9) in Lèbre and Michel, 2010, and Proposition 3 in Lèbre and Michel, 2012). In the inverse evolution direction, $P(t)$ and $P(l)$ do not converge when $t$ and $l$ increase, and the only constraint to be respected is that $P(t)$ and $P(l)$ remain probability vectors. This condition becomes not verified when a residue probability has a negative value.

**Remark 2.** The sum of the matrices $(O_k)_{1 \leq k \leq K} = \sum_{k=1}^K O_k = Q \cdot Q^{-1} = I$. Indeed, for all $i, j, \sum_{1 \leq k \leq K} O_{i,j} = \sum_{1 \leq k \leq K} Q_{i,j} \cdot Q^{-1} = [Q, k, j]$. It is the term in row $i$ and column $j$ of the matrix product $Q \cdot Q^{-1}$.

**Remark 3.** The non-zero condition for the vector $R$ of insertion rates ensures that $r = \sum_{1 \leq i \leq K} r_i > 0$. Thus, the denominator of
Remark 4. As in all the current insertion–deletion models for gene evolution, the deletion rate $d_i$ of each residue $i$ is equal to $d$. It is classically assumed that there is no distinction among residue for deletion. Moreover, the derivation of an analytical expression is not ensured with specific deletion rate $d_i$ for each residue $i$.

Particular cases such as the “substitution only” model (Proposition 3) and the “insertion–deletion only” model (Proposition 4) can be derived from the general formula (2.6).

**Proposition 3.** “Substitution only” model. The residue occurrence probability $P(t)$ at time $t$ is equal to $P(t) = \sum_{k=1}^{K} O_k e^{-\lambda_k t} \cdot P(t_0)$ where $(\lambda_k)_{1\leq k \leq K}$ are real eigenvalues of the substitution probability matrix $M$ and matrices $(O_k)_{1\leq k \leq K}$ of size $K \times K$ are defined by $O_k = Q \cdot 1_k \cdot W_k \cdot Q^{-1}$ with $Q$ the eigenvector matrix of matrix $M$, and $1_k = \langle 1_k \rangle = \langle 0, 0, 0, 1, 0, 0, 0 \rangle$, a vector having $1$ in $k$th row and $0$ otherwise, and $(1_k)^T$, its transpose vector.

**Proposition 4.** “Insertion–deletion only” model. The residue occurrence probability $P(t)$ at time $t$ or sequence length $x = t$ with the following convention $(x, x_0, h(x, x_0)) = (t, t_0, e^{-t_0} \cdot P(t_0))$ for time expression and $(x, x_0, h(x, x_0)) = (t, t_0, (\frac{1}{2} t_0)^{-\frac{1}{2}})$ for sequence length expression is equal to $P(x) = \frac{R}{t} + \left( P(x_0) - \frac{R}{t} \right) h(x, x_0)^y$ where $R = [r_i]_{1\leq i \leq K}, 1 \leq i \leq K, r_i \geq 0$, is the residue insertion rate vector, $r = \sum_{1\leq i \leq K} r_i > 0$ is the total insertion rate and $d \geq 0$ is the deletion rate.

2.2. Substitution models for motifs (Benard and Michel, 2011)

A Kronecker property was identified for constructing symmetric substitution matrices for genetic motifs of size $n$ containing up to three substitution parameters per motif site and for solving their eigeneigen analytically. It was found by Benard and Michel (2011) after a detailed analysis of the dinucleotide matrix $\delta$ (Fig. 1 in Michel, 2007c) and the trinucleotide matrix $\delta$ (Fig. B.1 in Michel, 2007b). It allows to derive analytical solutions giving the occurrence probabilities of genetic motifs of size $n$ at time $t$ with 3-parameter symmetric substitution matrices. Thus, it extends the classical 3-parameter symmetric substitution model of nucleotides (Kimura, 1981) to any genetic motif of size $n$.

We propose here a new and simplified proof for the recursive construction of a motif substitution matrix $A_n$ by applying the Kronecker operators to nucleotide substitution matrices $N_i$ associated to each site $i$ of genetic motifs of size $n$.

Let $s$ be the nucleotide site of a genetic motif of size $n, 1 \leq s \leq n$. For a given site $s$, let $a_i$, $b_i$ and $c_i$ be the parameters of transitions $A$ $\longleftrightarrow$ $G$ and $C$ $\longleftrightarrow$ $T$, transversions $I A$ $\longleftrightarrow$ $T$ and $C$ $\longleftrightarrow$ $G$ and transversions $II A$ $\longleftrightarrow$ $C$ and $G$ $\longleftrightarrow$ $T$, respectively. For example, when considering a dinucleotide $w = I_1$ $I_2$ then $a_1$, $b_1$ and $c_1$ are the transitions, transversions I and transversions II in the 1st site $l_1$ of $w$, respectively, and $a_2$, $b_2$ and $c_2$ are the transitions, transversions I and transversions II in the 2nd site $l_2$ of $w$, respectively. Thus, a motif of size $n$ has $3^n$ substitution parameters. Let us denote by $A_n = M_n - I_n$ of size $(4^n, 4^n)$ the symmetric substitution rate matrix of motif of size $n$ where $M_n$ is the instantaneous substitution probability matrix for motifs of length $n$ and $I_n$ is the identity matrix of size $(4^n, 4^n)$ (see Eq. (2.2)). The columns and lines of $A_n$ sum to 0. Matrix $A_n$ is a block matrix which is classically constructed recursively by varying $s = n$ to $s = 1$ as follows (Michel, 2007b,c)

$$A_n = \begin{pmatrix} A_{n-1} & c_{n-1} & a_{n-1} & b_{n-1} \\ c_{n-1} & A_{n-1} & b_{n-1} & a_{n-1} \\ a_{n-1} & b_{n-1} & A_{n-1} & c_{n-1} \\ b_{n-1} & a_{n-1} & c_{n-1} & A_{n-1} \end{pmatrix}$$

where $I_{n-1}$ is the identity matrix of size $(4^{n-1}, 4^{n-1})$ with $I_1 = 1$, $A_{n-1}$ is the recursive matrix of size $(4^{n-1}, 4^{n-1})$ with $A_0 = -\sum_{s=1}^{n} (a_s + b_s + c_s)$ and $a_s$, $b_s$, $c_s, 1 \leq s \leq n$, are the substitution parameters for the $s$th motif site.

As the matrix $A_n$ is real and symmetric, $A_n$ is diagonalizable, i.e. $A_n = Q_n \cdot D_n \cdot Q_n^{-1}$ where $D_n$ is the spectral matrix of $A_n$ and $Q_n$ is its associated eigenvector matrix. This property will allow the occurrence probability $P(x)$ of residues in Eq. (2.6) to be extended to genetic motifs.

Let $N_i, 1 \leq s \leq n$, the nucleotide substitution rate matrix of size $(4,4)$ of a site $s$ of a motif of size $n$

$$N_s = \begin{pmatrix} d_1 & c_1 & a_1 & b_1 \\ c_2 & d_2 & b_2 & a_2 \\ a_3 & b_3 & d_3 & c_3 \\ b_4 & a_4 & c_4 & d_4 \end{pmatrix}$$

with $d_s = -(a_s + b_s + c_s)$. As the matrix $N_i$ is real and symmetric, $N_i$ is diagonalizable for all $1 \leq s \leq n$

$$N_s = Q \cdot S_s \cdot Q^{-1}$$

where the nucleotide spectral matrix $S_s$ of $N_i$ is

$$S_s = \begin{pmatrix} 0 & 0 & 2(a_s + c_s) & 0 \\ 0 & 0 & 0 & 2(b_s + c_s) \end{pmatrix}$$

and its associated nucleotide eigenvector matrix $Q$ is

$$Q = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Remark 5. For the substitution rate matrix of nucleotides $(n = 1), A_1 = N_1 = Q \cdot D_1 \cdot Q^{-1} = Q \cdot S_1 \cdot Q^{-1}$ leading to $D_1 = S_1$ and $Q_1 = Q$.

