

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

ARTICLE INFO

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

ABSTRACT

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/>.

© 2015 Elsevier Ireland Ltd. All rights reserved.

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 $\{R, Y\}$ ($R = \{A, G\}$, $Y = \{C, T\}$) (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

* Corresponding author. Tel.: +33 368854462.

E-mail addresses: emmanuel.benard@tuebingen.mpg.de (E. Benard), slebre@unistra.fr (S. Lèbre), c.michel@unistra.fr (C.J. Michel).

¹ Present address: Research Group Neher, Max Planck Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany.

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 $P_t(i, j)$ of residue i by residue j 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_s, b_s, c_s) 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 $P(t_0)$ of initial motif content in the sequence at evolution time t_0 or the vector $P(n_0)$ of initial motif content at sequence length n_0 . 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*

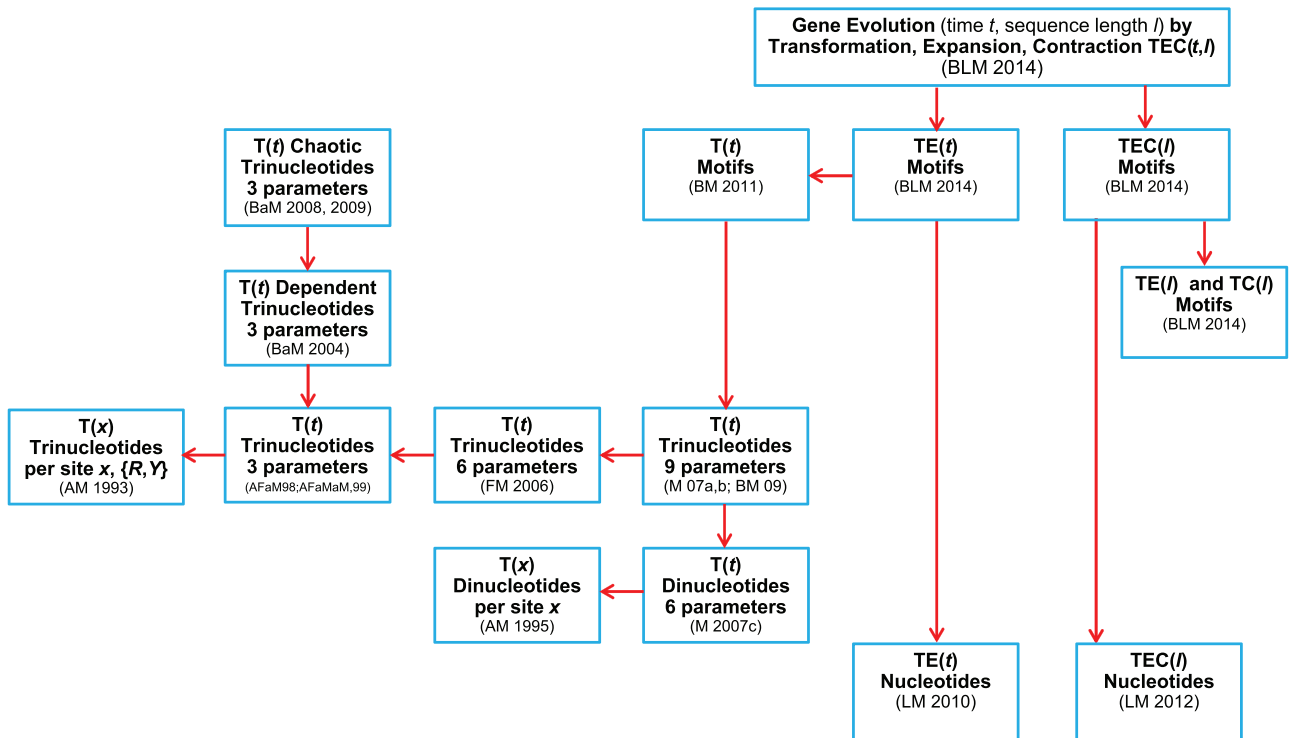


Fig. 1. Substitution models of motifs and substitution–insertion–deletion models of nucleotides. The hierarchy of models is given from top (more general models) to bottom (particular models). The bottom “row” corresponds to the nucleotide evolution models, the rows above to dinucleotide and trinucleotide evolution models and upper rows feature motif evolution models.

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 (Jukes and Cantor, 1969) 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

2.1. Substitution–insertion–deletion models for nucleotides (Lèbre and Michel, 2010, 2012)

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 $t \geq 0$ 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,

$$P'(t) = \underbrace{(M - I) \cdot P(t)}_{\text{Substitution}} + \underbrace{(-rP(t) + R)}_{\text{Insertion-Deletion}} \quad (2.1)$$

$$= A \cdot P(t) + R$$

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)

$$P'(t) = M \cdot P(t) - P(t) = (M - I) \cdot P(t) \quad (2.2)$$

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 $\pi = [\Pr(i \rightarrow j)]_{1 \leq i, j \leq K}$ which is stochastic in line (e.g. Kimura, 1980, 1981), i.e. $\pi_{ij} = \Pr(i \rightarrow j) = m_{ji}$.

(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 $n'_i(t) = \frac{\partial n_i(t)}{\partial t}$ 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 $1 \leq i \leq K$,

$$n'_i(t) = r_i \times n(t) - d \times n_i(t). \quad (2.3)$$

The derivative $P'(t)$ of the occurrence probability of residue i at time t can be written

$$\begin{aligned} p'_i(t) &= \frac{\partial}{\partial t} \left(\frac{n_i(t)}{n(t)} \right) \\ &= \frac{1}{n^2(t)} \left[(r_i n(t) - d n_i(t)) n(t) - n_i(t) \sum_{1 \leq j \leq K} n'_j(t) \right]. \end{aligned}$$

By replacing $n'_j(t)$ 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

$$P'(t) = -rP(t) + R. \quad (2.4)$$

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 IDIS 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 eigenvalue λ_k . Then, for any (non-zero) residue insertion rate vector $R = [r_i]_{1 \leq i \leq K}$ such that $\forall 1 \leq i \leq K$, $r_i \geq 0$, and the total insertion rate $r = \sum_{1 \leq i \leq K} r_i > 0$, the residue occurrence probability $P(t)$ at time t is

$$P(t) = Q \cdot D_1(t) \cdot Q^{-1} \cdot P(t_0) + Q \cdot D_2(t) \cdot Q^{-1} \cdot R \quad (2.5)$$

where $D_1(t) = \text{Diag} \left((e^{-(r+1-\lambda_k)(t-t_0)})_{1 \leq k \leq K} \right)$, $D_2(t) = \text{Diag} \left(\left(\frac{1}{r+1-\lambda_k} (1 - e^{-(r+1-\lambda_k)(t-t_0)}) \right)_{1 \leq k \leq K} \right)$ and $P(t_0) = [p_i(t_0)]_{1 \leq i \leq K}$ is the initial residue occurrence probability at time t_0 .