Remark 6. $Q^{-1} = \frac{1}{2} Q$.

**Proposition 5.** The spectral matrix $D_n$ and the eigenvector matrix $Q_n$ to be deduced from $S_s$ and $Q$, are respectively

$$D_n = \otimes_{s=1}^{n} S_s$$

$$Q_n = \otimes_{s=1}^{n} Q$$

$$Q_n^{-1} = (\otimes_{s=1}^{n} Q)^{-1} = \otimes_{s=1}^{n} Q^{-1} = \otimes_{s=1}^{n} \left( \frac{1}{Q} \right) = \frac{1}{4^n} \otimes_{s=1}^{n} Q = \frac{1}{4^n} Q_n$$

where the operators $\oplus$ and $\otimes$ are the Kronecker sum and the Kronecker product, respectively (defined e.g. in Laub (2005)). Thus, the motif substitution rate matrix $A_n$ can be directly determined from the Kronecker
Proposition 6. Let us denote by $M_n$ the instantaneous substitution probability matrix for motifs of length $n$ where $A_n$ is the symmetric substitution rate matrix for n-letter motifs (Eq. (2.7)) and $I_n$ is the identity matrix of size $(4^n, 4^n)$. Then, the substitution probability matrix $M_n$ is diagonalizable with diagonal spectral matrix $D_n + I_n$ and eigenvectors matrix $Q_n$, such that $M_n = \otimes_{n=1}^N S_n$ and $Q_n = \otimes_{n=1}^N Q$ where $S_n$ is the 3-parameter symmetric substitution matrix associated with site $s$ (Eq. (2.8)) and $Q$ is the eigenvectors matrix associated with any 3-parameter symmetric substitution matrix (Eq. (2.9)).

Proof. From Proposition 5, the motif substitution rate matrix $A_n$ is diagonalizable with real eigenvalues and decomposes as $A_n = Q_n \cdot D_n \cdot Q_n^{-1}$ where $D_n$ is the diagonal spectral matrix of $A_n$ and $Q_n$ is its associated eigenvectors matrix. Then, the substitution probability matrix $M_n$ satisfies

$$M_n = A_n + I_n$$

$$= Q_n \cdot D_n \cdot Q_n^{-1} + I_n$$

$$= Q_n \cdot D_n \cdot Q_n^{-1} + Q_n \cdot I_n \cdot Q_n^{-1}$$

$$= Q_n \cdot (D_n + I_n) \cdot Q_n^{-1}$$

where $D_n = \otimes_{n=1}^N S_n$ and $Q_n = \otimes_{n=1}^N Q$ results from Proposition 5. \qed

Proposition 7. The GETEC model for substitution, insertion and deletion of n-letter genetic motifs with symmetric substitution probability matrix $M_n$ defined in Proposition 6, n-letter genetic motif insertion rate vector $R = |r_i|_{1 \leq i \leq 4^n}$ with $\forall 1 \leq i \leq 4^n, r_i \geq 0$ and deletion rate $d$ satisfies Eq. (2.6) giving the occurrence probability of genetic motifs of size $n$ as function of time $t$ and sequence length $l$ with

$$\begin{cases}
\lambda_k = 1 + D_n[k, k] \\
O_k = \frac{1}{4^n} \otimes_{n=1}^N Q \cdot 1_k \cdot (1_k)^T \cdot \otimes_{n=1}^N Q
\end{cases}$$

(3.1)

where $D_n = (\otimes_{n=1}^N S_n)$, $S_n$ is the 3-parameter symmetric substitution matrix associated with site $s$ (Eq. (2.8)) and $Q$ is the eigenvectors matrix associated with any 3-parameter symmetric substitution matrix (Eq. (2.9)).

3. GETEC model

The GETEC (Genome Evolution by Transformation Expansion Contraction) model introduced here generalizes the motif substitution model (Section 2.2) to a motif substitution–insertion–deletion model. To our knowledge, it is the first biomathematical model of gene evolution in this research field analyzing transformation, expansion and contraction of genetic motifs during evolutionary time, and moreover, in both directions, direct (past–present) and inverse (present–past).

In the subsections below, we give two Propositions 6 and 7 for constructing the GETEC model. Then, we derive the analytical solutions from Propositions 6 and 7: model TECt (Transformation Expansion Contraction) at time $t$ and model TECI (Transformation Expansion Contraction) at sequence length $l$, and the particular cases: $Tt$ (Transformation) at time $t$, $EC$ (Expansion Contraction) at time $t$ and $EI$ (Expansion Contraction) at sequence length $l$.

3.2. Analytical solutions of the GETEC model

We give here the new analytical solutions which are derived from the GETEC model: $TECt$ (Transformation Expansion Contraction) at time $t$ and $TECI$ (Transformation Expansion Contraction) at sequence length $l$, and the particular cases: $Tt$ (Transformation) at time $t$, $EC$ (Expansion Contraction) at time $t$ and $EI$ (Expansion Contraction) at sequence length $l$.

3.2.1. Model TECt (Transformation Expansion Contraction) at time $t$

Using Eq. (2.6) and the relations (3.1), the occurrence probability $P(t)$ of genetic motifs of size $n$ at time $t$ with the initial condition $P(0)$ at time $t_0 = 0$ is

$$P(t) = \left( \sum_{k=1}^{4^n} \frac{1}{r+1-\lambda_k} O_k \right) \cdot R$$

$$+ \sum_{k=1}^{4^n} O_k \cdot \left( P(0) - \frac{1}{r+1-\lambda_k} R \right) e^{-(r+1-\lambda_k)t}$$

(3.2)

where $R = [r_i]_{1 \leq i \leq 4^n}$ is the vector of n-letter genetic motif insertion rate with $\forall 1 \leq i \leq 4^n, r_i \geq 0$, $r = \sum_{1 \leq i \leq 4^n} r_i$ is the total motif genetic insertion rate with $r \geq 0$ and for all $1 \leq k \leq 4^n$

$$O_k = \frac{1}{4^n} \otimes_{n=1}^N Q \cdot 1_k \cdot (1_k)^T \cdot \otimes_{n=1}^N Q$$

with $1_k = (\delta_{1,k}) = (0, \ldots, 0, 1, 0, \ldots, 0)^T$, a vector having 1 in $k$th row and 0 otherwise and

$$\lambda_k = 1 + \left( \otimes_{n=1}^N S_n \right) [k, k],$$

with

$$S_n = \begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & -2(a_n + b_n) & 0 & 0 \\
0 & 0 & -2(a_n + c_n) & 0 \\
0 & 0 & 0 & -2(b_n + c_n)
\end{pmatrix}$$

and

$$Q = \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 \\
1 & -1 & 1 & -1
\end{pmatrix}$$

defined in Eqs. (2.8) and (2.9).

The time inversion proposition (Lèbre and Michel, 2010, Section 3.3) allows the evolution time direction to be inverted for the substitution–insertion–deletion model. If $t \geq 0$ then the evolution direction is direct else inverse. From a computational point of view, the analytical formulas in the inverse evolution direction...
From Eq. (3.2), the occurrence probability \( P_i(t) \) of a chosen genetic motif \( i \) at time \( t \), implemented in the GETEC software, is easily obtained by

\[
P_i(t) = \left( \sum_{k=1}^{4^n} \frac{1}{r+1-\lambda_k} O_k[i, j] \right) \cdot R + \sum_{k=1}^{4^n} O_k[i, j] \left( P(0) - \frac{1}{r+1-\lambda_k} R \right) e^{-r+1-\lambda_k} t.
\]

### 3.3. Particular cases of the GETEC model

#### 3.3.1. “Substitution only” model \( Tt \) (Transformation) at time \( t \)

The occurrence probability \( P(t) \) of genetic motifs of size \( n \) at time \( t \) with the initial condition \( P(0) \) is

\[
P(t) = \left( \sum_{k=1}^{4^n} O_k \right) \cdot R + \sum_{k=1}^{4^n} O_k \left( P(0) - \frac{1}{r+1-\lambda_k} R \right) \left( \frac{1}{l_0} \right)^{r+1-\lambda_k} t.
\]

### 3.4. Relation between time \( t \) and sequence length \( l \) in the GETEC model

From the growth rate \( n'(t) \) of residue \( i \) at time \( t \) resulting from the insertion–deletion process (Eq. (2.3)), the number \( l = n(t), l \geq 1 \), of residues in the sequence at time \( t \) is

\[
\forall t \geq 0, \quad l = l_0 e^{r-d} t.
\]

### 4. Development of the research software GETEC

We present here the different functionalities of the research software GETEC (Genome Evolution by Transformation Expansion Contraction) freely accessible at [http://icube-bioinfo.u-strasbg.fr/webMathematica/GETEC](http://icube-bioinfo.u-strasbg.fr/webMathematica/GETEC) or via the web site [http://dpt-info.u-strasbg.fr/~michel](http://dpt-info.u-strasbg.fr/~michel) (Fig. 2).
Fig. 2. Home page of the research software GETEC. Evolution Plots functionalities and Formal Analytical Solutions functionalities are available for the five classes of evolution models TECt, TECl, Tt, ECT and ECl.
description of the GETEC functionalities is given here for the computer user.

4.1. Gene evolution models available in the GETEC software

Five gene evolution models are proposed in the GETEC software to compute evolution of occurrence probabilities of genetic motifs. The most general models are the substitution, insertion and deletion models TEC (Transformation Expansion Contraction; Eq. (3.2)) at time $t$ and TEC (Transformation Expansion Contraction; Eq. (3.3)) at sequence length $l$. The particular models are the “substitution only” model $T_t$ (Transformation; Eq. (3.4)) at time $t$ and the “insertion–deletion only” models $EC_t$ (Expansion Contraction; Eq. (3.5)) at time $t$ and $EC_l$ (Expansion Contraction; Eq. (3.6)) at sequence length $l$. For the five models TEC, TEC, $T_t$, $EC_t$ and $EC_l$, formal and numerical analytical solutions and evolution plots are available in the Evolution Plots functionality and the general formal analytical solutions are given in the Formal Analytical Solutions functionality.

4.2. Size of genetic motifs

The computation complexity (time and space) of the analytical solutions depend on the gene evolution model and the motif size. For the general models $TEC_t$ and $TEC_l$, the genetic motif sizes allowed are length 1, i.e. the four genetic motifs {A, . . . , T}, to 4, i.e. the 256 genetic motifs {AAAA, . . . , TTTT}. For the particular models $T_t$, $EC_t$ and $EC_l$, the genetic motifs can have a size up to 5, i.e. the 1024 genetic motifs {AAAAA, . . . , TTTTT}. This motif limitation is not related to the mathematical model but to the GETEC software which is currently hosted on a simple PC with a Core i7-4770 at 3.4 GHz and 8 Go RAM.

4.3. Formal Analytical Solutions functionality

The Formal Analytical Solutions functionality proposes the general formal analytical solution of one particular genetic motif for the models $T_t$, $EC_t$ and $EC_l$ at time $t \geq 0$ (in the direct evolution direction), and for the models $TEC_l$ and $EC_l$ at sequence length $l \geq l_0$ when $(r - d) \geq 0$. An example with the dinucleotide $GT$ for the model $TEC_t$ is given in Fig. 3.

For each model, three options permit to obtain the general formal analytical solutions: choice of the motif size $n$; choice of the genetic motif among the $4^n$ possible motifs; and choice of the output format (Standard, C, Fortran or TeX) for the solution which is displayed in the Results interface and can be saved in a text file.

4.4. Evolution Plots functionality

The Evolution Plots functionality allows for the five models $TEC_t$, $TEC_l$, $T_t$, $EC_t$ and $EC_l$ to compute the analytical occurrence probabilities of genetic motifs and plot their evolution at time $t$.
or sequence length \( l \). An example for the model \( \text{TECt} \) is shown in Fig. 4.

### 4.4.1. Initial motif occurrence probabilities, motif insertion rates and deletion rate

The first user step consists in selecting the genetic motif size \( n \) and uploading a parameter file containing the initial motif occurrence probabilities, the motif insertion rates and the deletion rate. An example file of initial dinucleotide occurrence probabilities, dinucleotide insertion rates and deletion rate for the model \( \text{TECt} \) is presented in Fig. 5.

This file must contain \( 4^n \) lines, i.e. one line per motif of size \( n \). Whatever the model chosen, the two first elements of each line \( k \), \( 1 \leq k \leq 4^n \), are the type and the initial occurrence probability of the \( k \)th motif of size \( n \) in lexicographical order. These \( 4^n \) initial motif occurrence probabilities in the five models \( \text{TECt, TECi, Tt, ECl and ECI} \) are the elements of the vectors \( P(0) \) in Eq. (3.2), \( P(0) \) in Eq. (3.3), \( P(0) \) in Eq. (3.4), \( \tau(0) \) in Eq. (3.5) and \( P(0) \) in Eq. (3.6), respectively. In the four models \( \text{TECt, TECi, Tt and ECI} \), the two next elements of a line \( k \) are the type and the insertion rate of the \( k \)th motif of size \( n \). These \( 4^n \) motif insertion rates in the four models \( \text{TECt, TECi, Tt and ECI} \) are the elements of the vector \( R \) in Eqs. (3.2), (3.3), (3.5) and (3.6), respectively. In the two models \( \text{TCl and ECI} \), the last two elements of the first line are the symbol “d” and the deletion rate. This deletion rate in the two models \( \text{TCl and ECI} \) is the term \( d \) in Eqs. (3.3) and (3.6), respectively. Note that for each line the element separator is a tabulation. A link to a pattern parameter file is available in the Evolution Plots Upload interface for each model and motif size (line above the submit button in Fig. 4).

According to the model chosen, different validity conditions on the initial motif occurrence probabilities, motif insertion rates and deletion rate are given (Fig. 4).

**Remark 9.** Initial motif occurrence probabilities, substitution, insertion and deletion parameters, and time value can be given in decimal or rational format or both. Exact analytical solutions are obtained when all the values are rational.