Proof. By extending Eq. (2.11) in Lèbre and Michel (2010) to any initial time t_0 . \square

From Eq. (2.5), we derive in Proposition 2 a general formula giving an analytical expression of the residue occurrence probability as a function of time t or sequence length l 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 to $h(x, x_0) = \left(\frac{l}{l_0} \right)^{-\frac{1}{r-d}}$ 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 $r = \sum_{1 \leq i \leq K} r_i > 0$, deletion rate $d \geq 0$ and initial residue occurrence probability $P(t_0) = [p_i(t_0)]_{1 \leq i \leq K}$ at time t_0 , the residue occurrence probability $P(x)$ as function of a variable x representing time $x=t$ or sequence length $x=l$ with the following convention $(x, x_0, h(x, x_0)) = (t, t_0, e^{-(t-t_0)})$ for time expression and $(x, x_0, h(x, x_0)) = \left(l, l_0, \left(\frac{l}{l_0} \right)^{-\frac{1}{r-d}} \right)$ for sequence length expression is

$$\begin{aligned} P(x) &= \left(\sum_{k=1}^K \frac{1}{r+1-\lambda_k} O_k \right) \cdot R \\ &+ \sum_{k=1}^K O_k \cdot \left(P(x_0) - \frac{1}{r+1-\lambda_k} R \right) h(x, x_0)^{r+1-\lambda_k} \end{aligned} \quad (2.6)$$

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

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

with $1_k = (\delta_{i,k}) = (0, \dots, 0, 1, 0, \dots, 0)^T$, a vector having 1 in k th row and 0 otherwise, and $(1_k)^T$, 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 $e^{-(t-t_0)} = \left(\frac{n(t)}{n(t_0)} \right)^{-\frac{1}{r-d}} = \left(\frac{l}{l_0} \right)^{-\frac{1}{r-d}}$ (see also Eq. (10) in Lèbre and Michel, 2012, for the particular case $t_0=0$).

\square

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 ($t > t_0$, $l > l_0$) or inverse ($t < t_0$, $l < l_0$) direction of evolution. In the direct evolution direction, $P(t)$ and $P(l)$ converge to the residue equilibrium distribution when t 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 $\{O_k\}_k$ is $\sum_{k=1}^K O_k = Q \cdot Q^{-1} = I$. Indeed, for all i, j , $\sum_{k=1}^K O_k[i, j] = \sum_{k=1}^K Q[i, k] \cdot Q^{-1}[k, j]$ 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

the ratio $\frac{1}{r+1-\lambda_k}$ is different from zero as the eigenvalues of the stochastic matrix M satisfies $\lambda_k \leq 1, \forall 1 \leq k \leq K$. If the insertion rate vector R is null, then the residue occurrence probability $P(t)$ satisfies $P(t) = Q \cdot D_1(t) \cdot Q^{-1} \cdot P(t_0)$ with $D_1(t) = \text{Diag}((e^{-(1-\lambda_k)(t-t_0)})_{1 \leq k \leq K})$ as in the “substitution only” model (Michel, 2007a).

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

$$P(t) = \left(\sum_{k=1}^K O_k e^{-(1-\lambda_k)(t-t_0)} \right) \cdot P(t_0)$$

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 $O_k = Q \cdot 1_k \cdot (1_k)^T \cdot Q^{-1}$ with Q , the eigenvector matrix of matrix M , and $1_k = (\delta_{i,k}) = (0, \dots, 0, 1, 0, \dots, 0)^T$, 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 $P(x)$ at time $x=t$ or sequence length $x=l$ with the following convention $(x, x_0, h(x, x_0)) = (t, t_0, e^{-(t-t_0)})$ for time expression and $(x, x_0, h(x, x_0)) = (l, l_0, (\frac{l}{l_0})^{-\frac{1}{r-d}})$ for sequence length expression is equal to

$$P(x) = \frac{R}{r} + \left(P(x_0) - \frac{R}{r} \right) h(x, x_0)^r$$

where $R = [r_i]_{1 \leq i \leq K}, \forall 1 \leq i \leq K, r_i \geq 0$, 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 eigenelements analytically. It was found by Benard and Michel (2011) after a detailed analysis of the dinucleotide matrix δ (Fig. 1 in Michel, 2007c) and the trinucleotide matrix δ (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_s associated to each site s 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_s, b_s and c_s be the parameters of transitions $A \leftrightarrow G$ and $C \leftrightarrow T$, transversions I $A \leftrightarrow T$ and $C \leftrightarrow G$ and transversions II $A \leftrightarrow C$ and $G \leftrightarrow T$, respectively. For example, when considering a dinucleotide $w = l_1 l_2$ 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 $3n$ substitution parameters. Let us denote by

$A_n = M_n - I_n$ of size $(4^n, 4^n)$ the symmetric substitution rate matrix of motifs 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)

$$A_s = \begin{pmatrix} A_{s-1} & c_{n-s+1} I_{s-1} & a_{n-s+1} I_{s-1} & b_{n-s+1} I_{s-1} \\ c_{n-s+1} I_{s-1} & A_{s-1} & b_{n-s+1} I_{s-1} & a_{n-s+1} I_{s-1} \\ a_{n-s+1} I_{s-1} & b_{n-s+1} I_{s-1} & A_{s-1} & c_{n-s+1} I_{s-1} \\ b_{n-s+1} I_{s-1} & a_{n-s+1} I_{s-1} & c_{n-s+1} I_{s-1} & A_{s-1} \end{pmatrix} \quad (2.7)$$

where I_{s-1} is the identity matrix of size $(4^{s-1}, 4^{s-1})$ with $I_0 = 1, A_{s-1}$ is the recursive matrix of size $(4^{s-1}, 4^{s-1})$ 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 $N_s, 1 \leq s \leq n$, be the nucleotide substitution rate matrix of size $(4,4)$ of a site s of a motif of size n

$$N_s = \begin{pmatrix} d_s & c_s & a_s & b_s \\ c_s & d_s & b_s & a_s \\ a_s & b_s & d_s & c_s \\ b_s & a_s & c_s & d_s \end{pmatrix}$$

with $d_s = -(a_s + b_s + c_s)$. As the matrix N_s is real and symmetric, N_s 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_s is

$$S_s = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -2(a_s + b_s) & 0 & 0 \\ 0 & 0 & -2(a_s + c_s) & 0 \\ 0 & 0 & 0 & -2(b_s + c_s) \end{pmatrix} \quad (2.8)$$

and its associated nucleotide eigenvector matrix Q is

$$Q = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix}. \quad (2.9)$$

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

Remark 6. $Q^{-1} = \frac{1}{4} Q$.

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

$$\begin{cases} D_n = \oplus_{s=1}^n S_s \\ Q_n = \otimes_{s=1}^n Q \\ Q_n^{-1} = (\otimes_{s=1}^n Q)^{-1} = \otimes_{s=1}^n Q^{-1} = \otimes_{s=1}^n \left(\frac{1}{4} Q \right) = \frac{1}{4^n} \otimes_{s=1}^n Q = \frac{1}{4^n} Q_n \end{cases}$$

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

sum of the n nucleotide spectral matrices S_s and the Kronecker product of the n nucleotide eigenvector matrix Q as follows

$$A_n = \otimes_{s=1}^n Q \cdot \oplus_{s=1}^n S_s \cdot \otimes_{s=1}^n Q^{-1}.$$

Moreover,

$$A_n = \oplus_{s=1}^n N_s. \tag{2.10}$$

Appendix A gives the proof of Proposition 5 and an explicit example of construction of a dinucleotide substitution matrix with the Kronecker operators, i.e. with the Kronecker sum \oplus and the Kronecker product \otimes which are rarely used in the bioinformatics research field.

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 evolution 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 TECl (Transformation Expansion Contraction) at sequence length l , and for its particular cases: “substitution only” model Tt (Transformation) at time t , “insertion–deletion only” model ECt (Expansion Contraction) at time t and “insertion–deletion only” model ECl (Expansion Contraction) at sequence length l . All these models are implemented in the research software GETEC (Section 4).