**Remark 10.** The deletion process, i.e. the deletion rate \( d \), is not involved in the three models \( \text{TCl} \) (Eq. (3.2)), \( \text{Tt} \) (Eq. (3.4)) and \( \text{ECt} \) (Eq. (3.5)). Thus, with these three models, there is no deletion rate in the parameter file.

The values of the parameter file, after its upload, are verified by \( \text{GETEC} \), e.g. a probability value must be decimal or rational in the interval \([0,1]\), the sum of probabilities must be equal to 1, the insertion and deletion values must be positive, etc. If errors are detected, descriptive messages are displayed and the user is invited to upload a new parameter file. In the absence of error, the main interface of the Evolution Plots functionality is displayed with the different functionalities listed below according to the selected model (Fig. 6).

#### 4.4.2. Functionalities for the models TECt, Tt and ECt at time \( t \)

**4.4.2.1. Time direction** (models \( \text{TCl, Tt, ECI} \)). The computation of the analytical occurrence probabilities \( P(t) \) (Eq. (3.2)), \( \tau(t) \) (Eq. (3.4)) and \( P(t) \) (Eq. (3.5)) at time \( t \) can be carried out in direct (past–present) or inverse (present–past) time directions (Fig. 6). By default, the analytical solutions are computed in direct time direction.
Fig. 6. Main interface of the Evolution Plots functionality for the model TECr: (1) choice of the evolution time direction; (2) selection of the number of substitution parameters per motif site; and (3) input values of substitution parameters (decimal or rational format or both).
4.4.2.2. Number of substitution parameters per site (models TECT, Tt).
The number of substitution parameters per motif site \( s \), \( 1 \leq s \leq n \), can be chosen (Fig. 6).

The 3-parameter substitution model (Kimura, 1981) distinguishes the three types of substitution for each motif site \( s \): transitions \( a_s \) (\( A \leftrightarrow G \) and \( C \leftrightarrow T \)), transversions \( b_s \) (\( A \leftrightarrow T \) and \( C \leftrightarrow G \)) and transversions \( c_s \) (\( A \leftrightarrow C \) and \( G \leftrightarrow T \)). This most general substitution model is chosen by default.

The particular substitution models of the 3-parameter model can also be selected. The 2-parameter substitution model (Kimura, 1980) has transitions \( u_s = a_s \) (\( A \leftrightarrow G \) and \( C \leftrightarrow T \)) and transversions \( v_s / 2 = b_s = c_s \) (\( A \leftrightarrow C \), \( A \leftrightarrow T \), \( C \leftrightarrow G \) and \( G \leftrightarrow T \)) for each motif site \( s \). The 1-parameter substitution model (Jukes and Cantor, 1969) has substitutions \( p_s / 3 = a_s = b_s = c_s \) for each motif site \( s \).

4.4.2.3. Values of substitution parameters (models TECT, Tt). The values of substitution parameters can be set formal, rational or decimal or any combination type (Fig. 6). They must be positive and their sum must be less than or equal to 1, otherwise descriptive error messages are displayed. By default, they are left formal.

4.4.2.4. Genetic motifs (models TECT, Tt, ECT). Evolution of up to four genetic motifs can be studied simultaneously (Fig. 7). By default, only one genetic motif is selected, the motif \( A^n = \underbrace{A \cdots A}_{n} \) for the chosen size \( n \).

4.4.2.5. Output format (models TECT, Tt, ECI). The analytical occurrence probabilities can be displayed in four different formats to facilitate their integration in external user-programs: Standard (human-readable), C, Fortran and TeX (Fig. 7). By default, the Standard format is selected (Fig. 8).

4.4.2.6. Optional functionalities for plots and numerical solutions (models TECT, Tt, ECT). When all the model parameters are non-formal, two evolution plots are displayed as function of time: a plot drawing the evolution curves of the studied genetic motifs, i.e. containing up to four curves, and a plot drawing the evolution curve of their sum (see an example in Fig. 9).

(i) Time interval for plots (Fig. 7): the parameters \( t_{\text{min}} \) and \( t_{\text{max}} \) of the time interval \([t_{\text{min}}, t_{\text{max}}]\) can be chosen. They must always be positive in the direct and inverse time directions. By default, plots are drawn in the time interval \([t_{\text{min}}, t_{\text{max}}] = [0, 5]\).

(ii) Scale of y-axis for plots (Fig. 7): the vertical zoom can be selected: full validity range or automatic rescale.

(iii) Time value (Fig. 7): a particular numerical value for the time \( t \geq 0 \) gives the numerical solutions of the occurrence probabilities of the studied genetic motifs and their probability sum.

4.4.3. Functionalities for the models TECI and ECI at sequence length \( l \)

4.4.3.1. Initial sequence length (models TECI, ECI). The analytical occurrence probabilities \( P(l) \) (Eq. (3.3)) and \( P(t) \) (Eq. (3.6)) at sequence length \( l \) are functions of the initial sequence length \( l_0 \) which can be formal or a strictly positive integer. By default, the initial sequence length \( l_0 \) is left formal.

4.4.3.2. Number of substitution parameters per site (model TECI). Similar to the models TECT and Tt (see Section 4.4.2.2).

4.4.3.3. Values of substitution parameters (model TECI). Similar to the models TECT and Tt (see Section 4.4.2.3).

4.4.3.4. Genetic motifs (models TECI, ECI). Similar to the models TECT, Tt and ECT (see Section 4.4.2.4).
Fig. 9. Plots with the model TECt drawing the evolution curves for the dinucleotides AA (up) and AC (bottom) and the evolution curve of their sum (bottom).

4.4.3.5. Output format (models TECt, ECl). Similar to the models TECt, Tr and ECl (see Section 4.4.2.5).

4.4.3.6. Optional functionalities for plots and numerical solutions (models TECt, ECl). When all the model parameters are non-formal, two evolution plots are displayed as function of sequence length: a plot drawing the evolution curves of the studied genetic motifs, i.e. containing up to four curves, and a plot drawing the evolution curve of their sum.

(i) Sequence length interval for plots: the parameters \( l_{\text{min}} \) and \( l_{\text{max}} \) of the sequence length interval \([l_{\text{min}}, l_{\text{max}}]\) can be chosen. They must be strictly positive integers. By default, plots are drawn in the sequence length interval \([1, 10]\).

(ii) Scale of y-axis for plots: similar to the models TECt, Tr and ECl (see Section 4.4.2.6).

(iii) Sequence length value: a particular integer value for the sequence length \( l > 0 \) gives the numerical solutions of the occurrence probabilities of the studied genetic motifs and their probability sum.

5. A bioinformatics application

We provide the detailed procedures of the research software GETEC to retrieve the classical formulas of the 1-parameter substitution model (Jukes and Cantor, 1969) and the 2-parameter substitution model (Kimura, 1980). The functionality Evolution Plots of the model Tr (Transformation; Eq. (3.4)) at time \( t \) allows these classical analytical solutions to be retrieved easily.

5.1. Analytical solutions of the 2-parameter substitution model (Kimura, 1980) using the research software GETEC

The 2-parameter substitution model (Kimura, 1980) is based on a symmetric substitution matrix with two formal parameters for the nucleotide transitions and transversions (Section 4.4.2.2).

5.1.1. First user interface of the model Tr

The following parameters must be selected:

1. Choose the motif size: Nucleotides (1).
2. Upload the initial occurrence probability file: the file Prob0TrL1.txt (Fig. 10) must contain an initial nucleotide occurrence probability equal to 1, e.g. \( P_A(0) = 1 \), and thus, the three other initial nucleotide occurrence probabilities are equal to 0, i.e. \( P_C(0) = P_G(0) = P_T(0) = 0 \).