3.1. Construction of the GETEC model

Proposition 6. Let us denote by $M_n = A_n + I_n$ 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 spectral matrix $D_n + I_n$ and eigenvectors matrix Q_n such that $D_n = \oplus_{s=1}^n S_s$ and $Q_n = \otimes_{s=1}^n Q$ where S_s 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

$$\begin{aligned} M_n &= A_n + I_n \\ &= Q_n \cdot D_n \cdot Q_n^{-1} + I_n \\ &= Q_n \cdot D_n \cdot Q_n^{-1} + Q_n \cdot I_n \cdot Q_n^{-1} \\ &= Q_n \cdot (D_n + I_n) \cdot Q_n^{-1} \end{aligned}$$

where $D_n = \oplus_{s=1}^n S_s$ and $Q_n = \otimes_{s=1}^n Q$ results from Proposition 5. \square

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

$$\begin{cases} \lambda_k = 1 + D_n[k, k] \\ O_k = \frac{1}{4^n} \otimes_{s=1}^n Q \cdot 1_k \cdot (1_k)^T \cdot \otimes_{s=1}^n Q \end{cases} \tag{3.1}$$

where $D_n = (\oplus_{s=1}^n S_s)$, S_s 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)).

3.2. 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 TECl (Transformation Expansion Contraction) at sequence length l , and the particular cases: Tt (Transformation) at time t , ECt (Expansion Contraction) at time t and ECl (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

$$\begin{aligned} P(t) &= \left(\sum_{k=1}^{4^n} \frac{1}{r+1-\lambda_k} O_k \right) \cdot R \\ &+ \sum_{k=1}^{4^n} O_k \cdot \left(P(0) - \frac{1}{r+1-\lambda_k} R \right) e^{-(r+1-\lambda_k)t} \end{aligned} \tag{3.2}$$

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 genetic motif insertion rate with $r > 0$ and for all $1 \leq k \leq 4^n$

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

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

$$\lambda_k = 1 + (\oplus_{s=1}^n S_s) [k, k],$$

with

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

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

(present–past) can be deduced from the direct evolution direction (past–present) in Eq. (3.2) by replacing t by $-t$.

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

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

3.2.2. Model TECl (Transformation Expansion Contraction) at sequence length l

Using Eq. (2.6) and the relations (3.1), the occurrence probability $P(l)$ of genetic motifs of size n at sequence length l with the initial condition $P(l_0)$ at sequence length l_0 is

$$P(l) = \left(\sum_{k=1}^{4^n} \frac{1}{r+1-\lambda_k} O_k \right) \cdot R + \sum_{k=1}^{4^n} O_k \cdot \left(P(l_0) - \frac{1}{r+1-\lambda_k} R \right) \left(\frac{l}{l_0} \right)^{-\frac{r+1-\lambda_k}{r-d}} \quad (3.3)$$

where λ_k , O_k , R and r are defined in Eq. (3.2) and d is the deletion rate.

Remark 7. The length l_0 cannot be equal to 0, in contrast to the time t_0 .

If $(r-d) > 0$ then the sequence length l increases else decreases.

The occurrence probabilities of a given genetic motif i with the *TECl* model and the particular cases of the *GETEC* model described below in Section 3.3, implemented in the *GETEC* software, are given in Appendix B.

The substitution probability matrix $M = [\text{Pr}(j \rightarrow i)]_{1 \leq i, j \leq K}$, the insertion rates $R = [r_i]_{1 \leq i \leq K}$ and the deletion rate d are specific parameters of the *GETEC* model which can be determined from genomic data extracted, e.g. from databases, using these analytical expressions. For example, a best fit curve minimizing the error RSS (Residual Sum of Squares) can be estimated from genomic data, such as the GC content as function of the genome length l (see Lèbre and Michel, 2013, Section 7).

3.3. Particular cases of the *GETEC* model

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

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

$$\mathcal{P}(t) = \left(\sum_{k=1}^{4^n} O_k e^{-(1-\lambda_k)t} \right) \cdot \mathcal{P}(0) \quad (3.4)$$

where λ_k and O_k are defined in Eq. (3.2).

Proof. In absence of insertion, then $r=0$ and vector R is null. Eq. (3.2) leads to Eq. (3.4) immediately. □

If $t \geq 0$ then the evolution direction is direct else inverse.

3.3.2. “Insertion–deletion only” models Ect and ECl

3.3.2.1. Model Ect (Expansion Contraction) at time t . The occurrence probability $\mathbf{P}(t)$ of genetic motifs of size n at time t with the initial condition $\mathbf{P}(0)$ is

$$\mathbf{P}(t) = \frac{R}{r} + \left(\mathbf{P}(0) - \frac{R}{r} \right) e^{-rt} \quad (3.5)$$

where R and r are defined in Eq. (3.2).

Proof. The absence of substitution is associated to a substitution matrix equal to the identity matrix. Then, for all $1 \leq k \leq K$, $\lambda_k = 1$. Hence, $\frac{1}{r+1-\lambda_k} = \frac{1}{r}$. From Remark 2, the sum of the matrices $\{O_k\}_k$, is $\sum_{k=1}^K O_k = Q \cdot Q^{-1} = I$. Consequently, Eq. (3.2) leads to Eq. (3.5). □

If $t \geq 0$ then the evolution direction is direct else inverse.

3.3.2.2. Model ECl (Expansion Contraction) at sequence length l . The occurrence probability $\mathbf{P}(l)$ of genetic motifs of size n at sequence length l with the initial condition $\mathbf{P}(l_0)$ is

$$\mathbf{P}(l) = \frac{R}{r} + \left(\mathbf{P}(l_0) - \frac{R}{r} \right) \left(\frac{l}{l_0} \right)^{-\frac{r}{r-d}} \quad (3.6)$$

where R and r are defined in Eq. (3.2).

Proof. Similar to the proof with the model *Ect* applied to Eq. (3.3). □

If $(r-d) > 0$ then the sequence length l increases else decreases.

Remark 8. The Kronecker operators are absent in the models *Ect* and *ECl*.

The analytical solutions of the models *TEct* at time t and *TECl* at sequence length l , the “substitution only” model *Tt* at time t and the “insertion–deletion only” models *Ect* at time t and *ECl* at sequence length l , all in both directions (direct and inverse), are implemented in the *GETEC* software. Thus, five classes of analytical formulas of genetic motif evolution are available to the biological community for analyzing their own gene evolution problems.

3.4. Relation between time t and sequence length l in the *GETEC* model

From the growth rate $n'_i(t)$ 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

$$\forall t \geq 0, l = l_0 e^{(r-d)t}$$

where l_0 is the sequence length at time $t=0$. Thus,

$$t = \frac{\ln l - \ln l_0}{(r-d)}.$$

In an insertion–deletion process with dominant insertion, i.e. $(r-d) > 0$, then $\ln l > \ln l_0$ and the sequence length l increases. In contrast, in an insertion–deletion process with dominant deletion, i.e. $(r-d) < 0$, then $\ln l < \ln l_0$ and the sequence length l decreases.

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/> or via the web site <http://dpt-info.u-strasbg.fr/~michel/> (Fig. 2). It is a major extension of the research software *SEGM* (Stochastic Evolution of Genetic Motifs) (Benard and Michel, 2011). To our knowledge, it is to date the only computational biological software in this evolution field. Thus, a brief



Emmanuel Benard, Sophie Lèbre and Christian J. Michel

Theoretical Bioinformatics, ICube, University of Strasbourg, CNRS

Select a model

Model TEC at time t :

Transformation Expansion Contraction <i>time</i>	Evolution Plots
	Formal Analytical Solutions

Model TEC at sequence length l :

Transformation Expansion Contraction <i>length</i>	Evolution Plots
	Formal Analytical Solutions

Model T at time t :

Transformation <i>time</i>	Evolution Plots
	Formal Analytical Solutions

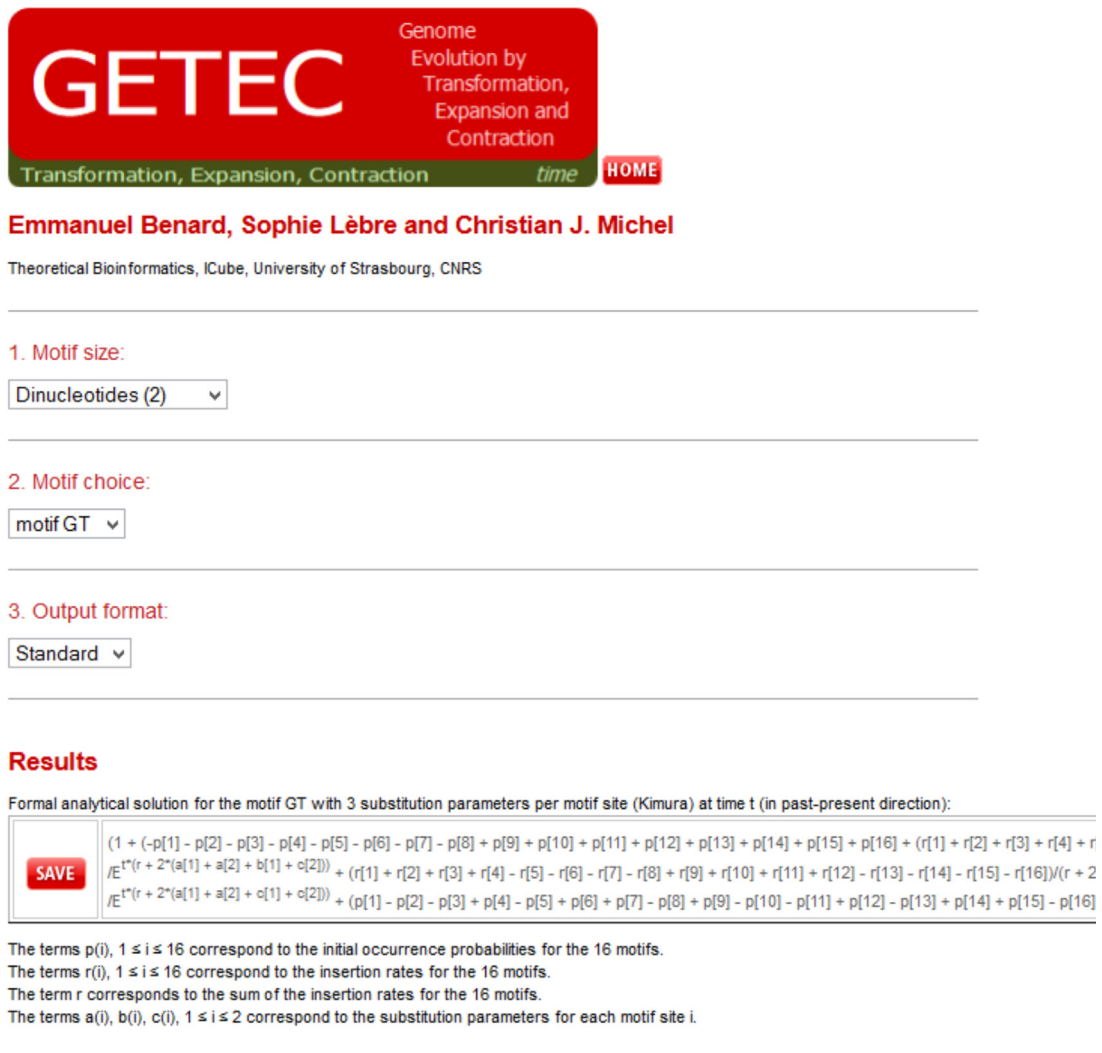
Model EC at time t :

Expansion Contraction <i>time</i>	Evolution Plots
	Formal Analytical Solutions

Model EC at sequence length l :

Expansion Contraction <i>length</i>	Evolution Plots
	Formal Analytical Solutions

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 *TEC*, *TECl*, *Tt*, *ECt* and *ECl*.



GETEC Genome Evolution by Transformation, Expansion and Contraction
Transformation, Expansion, Contraction time HOME

Emmanuel Benard, Sophie Lèbre and Christian J. Michel
Theoretical Bioinformatics, ICube, University of Strasbourg, CNRS

1. Motif size:
Dinucleotides (2) ▾

2. Motif choice:
motif GT ▾

3. Output format:
Standard ▾

Results

Formal analytical solution for the motif GT with 3 substitution parameters per motif site (Kimura) at time t (in past-present direction):

SAVE

$$\frac{(1 + (-p[1] - p[2] - p[3] - p[4] - p[5] - p[6] - p[7] - p[8] + p[9] + p[10] + p[11] + p[12] + p[13] + p[14] + p[15] + p[16] + (r[1] + r[2] + r[3] + r[4] + r[5] + r[6] + r[7] + r[8] + r[9] + r[10] + r[11] + r[12] - r[13] - r[14] - r[15] - r[16]))/r + 2)}{E^{t(r + 2^*(a[1] + a[2] + b[1] + c[2]))} + (p[1] - p[2] - p[3] + p[4] - p[5] + p[6] + p[7] - p[8] + p[9] - p[10] - p[11] + p[12] - p[13] + p[14] + p[15] - p[16])}$$

The terms $p(i)$, $1 \leq i \leq 16$ correspond to the initial occurrence probabilities for the 16 motifs.
The terms $r(i)$, $1 \leq i \leq 16$ correspond to the insertion rates for the 16 motifs.
The term r corresponds to the sum of the insertion rates for the 16 motifs.
The terms $a(i)$, $b(i)$, $c(i)$, $1 \leq i \leq 2$ correspond to the substitution parameters for each motif site i .

Fig. 3. Screenshot of the Formal Analytical Solutions interface for the model *TEct*. Example with the dinucleotide *GT* in Standard output format.

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 *TEct* (Transformation Expansion Contraction; Eq. (3.2)) at time t and *TECl* (Transformation Expansion Contraction; Eq. (3.3)) at sequence length l . The particular models are the “substitution only” model *Tt* (Transformation; Eq. (3.4)) at time t and the “insertion–deletion only” models *Ect* (Expansion Contraction; Eq. (3.5)) at time t and *ECl* (Expansion Contraction; Eq. (3.6)) at sequence length l . For the five models *TEct*, *TECl*, *Tt*, *Ect* and *ECl*, 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 *TEct* and *TECl*, the genetic motif sizes allowed are length 1, i.e. the four genetic motifs $\{A, \dots, T\}$, to 4, i.e.

the 256 genetic motifs $\{AAAA, \dots, TTTT\}$. For the particular models *Tt*, *Ect* and *ECl*, the genetic motifs can have a size up to 5, i.e. the 1024 genetic motifs $\{AAAAA, \dots, TTTTT\}$. 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 *TEct*, *Tt* and *Ect* at time $t \geq 0$ (in the direct evolution direction), and for the models *TECl* and *ECl* at sequence length $l \geq l_0$ when $(r - d) \geq 0$. An example with the dinucleotide *GT* for the model *TEct* 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 *TEct*, *TECl*, *Tt*, *Ect* and *ECl* to compute the analytical occurrence probabilities of genetic motifs and plot their evolution at time t

Fig. 4. Screenshot of the Evolution Plots interface for the model *TECt*: selection of the genetic motif size (1–4) and upload of the parameter file containing the initial motif occurrence probabilities, the motif insertion rates and the deletion rate.

or sequence length l . An example for the model *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 *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 *TECt*, *TECl*, *Tt*, *ECt* and *ECl* are the elements of the vectors $P(0)$ in Eq. (3.2), $P(l_0)$ in Eq. (3.3), $\mathcal{P}(0)$ in Eq. (3.4), $\mathbf{P}(0)$ in Eq. (3.5) and $\mathbf{P}(l_0)$ in Eq. (3.6), respectively. In the four models *TECt*, *TECl*, *ECt* and *ECl*, 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 *TECt*, *TECl*, *ECt* and *ECl* are the elements of the vector R in Eqs. (3.2), (3.3), (3.5) and (3.6), respectively. In the two models *TECl* and *ECl*, the last two elements of the first line are the symbol “d” and the deletion rate. This deletion rate in the two models *TECl* and *ECl* is the term

POAA	1/5	rAA	0.1	d	0.1
POAC	0	rAC	2/235		
POAG	0.2	rAG	0.5		

Fig. 5. Three first lines of an example parameter file of initial dinucleotide occurrence probabilities, dinucleotide insertion rates and deletion rate for the model *TECl*. Element separator is a tabulation. Values can be rational, decimal or both.