After having pressed the submit button, a second user interface is available.

5.1.2. Second user interface of the model Tr

The following parameters must be selected:

1. Evolutionary time direction: Direct (past -> present). The user has two possible ways to solve this problem.
   (i) With the model Tr at three parameters:
   2. Number of substitution parameters per motif site: 3 parameters.
   3. Substitution parameters: \( a[1]: a \), \( b[1]: b \) and \( v[1]: v \).
   (ii) With the model Tr at two parameters:
   2. Number of substitution parameters per motif site: 2 parameters.
   3. Substitution parameters: \( u[1]: a \) and \( v[1]: v \).
Results

Analytical solutions (Standard format):

\[
\begin{align*}
\text{ProbtA}(t) &= \frac{1 + E^{-4bt}}{4} \\
\text{ProbtC}(t) &= \frac{1 - E^{-4bt}}{4} \\
\text{ProbtG}(t) &= \frac{1 + E^{-4bt} - 2E^{-2(a+b)t}}{4} \\
\text{ProbtT}(t) &= \frac{1 - E^{-4bt}}{4}
\end{align*}
\]

Fig. 11. The classical analytical solutions of the 2-parameter substitution model (Kimura, 1980; Eq. (1.10) in Yang, 2006) retrieved by the research software GETEC.

Note that the formal writing “v[1]: 2b” is also possible. Note also that the formal writing “v[1]: 2v” or “v[1]: 2v” is not allowed as a Mathematica recursion is generated.

Remark 11. For the 2-parameter substitution model (Kimura, 1980), the parameters are defined as follows: transitions \( u_1 = a \) (\( A \leftarrow\rightarrow G \) and \( C \leftarrow\rightarrow T \)) and transversions \( v_1 = b = c \) (\( A \leftarrow C \), \( A \leftarrow T \), \( C \leftarrow G \) and \( G \leftarrow T \)) for each motif site. In order to express this model as a particular case of the 3-parameter substitution model (Kimura, 1981) (see Section 4.4.2.2). Thus, in order to retrieve the formulas of the 2-parameter substitution model with the model \( Tt \) at two parameters, transversions must be multiplied by 2.

The end of the procedure is identical for the model \( Tt \) at three and two parameters.

4. Choice of the probabilities to study and plot: motif A motif C motif G motif T.
5. Choice of the analytical solutions output format: Standard.

The submit button leads to the following results (Fig. 11) which are the classical analytical solutions of the 2-parameter substitution model (Kimura, 1980; Eq. (1.10) in Yang, 2006). Note that Mathematica puts some positive exponential terms in denominator.

5.2. Analytical solutions of the 1-parameter substitution model (Jukes and Cantor, 1969) using the research software GETEC

The 1-parameter substitution model (Jukes and Cantor, 1969) is based on a symmetric substitution matrix with one formal parameter for all nucleotide substitution types (Section 4.4.2.2).

5.2.1. First user interface of the model \( Tt \)

The procedure is identical to the first user interface of the model \( Tt \) in Section 5.1.1.

5.2.2. Second user interface of the model \( Tt \)

The procedure is similar to the second user interface of the model \( Tt \) in Section 5.1.2 with three possible ways to solve this problem.

(i) With the model \( Tt \) at three parameters:
2. Number of substitution parameters per motif site: 3 parameters.
3. Substitution parameters: \( a[1]: a \), \( b[1]: a \) and \( v[1]: a \).

(ii) With the model \( Tt \) at two parameters:
2. Number of substitution parameters per motif site: 2 parameters.
3. Substitution parameters: \( u[1]: a \) and \( v[1]: 2a \).

(iii) With the model \( Tt \) at one parameter:
2. Number of substitution parameters per motif site: 1 parameter.
3. Substitution parameters: \( p[1]: 3a \).

Fig. 12. The classical analytical solutions of the 1-parameter substitution model (Jukes and Cantor, 1969; Eq. (1.3) in Yang, 2006) retrieved by the research software GETEC.

The 1-parameter substitution model (Jukes and Cantor, 1969) is based on a symmetric substitution matrix with one formal parameter for all nucleotide substitution types (Section 4.4.2.2).

3. Substitution parameters:

Enter values for the substitution parameters. Decimal or integer values will be replaced by the name of the corresponding parameter. All the substitution parameters must be decimal or integer values. All the substitution parameters and their sum must be 3 and 1.

Fig. 13. Partial screenshot of the Evolution Plots functionality of the GETEC software for the models \( Tt \) and \( TECt \) showing the substitution parameter settings \( \kappa_2 \) and \( \kappa_2 \) respectively. 1-parameter substitution model (top of the figure), substitution rates equal to 1 for the site 2 and equal to 0 for the sites 1 and 3. The bottom of the figure shows the selection of the four codons GGA, GGC, GGG and GGT coding the amino acid glycine.
Fig. 14. Evolution curves in bacterial genes of the four codon occurrence probabilities \(P_{GGA}(t), P_{GGC}(t), P_{GGG}(t)\) and \(P_{GGT}(t)\) (top figure) and their probability sum \(P_{Gly}(t)\) of glycine (bottom figure) in the time interval \([0,5]\) with the model \(Tt\) and the substitution configuration \(c_{S1}\) (substitution rates equal to 1 for the codon site 1 and equal to 0 for the codon sites 2 and 3).

The submit button leads to the following results (Fig. 12) which are the classical analytical solutions of the 1-parameter substitution model (Jukes and Cantor, 1969; Eq. (1.3) in Yang, 2006).

Fig. 15. Evolution curves in bacterial genes of the four codon occurrence probabilities \(P_{GGA}(t), P_{GGC}(t), P_{GGG}(t)\) and \(P_{GGT}(t)\) (top figure) and their probability sum \(P_{Gly}(t)\) of glycine (bottom figure) in the time interval \([0,5]\) with the model \(TEC_t\) and the substitution–insertion configuration \(c_{SI1}\) (substitution rates equal to 1 for the codon site 1 and equal to 0 for the codon sites 2 and 3, and codon insertion rate according to Eq. (6.1)).

6. A biological application: evolution of the amino acid glycine in bacterial genes

The research software GETEC allows evolution of genetic motifs to be studied. Thus, it is a general approach as several databases of
genetic motifs are available and many software have been developed for identifying genetic motifs, e.g., the MEME Suite (Bailey et al., 2009). As an example of biological application with the GETEC software, we propose here an evolution study of glycine and its four encoded codons GGA, GGC, GGG, and GGT, in bacterial genes. The main purpose of this example is to provide a sketch of the consequences of adding an insertion process beside a site-specific substitution process on the evolution of glycine and its four encoded codons. Thus, the models \( Tt \) (Transformation) at time \( t \) and \( TECT \) (Transformation Expansion Contraction) at time \( t \) are used for this application. The occurrence probability \( P_{Gly}(t) \) of glycine at time \( t \) in the model \( Tt \) (Eq. (3.4)) is the sum of occurrence probabilities of the four codons coding glycine at time \( t \), i.e. \( P_{Gly}(t) = P_{GGA}(t) + P_{GGC}(t) + P_{GGG}(t) + P_{GGT}(t) \). The occurrence probability \( P_{Gly}(t) \) of glycine at time \( t \) in the model \( TECT \) (Eq. (3.2)) is defined similarly.