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 *TECt* (Eq. (3.2)), *Tt* (Eq. (3.4)) and *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 *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 *TECt*, *Tt*, *ECt*). The computation of the analytical occurrence probabilities $P(t)$ (Eq. (3.2)), $\mathcal{P}(t)$ (Eq. (3.4)) and $\mathbf{P}(t)$ (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.



Stochastic Evolution of Dinucleotides HOME

Emmanuel Benard, Sophie Lèbre and Christian J. Michel

Theoretical Bioinformatics, ICube, University of Strasbourg, CNRS

Upload new initial occurrence probabilities and insertion rates?

Uploaded file info:

✓

Initial occurrence probabilities and insertion rates file valid

1. Evolutionary time direction:

Direct (past -> present) ▾

2. Number of substitution parameters per motif site $1 \leq x \leq 2$:

3 parameters:

- **a[x]: 1 transition rate ($A \leftrightarrow G = C \leftrightarrow T$)**
- **b[x]: 1 transversion I rate ($A \leftrightarrow T = C \leftrightarrow G$)**
- **c[x]: 1 transversion II rate ($A \leftrightarrow C = G \leftrightarrow T$)**

2 parameters:

- **u[x]=a[x]: 1 transition rate ($A \leftrightarrow G = C \leftrightarrow T$)**
- **v[x]/2=b[x]=c[x]: 1 transversion rate ($A \leftrightarrow T = A \leftrightarrow C = C \leftrightarrow G = G \leftrightarrow T$)**

1 parameter:

- **p[x]/3=a[x]=b[x]=c[x]: 1 substitution rate ($A \leftrightarrow C = A \leftrightarrow G = A \leftrightarrow T = C \leftrightarrow G = C \leftrightarrow T = G \leftrightarrow T$)**

3 parameters ▾

3. Substitution parameters:

Enter values for the substitution parameters.
Non decimal or rational values will be replaced by the name of the corresponding parameter.

All the substitution parameters must be decimal or rational to get plots.
All the substitution parameters and their sum must be ≥ 0 and < 1 .

Site 1	a[1]: <input style="width: 50px;" type="text" value="0.2"/>	b[1]: <input style="width: 50px;" type="text" value="0.1"/>	c[1]: <input style="width: 50px;" type="text" value="0.05"/>
Site 2	a[2]: <input style="width: 50px;" type="text" value="0.3"/>	b[2]: <input style="width: 50px;" type="text" value="0.15"/>	c[2]: <input style="width: 50px;" type="text" value="0.2"/>

Substitution parameters info:

Parameters sum = 1.

Fig. 6. Main interface of the Evolution Plots functionality for the model *TECT*: (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).

Fig. 7. Main interface of the Evolution Plots functionality for the model *TECt*: (4) choice of the genetic motifs; (5) selection of the output format; (6) time interval for plots (optional); (7) y-axis scale (optional); and (8) time value (optional).

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 I b_s ($A \leftrightarrow T$ and $C \leftrightarrow G$) and transversions II 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 $A^n = \underbrace{A \cdot \dots \cdot A}_n$ for the chosen size n .

4.4.2.5. Output format (models *TECt*, *Tt*, *ECt*). 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_{\min} and t_{\max} of the time interval $[t_{\min}, t_{\max}]$ can be chosen. They must always be positive in the direct and inverse time directions. By default, plots are drawn in the time interval $[t_{\min}, t_{\max}] = [0, 5]$.
- (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 *TECt* and *ECl* at sequence length l

4.4.3.1. Initial sequence length (models *TECt*, *ECl*). The analytical occurrence probabilities $P(l)$ (Eq. (3.3)) and $\mathbf{P}(l)$ (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 *TECt*). Similar to the models *TECt* and *Tt* (see Section 4.4.2.2).

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

4.4.3.4. Genetic motifs (models *TECt*, *ECl*). Similar to the models *TECt*, *Tt* and *ECt* (see Section 4.4.2.4).

Results

Analytical solutions (Standard format):

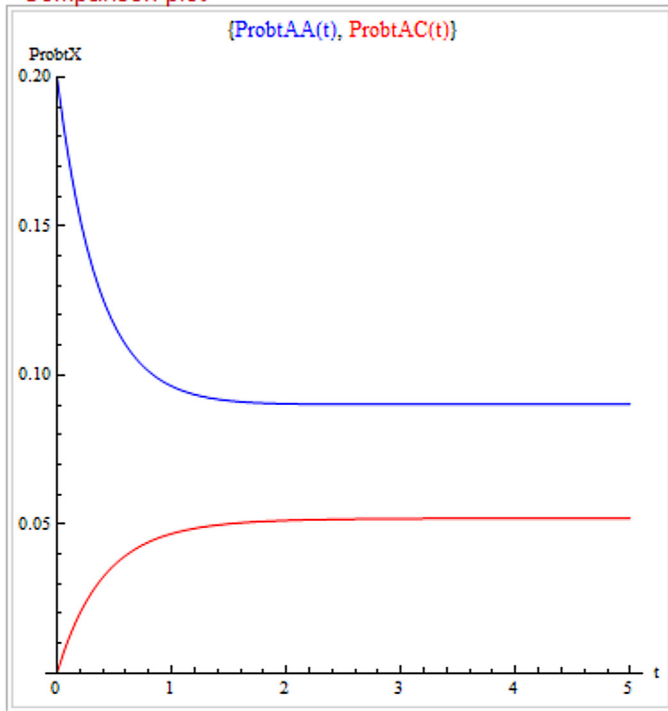
ProbAA(t)	<input type="button" value="SAVE"/>	$(1.44428 + 0.457025/E^{-3.06338t} - 0.255444/E^{-2.96338t} + 0.635276/E^{-2.86338t} - 0.466215/E^{-2.76338t} + 0.816513/E^{-2.66338t} + 0.75318/E^{-2.46338t} - 0.108345/E^{-2.36338t} - 0.109859/E^{-2.16338t} + 0.525542/E^{-2.06338t} - 0.0208617/E^{-1.96338t} - 0.471087/E^{-1.76338t})/16$
ProbAC(t)	<input type="button" value="SAVE"/>	$(0.830071 - 0.457025/E^{-3.06338t} - 0.292354/E^{-2.96338t} + 0.635276/E^{-2.86338t} + 0.466215/E^{-2.76338t} - 0.46411/E^{-2.66338t} - 0.75318/E^{-2.46338t} - 0.108345/E^{-2.36338t} + 0.109859/E^{-2.16338t} + 0.525542/E^{-2.06338t} - 0.0208617/E^{-1.96338t} - 0.471087/E^{-1.76338t})/16$

Fig. 8. Analytical solutions with the model *TECt* for the dinucleotides AA and AC in Standard output format.

Plots:

POA	1
POC	0
POG	0
POT	0

- Comparison plot



- Sum plot

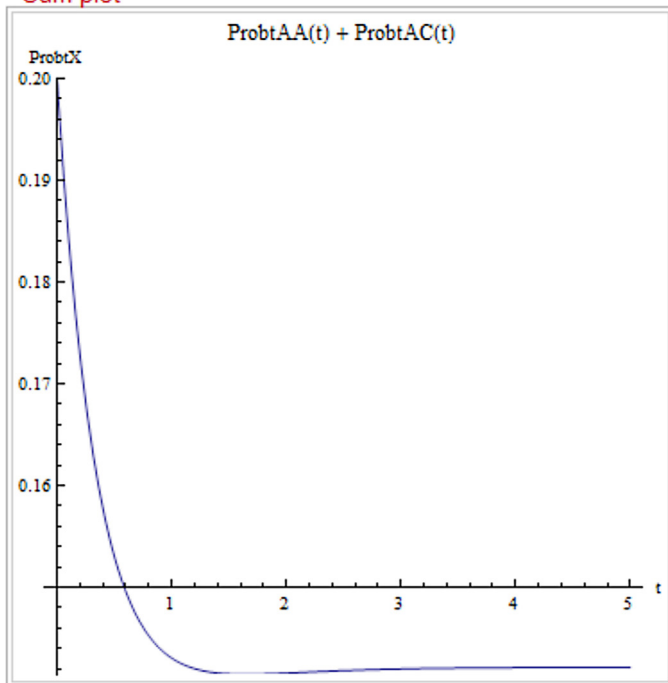


Fig. 9. Plots with the model *TECt* drawing the evolution curves for the dinucleotides AA and AC (up) and the evolution curve of their sum (bottom).

4.4.3.5. *Output format (models TECI, ECI).* Similar to the models *TECt*, *Tt* and *ECt* (see Section 4.4.2.5).

4.4.3.6. *Optional functionalities for plots and numerical solutions (models TECI, ECI).* 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,

Fig. 10. File Prob0TL1.txt of initial occurrence probabilities of nucleotides of the model *Tt* (Transformation) at time *t*.

i.e. containing up to four curves, and a plot drawing the evolution curve of their sum.

- Sequence length interval for plots: the parameters l_{\min} and l_{\max} of the sequence length interval $[l_{\min}, l_{\max}]$ can be chosen. They must be strictly positive integers. By default, plots are drawn in the sequence length interval $[l_{\min}, l_{\max}] = [1, 10]$.
- Scale of y-axis for plots: similar to the models *TECt*, *Tt* and *ECt* (see Section 4.4.2.6).
- 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 *Tt* (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 *Tt*

The following parameters must be selected:

- Choose the motif size: Nucleotides (1).
- Upload the initial occurrence probability file: the file Prob0TL1.txt (Fig. 10) must contain an initial nucleotide occurrence probability equal to 1, e.g. $\mathcal{P}_A(0) = 1$, and thus, the three other initial nucleotide occurrence probabilities are equal to 0, i.e. $\mathcal{P}_C(0) = \mathcal{P}_G(0) = \mathcal{P}_T(0) = 0$.

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

5.1.2. Second user interface of the model *Tt*

The following parameters must be selected:

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

Results

Analytical solutions (Standard format):

ProbtA(t)	SAVE	$(1 + E^{-4*b*t} + 2/E^{2*(a+b)*t})/4$
ProbtC(t)	SAVE	$(1 - E^{-4*b*t})/4$
ProbtG(t)	SAVE	$(1 + E^{-4*b*t} - 2/E^{2*(a+b)*t})/4$
ProbtT(t)	SAVE	$(1 - E^{-4*b*t})/4$

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]: 2*v” 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 $u_s = a_s$ ($A \leftrightarrow G$ 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 in order to express this model has a particular case of 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.

- Choice of the probabilities to study and plot: motif A motif C motif G motif T.
- 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.

Results

Analytical solutions (Standard format):

ProbtA(t)	SAVE	$(1 + 3/E^{4*a*t})/4$
ProbtC(t)	SAVE	$(1 - E^{-4*a*t})/4$
ProbtG(t)	SAVE	$(1 - E^{-4*a*t})/4$
ProbtT(t)	SAVE	$(1 - E^{-4*a*t})/4$

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.

- With the model Tt at three parameters:
 - Number of substitution parameters per motif site: 3 parameters.
 - Substitution parameters: a[1]: a, b[1]: a and v[1]: a.
- With the model Tt at two parameters:
 - Number of substitution parameters per motif site: 2 parameters.
 - Substitution parameters: u[1]: a and v[1]: 2*a.
- With the model Tt at one parameter:
 - Number of substitution parameters per motif site: 1 parameter.
 - Substitution parameters: p[1]: 3*a.

1 parameter ▾

3. Substitution parameters:
Enter values for the substitution parameters.
 Non decimal or rational values will be replaced by the name of the corresponding parameter.
 All the substitution parameters must be decimal or rational to get plots.
 All the substitution parameters and their sum must be ≥ 0 and ≤ 1 .

Site 1 p[1]:

Site 2 p[2]:

Site 3 p[3]:

Substitution parameters info:
 Parameters sum = 1

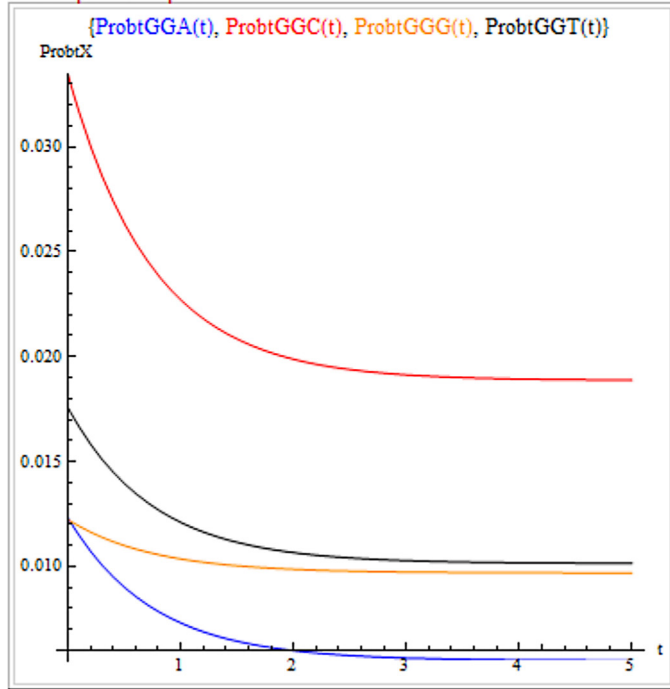
4. Choice of the probabilities to study and plot:
Choose up to 4 analytical solutions.
 By default, only the analytical solution of the motif AAA is displayed and plotted.

motif GGA ▾ motif GGC ▾ motif GGG ▾ motif GGT ▾

Fig. 13. Partial screenshot of the Evolution Plots functionality of the GETEC software for the models Tt and $TECt$ showing the substitution parameter settings cs_2 and cs_{i2} , 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.