6.1. Codon usage in bacterial genes

The codon usage chosen in this example on a large population of bacterial genes (7,851,762 genes, 2,481,566,882 trinucleotides, from Table 2a in Michel, 2015) is given in Appendix C. It is used for defining the initial vectors \( P(0) = P(0) \) of codon occurrence probabilities at time \( t = 0 \) in the models \( Tt \) (Eq. (3.4)) and \( TECT \) (Eq. (3.2)). Thus, the initial occurrence probability \( P_{Gly}(0) = P_{Gly}(0) \) of glycine at time 0 in both models \( Tt \) and \( TECT \) is equal to \( P_{Gly}(0) = P_{GGA}(0) + P_{GGC}(0) + P_{GGG}(0) + P_{GGT}(0) = 0.0123 + 0.0335 + 0.0122 + 0.0176 = 0.0756 \) and is the initial value of glycine in the plot curves (see the Figs. 14 and 15).

6.2. Parameter settings of the models \( Tt \) and \( TECT \)

Both models \( Tt \) and \( TECT \) involve a site-specific substitution process. For the current example, we choose the 1-parameter substitution model (Jukes and Cantor, 1969) extended to codons, i.e., one substitution parameter per codon site. The model \( TECT \) also involves an insertion process. For the current example, we set the codon-specific insertion rate \( r_i \) as follows

\[
r_i = \begin{cases} 
1/64 & \text{if } i \in \{GGA, GGC, GGG, GGT\} \\
0 & \text{otherwise} 
\end{cases} 
\]  

(6.1)

Let \( cs \) (s standing for substitution) and \( csi \) (si standing for substitution–insertion) be the two configurations of the models \( Tt \) and \( TECT \), respectively. Moreover, since the 1-parameter substitution model is chosen, one substitution parameter has to be set per codon site. We make the codon sites to evolve one at a time for both configurations. Thus, three parameter settings per configuration are defined: (i) for the substitution configuration \( cs \): \( cs_1 \) (substitution rates equal to 1 for the codon site 1 and equal to 0 for the codon sites 2 and 3), \( cs_2 \) (substitution rates equal to 1 for the codon site 2 and equal to 0 for the codon sites 1 and 3) and \( cs_3 \) (substitution rates equal to 1 for the codon sites 3 and equal to 0 for the codon sites 1 and 2); (ii) for a substitution–insertion configuration \( csi \): \( csi_1 \), \( csi_2 \) and \( csi_3 \) defined similarly to \( cs_1 \), \( cs_2 \) and \( cs_3 \), respectively, and with a codon insertion rate according to Eq. (6.1). Fig. 13 shows the substitution parameter settings \( cs_2 \) and \( csi_2 \) for the models \( Tt \) and \( TECT \), respectively, in the Evolution Plots functionality of the GETEC software.

6.3. Results

Evolution in bacterial genes of the occurrence probabilities of glycine and its four encoded codon in the models \( Tt \) and \( TECT \) are studied for the six parameter settings \( cs_1 \) and \( csi_1 \), \( 1 \leq j \leq 3 \), respectively. Appendix D gives the numerical solutions of occurrence probabilities \( P_{Gly}(t) \) and \( P_{Gly}(t) \) of glycine in bacterial genes at time \( t \) in the models \( Tt \) and \( TECT \), respectively, with the six parameter settings \( cs_j \) and \( csi_j \), respectively. Fig. 14 generated by the Evolution Plots functionality of the GETEC software represents the evolution curves in bacterial genes of the four codon occurrence probabilities \( PGAA(t) \), \( PGCC(t) \), \( PGCC(t) \) and \( PGGT(t) \), and their probability sum \( P_{Gly}(t) \) of glycine in the time interval \([0,5]\) with the model \( Tt \) and the substitution configuration \( cs_1 \) (substitution rates equal to 1 for the codon site 1 and equal to 0 for the codon sites 2 and 3), chosen as example. Similarly, Fig. 15 represents the evolution curves in bacterial genes of the four codon occurrence probabilities \( PGAA(t) \), \( PGCC(t) \), \( PGGC(t) \) and \( PGGT(t) \), and their probability sum \( P_{Gly}(t) \) of glycine in the time interval \([0,5]\) with the model \( TECT \) and the substitution–insertion configuration \( csi_1 \) (substitution rates equal to 1 for the codon site 1 and equal to 0 for the codon sites 2 and 3, and codon insertion rate according to Eq. (6.1)). Overall, with the substitution model \( Tt \) and \( cs_1 \), the five probabilities \( P_{GAA}(t) \), \( P_{GCC}(t) \), \( P_{GCC}(t) \), \( P_{GCT}(t) \) and \( P_{GTT}(t) \) decrease with time \( t \) up to a horizontal asymptote (Fig. 14). In contrast, with the substitution–insertion model \( TECT \) and \( csi_1 \), these five probabilities \( P_{GAA}(t) \), \( P_{GCC}(t) \), \( P_{GCC}(t) \), \( P_{GCT}(t) \) and \( P_{GTT}(t) \) increase, after a minimum for \( P_{GCC}(t) \), with time \( t \) and tend to a higher limit (not shown in Fig. 15).

Fig. 16 summarizes evolution of the occurrence probabilities \( P_{Gly}(t) \) and \( P_{Gly}(t) \) of glycine in bacterial genes with the substitution
and substitution-insertion models \( Tt \) and \( TEC \), respectively. Evolutionary meaning of curves can be analyzed per evolution process or per motif site.

6.3.1. Evolution process comparison

The occurrence probability of glycine in bacterial genes at any time \( t \) is greater under substitution-insertion (model \( TEC \)) than under “substitution only” (model \( Tt \)), i.e. \( P_{\text{Gly}}(t) > P_{\text{Gly}}(t) \) for each couple of configurations \((cs_j, cs_i)\), \( 1 \leq j \leq 3 \) (Fig. 16), the difference resulting from the additional insertion process as shown previously for the couple \((cs_3, cs_1)\) (Figs. 14 and 15).

6.3.2. Motif site comparison

Under “substitution only” (model \( Tt \) and \( cs \)), the occurrence probability of glycine in bacterial genes for time \( t \) in \([0,5]\) has the lowest value with the 1st codon site and the highest value with the 2nd codon site (Fig. 16). Under substitution-insertion (model \( TEC \) and \( cs \)), \( P_{\text{Gly}}(t) \) in bacterial genes for time \( t \) in \([0.6, 5]\) has a similar probability behavior per site to “substitution only”, but for time \( t \) in \([0.6, 5]\), \( P_{\text{Gly}}(t) \) has the highest value with the 3rd codon site (Fig. 16).

7. Conclusion

The \( GETEC \) model developed here is a model of gene evolution based on substitution, insertion and deletion of genetic motifs. It represents a significant mathematical step for unifying two classes of evolution models which have been developed separately for 20 years: the models of substitution, insertion and deletion of nucleotides and the models of symmetric substitution of genetic motifs (see Introduction). It allows the analysis of genetic motif evolution without alignment or phylogenetic inference. The mathematical construction of the \( GETEC \) model has no relation with the mathematical formulation of alignment and phylogenetic methods. Indeed, the alignment methods (global, local, etc.) rely on a distance or similarity associated to residue costs, the phylogenetic methods are commonly based on parsimony, maximum likelihood (ML), MCMC-based Bayesian inference and distance matrix while the \( GETEC \) model is based on a probabilistic differential equation. Thus, the \( GETEC \) model is an alternative to the alignment and phylogenetic methods for studying gene and genome evolution as it can analyze evolution of genetic motifs in two time directions (past to present and present to past).

So far, the \( GETEC \) model is not able to derive expressions of the genetic motif occurrence probabilities as a function of time or sequence length with insertion, deletion and asymmetric instantaneous substitution probability matrices \( M = \{m_{ij}\}_{1 \leq i, j \leq K} \), where the substitution probability \( Pr(j \rightarrow i) = m_{ij} \) of residue \( i \) into residue \( j \) differs from the substitution probability \( Pr(i \rightarrow j) = m_{ji} \) of residue \( i \) into residue \( j \). Asymmetric substitution matrices constitute an interesting modelling tool for analyzing asymmetric substitution rates which may occur more frequently in some genomes. The \( DISL-HKY \) model (Lèbre and Michel, 2012) allows to derive nucleotide occurrence probabilities as function of time or sequence length with insertion, deletion and asymmetric instantaneous substitution probability matrices \( M \), e.g. with the classical substitution matrix \( HKY \) (Hasegawa et al., 1985). However, its extension to genetic motif occurrence probabilities is an open mathematical problem.

The research software \( GETEC \) we have developed allows the computation of the analytical solutions of this new model and its particular cases: models \( TECi \) (Transformation Expansion Contraction) at time \( t \), \( TEC \) (Transformation Expansion Contraction) at sequence length \( l \), \( Tt \) (Transformation) at time \( t \), \( ECl \) (Expansion Contraction) at time \( t \) and \( ECl \) (Expansion Contraction) at sequence length \( l \). It is freely accessible at \url{http://icube-bioinfo.u-strasbg.fr/webMathematica/GETEC/} or via the web site \url{http://dpt-info.u-strasbg.fr/~michel/}. It allows biologists and bioinformaticians to develop their own gene evolution models. The evolution analysis of nucleotides can now be extended to the evolution study of genetic motifs in two ways: (i) motifs on a given site in a set of sequences, e.g. the dinucleotides in the splice sites, the TATA box, etc., (ii) motifs in one or several sequences (content), e.g. codons in genes, amino acids, etc. In future, we will apply the \( GETEC \) model to study evolution of circular codes and bijective genetic codes (Michel, 2014).

Acknowledgement

We thank the five reviewers for their advice.

Appendix A.

A.1. Proof of Proposition 5

\textbf{Proof.} From Eq. (2.7), the substitution rate matrix \( A_{s} \) for motifs of size \( s \), with \( 1 \leq s \leq n \), can be decomposed into a sum of two matrices as follows

\[
A_{s} = \begin{pmatrix}
A_{s-1} & c_{n-s+1}I_{n-s-1} & a_{n-s+1}I_{n-s-1} & b_{n-s+1}I_{n-s-1} \\
0 & a_{n-s+1} & b_{n-s+1}I_{n-s-1} & a_{n-s+1}I_{n-s-1} \\
b_{n-s+1}I_{n-s-1} & a_{n-s+1}I_{n-s-1} & c_{n-s+1}I_{n-s-1} & A_{s-1} \\
0 & c_{n-s+1} & a_{n-s+1} & b_{n-s+1}
\end{pmatrix}
\]

where the diagonal block matrix with \( A_{s-1} \) on the main diagonal is a matrix of size \( (4^{s-1}, 4^{s-1}) \).

With \( I_{s-1} \) the identity matrix of size \( (4,4) \). Therefore, by definition of the Kronecker sum,

\[
A_{s} = \begin{pmatrix}
0 & c_{n-s+1} & a_{n-s+1} & b_{n-s+1} \\
a_{n-s+1} & 0 & b_{n-s+1} & a_{n-s+1} \\
b_{n-s+1} & a_{n-s+1} & 0 & c_{n-s+1} \\
0 & c_{n-s+1} & a_{n-s+1} & b_{n-s+1}
\end{pmatrix}
\]

\[
\oplus
\]

\[
A_{s-1} \]
Moreover,

\[
N_{n+s+1} = \left( \begin{array}{ccc}
0 & c_{n+s+1} & a_{n+s+1} & b_{n+s+1} \\
c_{n+s+1} & 0 & b_{n+s+1} & a_{n+s+1} \\
a_{n+s+1} & b_{n+s+1} & 0 & c_{n+s+1} \\
b_{n+s+1} & a_{n+s+1} & c_{n+s+1} & 0
\end{array} \right) + d_{n+s+1} \times I_1
\]

with \(d_{n+s+1} = -(a_{n+s+1} + b_{n+s+1} + c_{n+s+1})\). Then, \(A_s\) (Eq. (A.1)) can be expressed as function of \(N_s\) as follows

\[
A_s = (N_{n+s+1} - d_{n+s+1} \times I_1) \oplus A_{s-1}.
\]

Thus,

\[
A_n = \oplus_{s=1}^{n} \left( N_s - d_s \times I_1 + A_0 \times I_n \right)
= \oplus_{s=1}^{n} N_s - \sum_{s=1}^{n} d_s \times I_n + A_0 \times I_n
= \oplus_{s=1}^{n} N_s
\]

as \(A_0 = -\sum_{s=1}^{n} (a_s + b_s + c_s) = \sum_{s=1}^{n} d_s\). The recursive construction of the motif substitution rate matrix \(A_n\) can be written as a Kronecker sum of the nucleotide substitution rate matrices \(N_s\) associated with each site \(s\) (\(1 \leq s \leq n\)) of the motifs of size \(n\).

\[\text{A.2. Construction of the dinucleotide substitution matrix with the Kronecker operators}\]

Construction using Eq. (2.10) of the dinucleotide substitution matrix \(A_2\) (16,16) with the Kronecker operators applied to two matrices \(N_1\) and \(N_2\) of size (4,4) associated to the nucleotide substitution matrices at dinucleotide sites 1 and 2, respectively.

\[
N_1 \oplus N_2 = \left( \begin{array}{cc}
d_1 & c_1 \\
c_1 & d_1
\end{array} \right) \oplus \left( \begin{array}{cc}
d_2 & c_2 \\
c_2 & d_2
\end{array} \right)
\]

with \(d_1 = -(a_1 + b_1 + c_1)\) and \(d_2 = -(a_2 + b_2 + c_2)\). Then,

\[
N_1 \oplus N_2 = \left( \begin{array}{cccc}
d_1 & c_1 & a_1 & b_1 \\
c_1 & d_1 & b_1 & a_1 \\
a_1 & b_1 & d_1 & c_1 \\
b_1 & a_1 & c_1 & d_1
\end{array} \right) \oplus \left( \begin{array}{cccc}
d_2 & c_2 & a_2 & b_2 \\
c_2 & d_2 & b_2 & a_2 \\
a_2 & b_2 & d_2 & c_2 \\
b_2 & a_2 & c_2 & d_2
\end{array} \right)
\]

\[
+ \left( \begin{array}{cccc}
d_1 & c_1 & a_1 & b_1 \\
c_1 & d_1 & b_1 & a_1 \\
a_1 & b_1 & d_1 & c_1 \\
b_1 & a_1 & c_1 & d_1
\end{array} \right) \oplus \left( \begin{array}{cccc}
d_2 & c_2 & a_2 & b_2 \\
c_2 & d_2 & b_2 & a_2 \\
a_2 & b_2 & d_2 & c_2 \\
b_2 & a_2 & c_2 & d_2
\end{array} \right)
\]

where \(I_1\) is the identity matrix of size (4,4).