Plots:

- Comparison plot



- Sum plot

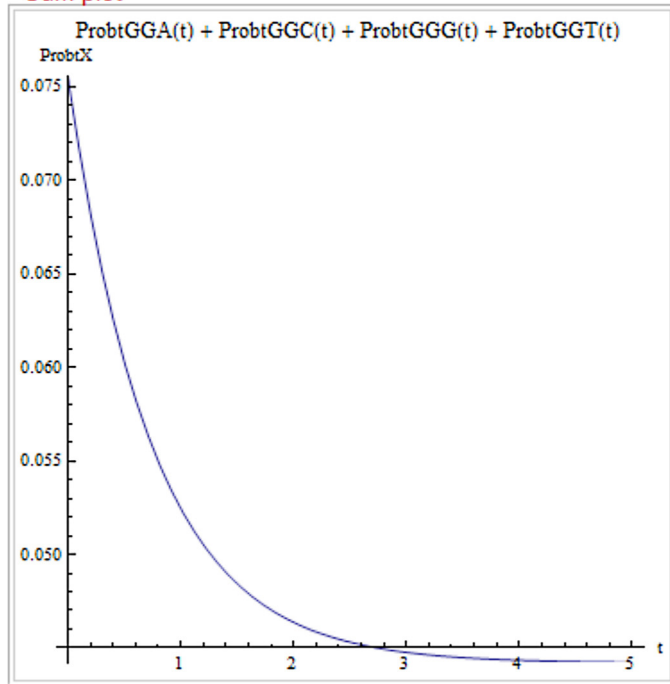
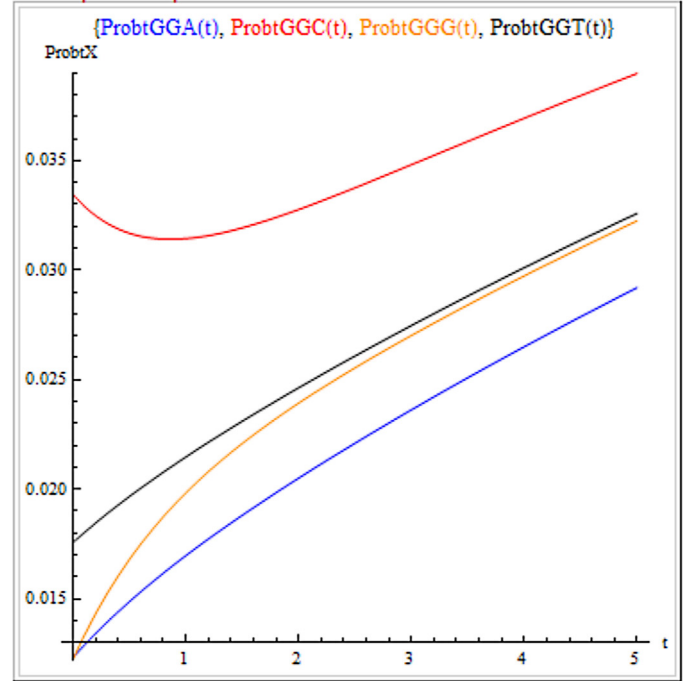


Fig. 14. Evolution curves in bacterial genes of the four codon occurrence probabilities $P_{GGA}(t)$, $P_{GGC}(t)$, $P_{GGG}(t)$ and $P_{GGT}(t)$ (top figure) and their probability sum $P_{Gly}(t)$ of glycine (bottom figure) in the time interval $[0,5]$ with the model Tt and the substitution configuration cs_1 (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).

Plots:

- Comparison plot



- Sum plot

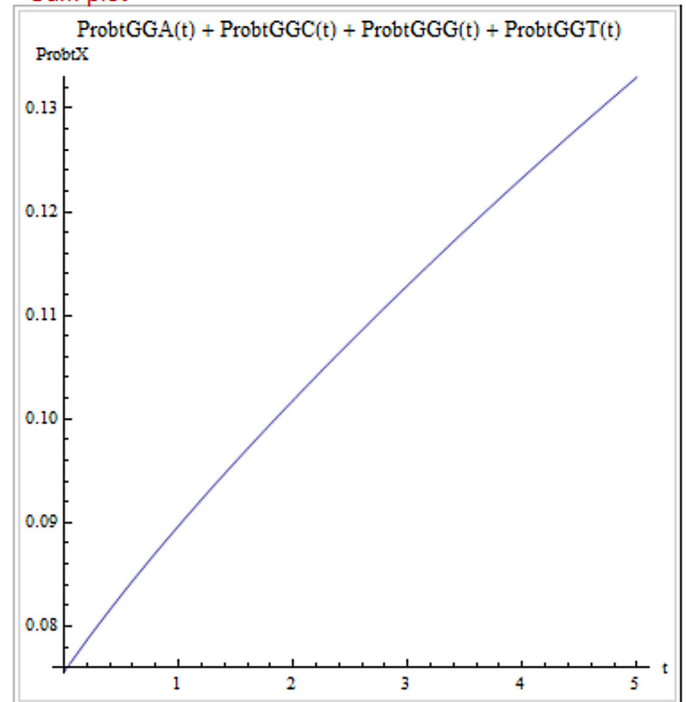


Fig. 15. Evolution curves in bacterial genes of the four codon occurrence probabilities $P_{GGA}(t)$, $P_{GGC}(t)$, $P_{GGG}(t)$ and $P_{GGT}(t)$ (top figure) and their probability sum $P_{Gly}(t)$ of glycine (bottom figure) in the time interval $[0,5]$ with the model TEt and the substitution–insertion 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)).

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

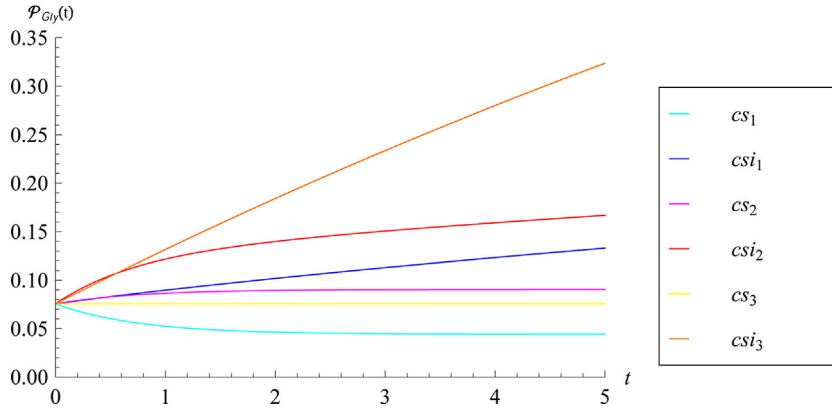
$\mathcal{P}_{Gly}(t)$ - Prokaryotes

Fig. 16. Evolution curves in bacterial genes of the occurrence probabilities $\mathcal{P}_{Gly}(t)$ and $P_{Gly}(t)$ of glycine in the time interval $[0,5]$ with the models Tt (substitution configurations cs_1 , cs_2 and cs_3) and $TEct$ (substitution-insertion configurations csi_1 , csi_2 and csi_3).

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 $\mathcal{P}_{Gly}(t)$ 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. $\mathcal{P}_{Gly}(t) = \mathcal{P}_{GGA}(t) + \mathcal{P}_{GGC}(t) + \mathcal{P}_{GGG}(t) + \mathcal{P}_{GGT}(t)$. 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 $\mathcal{P}(0) = P(0)$ 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 $\mathcal{P}_{Gly}(0) = P_{Gly}(0)$ of glycine at time 0 in both models Tt and $TEct$ is equal to $\mathcal{P}_{Gly}(0) = \mathcal{P}_{GGA}(0) + \mathcal{P}_{GGC}(0) + \mathcal{P}_{GGG}(0) + \mathcal{P}_{GGT}(0) = 0.0123 + 0.0335 + 0.0122 + 0.0176 = 0.0756$ 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

$$r_i = \begin{cases} 1/64 & \text{if } i \in \{GGA, GGC, GGG, GGT\} \\ 0 & \text{otherwise} \end{cases} \quad (6.1)$$

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 site 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_j and csi_j , $1 \leq j \leq 3$, respectively. Appendix D gives the numerical solutions of occurrence probabilities $\mathcal{P}_{Gly}(t)$ 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 $\mathcal{P}_{GGA}(t)$, $\mathcal{P}_{GGC}(t)$, $\mathcal{P}_{GGG}(t)$ and $\mathcal{P}_{GGT}(t)$, and their probability sum $\mathcal{P}_{Gly}(t)$ 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 $P_{GGA}(t)$, $P_{GGC}(t)$, $P_{GGG}(t)$ and $P_{GGT}(t)$, 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 $\mathcal{P}_{GGA}(t)$, $\mathcal{P}_{GGC}(t)$, $\mathcal{P}_{GGG}(t)$, $\mathcal{P}_{GGT}(t)$ and $\mathcal{P}_{Gly}(t)$ 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 $P_{GGA}(t)$, $P_{GGC}(t)$, $P_{GGG}(t)$, $P_{GGT}(t)$ and $P_{Gly}(t)$ increase, after a minimum for $P_{GGC}(t)$, with time t and tend to a higher limit (not shown in Fig. 15).