Thus,

\[
N_1 \oplus N_2 = \left( \begin{array}{cccc}
d_1 & c_1 & a_1 & b_1 \\
c_1 & d_1 & b_1 & a_1 \\
a_1 & b_1 & d_1 & c_1 \\
b_1 & a_1 & c_1 & d_1
\end{array} \right) \oplus \left( \begin{array}{cccc}
d_2 & c_2 & a_2 & b_2 \\
c_2 & d_2 & b_2 & a_2 \\
a_2 & b_2 & d_2 & c_2 \\
b_2 & a_2 & c_2 & d_2
\end{array} \right)
\]

\[
N_1 \oplus N_2 = \left( \begin{array}{cccc}
d_1 & c_1 & a_1 & b_1 \\
c_1 & d_1 & b_1 & a_1 \\
a_1 & b_1 & d_1 & c_1 \\
b_1 & a_1 & c_1 & d_1
\end{array} \right) \oplus \left( \begin{array}{cccc}
d_2 & c_2 & a_2 & b_2 \\
c_2 & d_2 & b_2 & a_2 \\
a_2 & b_2 & d_2 & c_2 \\
b_2 & a_2 & c_2 & d_2
\end{array} \right)
\]

\[
N_1 \oplus N_2 = \left( \begin{array}{cccc}
d_1 & c_1 & a_1 & b_1 \\
c_1 & d_1 & b_1 & a_1 \\
a_1 & b_1 & d_1 & c_1 \\
b_1 & a_1 & c_1 & d_1
\end{array} \right) \oplus \left( \begin{array}{cccc}
d_2 & c_2 & a_2 & b_2 \\
c_2 & d_2 & b_2 & a_2 \\
a_2 & b_2 & d_2 & c_2 \\
b_2 & a_2 & c_2 & d_2
\end{array} \right)
\]

\[
\text{Appendix B. Occurrence probability of a given genetic motif } w \text{ with the models TECl, Ti, ECT and ECI}
\]

Model TECl (Transformation Expansion Contraction) at sequence length \(l\). From Eq. (3.3), the occurrence probability \(P_l(i)\) of a genetic motif \(i\) at sequence length \(l\) is

\[
P_l(i) = \left( \sum_{k=1}^{4^l} \frac{1}{r+1-\lambda_k} \delta[O_k[i, j] - 1] \cdot R \right) \cdot \left( P(l) - \frac{1}{r+1-\lambda_k} R \right) \left( \frac{1}{l} \right)^{\frac{1}{r+1-\lambda_k}}.
\]
Appendix C. Codon usage in bacterial genes

See Table 1.

Table 1  Codon usage (%) in bacterial genes (7,851,762 genes, 2,481,566,882 trinucleotides, from Table 1 in Michel, 2015). It is used for defining the initial vectors \( P_l(0) \) of codon occurrence probabilities at time or in the models \( Tr(T) \) (Transformation–Expansion Contraction; Eq. (3.4)) and ECT (Transformation Expansion Contraction; Eq. (3.2)).

<table>
<thead>
<tr>
<th>Codon i</th>
<th>( P_l(0) )</th>
<th>Codon i</th>
<th>( P_l(0) )</th>
<th>Codon i</th>
<th>( P_l(0) )</th>
<th>Codon i</th>
<th>( P_l(0) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>2.87 CAA</td>
<td>1.61 GAA</td>
<td>3.47 TAA</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAC</td>
<td>1.79 CAG</td>
<td>1.05 CAG</td>
<td>2.63 CAG</td>
<td>1.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAG</td>
<td>1.97 CAG</td>
<td>2.18 CAG</td>
<td>2.62 CAG</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAT</td>
<td>1.93 CAT</td>
<td>1.06 CAT</td>
<td>2.80 TAT</td>
<td>1.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACA</td>
<td>1.00 CCA</td>
<td>0.77 CCA</td>
<td>1.69 TCA</td>
<td>0.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>2.12 CCC</td>
<td>1.08 CCC</td>
<td>3.54 TCC</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACG</td>
<td>1.39 CCG</td>
<td>1.88 CCG</td>
<td>3.06 TCG</td>
<td>1.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT</td>
<td>0.90 CCT</td>
<td>0.81 CTT</td>
<td>1.60 TCT</td>
<td>0.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGA</td>
<td>0.54 CGA</td>
<td>0.42 CGA</td>
<td>1.23 CGA</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGC</td>
<td>1.39 CCC</td>
<td>2.25 CCC</td>
<td>3.35 TGC</td>
<td>0.56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGG</td>
<td>0.32 CGG</td>
<td>1.08 CGG</td>
<td>1.22 CGG</td>
<td>1.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGT</td>
<td>0.81 CGT</td>
<td>1.10 CTT</td>
<td>1.76 GTT</td>
<td>0.38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATA</td>
<td>0.89 CTA</td>
<td>0.56 CTA</td>
<td>1.08 TTA</td>
<td>1.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATC</td>
<td>2.71 TCT</td>
<td>1.73 TCT</td>
<td>2.04 TTC</td>
<td>1.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATT</td>
<td>2.34 CTT</td>
<td>3.66 CTT</td>
<td>2.58 TGG</td>
<td>1.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATT</td>
<td>2.43 CTT</td>
<td>1.28 CTT</td>
<td>1.52 TTT</td>
<td>2.01</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Appendix D. Evolution of glycine in bacterial genes

See Table 2.

Table 2  Occurrence probabilities \( P_{CG}(t) \) and \( P_{CG}(t) \) of glycine in bacterial genes at time or in the models \( Tr(T) \) (Transformation; Eq. (3.4)) and ECT (Transformation Expansion Contraction; Eq. (3.2)), respectively, with the six configurations \( cs_1 \) and \( cs_3 \) substitution configurations \( cs_2 \) (substitution rates equal to 1 for the codon site 1 and equal to 0 for the codon sites 2 and 3), \( cs_1 \) (substitution rates equal to 1 for the codon site 2 and equal to 0 for the codon sites 1 and 3) and \( cs_3 \) (substitution rates equal to 1 for the codon site 3 and equal to 0 for the codon sites 1 and 2), and substitution–insertion configurations \( cs_1 \) and \( cs_2 \) defined similarly to \( cs_1 \) and \( cs_3 \) respectively, and with a codon insertion rate according to Eq. (6.1).

<table>
<thead>
<tr>
<th>Model</th>
<th>Configuration</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Tr )</td>
<td>( cs_1 )</td>
<td>( P_{CG}(t) = 0.0001327 \cdot 10^{-4} \cdot 10^{-4} \cdot 10^{-4} )</td>
</tr>
<tr>
<td>( Tr )</td>
<td>( cs_3 )</td>
<td>( P_{CG}(t) = 0.0001327 \cdot 10^{-4} \cdot 10^{-4} \cdot 10^{-4} )</td>
</tr>
<tr>
<td>( ECT )</td>
<td>( cs_1 )</td>
<td>( P_{CG}(t) = 0.0001327 \cdot 10^{-4} \cdot 10^{-4} \cdot 10^{-4} )</td>
</tr>
<tr>
<td>( ECT )</td>
<td>( cs_3 )</td>
<td>( P_{CG}(t) = 0.0001327 \cdot 10^{-4} \cdot 10^{-4} \cdot 10^{-4} )</td>
</tr>
</tbody>
</table>