Fig. 16 summarizes evolution of the occurrence probabilities $\mathcal{P}_{Gly}(t)$ and $P_{Gly}(t)$ of glycine in bacterial genes with the substitution

and substitution–insertion models Tt and $TECt$, 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 $TECt$) than under “substitution only” (model Tt), i.e. $P_{Gly}(t) > \mathcal{P}_{Gly}(t)$ for each couple of configurations (cs_j, csi_j) , $1 \leq j \leq 3$ (Fig. 16), the difference resulting from the additional insertion process as shown previously for the couple (cs_1, csi_1) (Figs. 14 and 15).

6.3.2. Motif site comparison

Under “substitution only” (model Tt and cs), the occurrence probability $\mathcal{P}_{Gly}(t)$ 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 $TECt$ and csi), $P_{Gly}(t)$ in bacterial genes for time t in $[0, 0.6]$ has a similar probability behavior per site to “substitution only”, but for time t in $[0.6, 5]$, $P_{Gly}(t)$ 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 j into residue i 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 *IDISL–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 *TECt* (Transformation Expansion Contraction) at time t , *TECI* (Transformation Expansion Contraction) at sequence length l , *Tt* (Transformation) at time t , *ECt* (Expansion Contraction) at time t and *ECl* (Expansion Contraction) at sequence

length l . 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. 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

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

$$A_s = \begin{pmatrix} A_{s-1} & c_{n-s+1}I_{s-1} & a_{n-s+1}I_{s-1} & b_{n-s+1}I_{s-1} \\ c_{n-s+1}I_{s-1} & A_{s-1} & b_{n-s+1}I_{s-1} & a_{n-s+1}I_{s-1} \\ a_{n-s+1}I_{s-1} & b_{n-s+1}I_{s-1} & A_{s-1} & c_{n-s+1}I_{s-1} \\ b_{n-s+1}I_{s-1} & a_{n-s+1}I_{s-1} & c_{n-s+1}I_{s-1} & A_{s-1} \end{pmatrix} \\ = \begin{pmatrix} 0 & c_{n-s+1} & a_{n-s+1} & b_{n-s+1} \\ c_{n-s+1} & 0 & b_{n-s+1} & a_{n-s+1} \\ a_{n-s+1} & b_{n-s+1} & 0 & c_{n-s+1} \\ b_{n-s+1} & a_{n-s+1} & c_{n-s+1} & 0 \end{pmatrix} \otimes I_{s-1} \\ + \begin{pmatrix} A_{s-1} & & & \\ & A_{s-1} & & \\ & & A_{s-1} & \\ & & & A_{s-1} \end{pmatrix}$$

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

$$A_s = \begin{pmatrix} 0 & c_{n-s+1} & a_{n-s+1} & b_{n-s+1} \\ c_{n-s+1} & 0 & b_{n-s+1} & a_{n-s+1} \\ a_{n-s+1} & b_{n-s+1} & 0 & c_{n-s+1} \\ b_{n-s+1} & a_{n-s+1} & c_{n-s+1} & 0 \end{pmatrix} \otimes I_{s-1} + I_1 \otimes A_{s-1}$$

with I_1 the identity matrix of size (4,4). Therefore, by definition of the Kronecker sum,

$$A_s = \begin{pmatrix} 0 & c_{n-s+1} & a_{n-s+1} & b_{n-s+1} \\ c_{n-s+1} & 0 & b_{n-s+1} & a_{n-s+1} \\ a_{n-s+1} & b_{n-s+1} & 0 & c_{n-s+1} \\ b_{n-s+1} & a_{n-s+1} & c_{n-s+1} & 0 \end{pmatrix} \oplus A_{s-1}. \quad (\text{A.1})$$

Moreover,

$$N_{n-s+1} = \begin{pmatrix} 0 & c_{n-s+1} & a_{n-s+1} & b_{n-s+1} \\ c_{n-s+1} & 0 & b_{n-s+1} & a_{n-s+1} \\ a_{n-s+1} & b_{n-s+1} & 0 & c_{n-s+1} \\ b_{n-s+1} & a_{n-s+1} & c_{n-s+1} & 0 \end{pmatrix} + \begin{pmatrix} d_{n-s+1} & 0 & 0 & 0 \\ 0 & d_{n-s+1} & 0 & 0 \\ 0 & 0 & d_{n-s+1} & 0 \\ 0 & 0 & 0 & d_{n-s+1} \end{pmatrix} = \begin{pmatrix} 0 & c_{n-s+1} & a_{n-s+1} & b_{n-s+1} \\ c_{n-s+1} & 0 & b_{n-s+1} & a_{n-s+1} \\ a_{n-s+1} & b_{n-s+1} & 0 & c_{n-s+1} \\ b_{n-s+1} & a_{n-s+1} & c_{n-s+1} & 0 \end{pmatrix} + d_{n-s+1} \times I_1$$

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,

$$\begin{aligned} A_n &= \oplus_{s=1}^n (N_s - d_s \times I_1) + A_0 \times I_n \\ &= \oplus_{s=1}^n N_s - \oplus_{s=1}^n (d_s \times I_1) + A_0 \times I_n \\ &= \oplus_{s=1}^n N_s - \sum_{s=1}^n d_s \times I_n + A_0 \times I_n \\ &= \oplus_{s=1}^n N_s \end{aligned}$$

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 . □

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.

$$N_1 \oplus N_2 = \begin{pmatrix} d_1 & c_1 & a_1 & b_1 \\ c_1 & d_1 & b_1 & a_1 \\ a_1 & b_1 & d_1 & c_1 \\ b_1 & a_1 & c_1 & d_1 \end{pmatrix} \oplus \begin{pmatrix} d_2 & c_2 & a_2 & b_2 \\ c_2 & d_2 & b_2 & a_2 \\ a_2 & b_2 & d_2 & c_2 \\ b_2 & a_2 & c_2 & d_2 \end{pmatrix}$$

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

$$\begin{aligned} N_1 \oplus N_2 &= \begin{pmatrix} d_1 & c_1 & a_1 & b_1 \\ c_1 & d_1 & b_1 & a_1 \\ a_1 & b_1 & d_1 & c_1 \\ b_1 & a_1 & c_1 & d_1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\ &+ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} d_2 & c_2 & a_2 & b_2 \\ c_2 & d_2 & b_2 & a_2 \\ a_2 & b_2 & d_2 & c_2 \\ b_2 & a_2 & c_2 & d_2 \end{pmatrix} \\ &= \begin{pmatrix} d_1 I_1 & c_1 I_1 & a_1 I_1 & b_1 I_1 \\ c_1 I_1 & d_1 I_1 & b_1 I_1 & a_1 I_1 \\ a_1 I_1 & b_1 I_1 & d_1 I_1 & c_1 I_1 \\ b_1 I_1 & a_1 I_1 & c_1 I_1 & d_1 I_1 \end{pmatrix} + \begin{pmatrix} N_2 & 0 & 0 & 0 \\ 0 & N_2 & 0 & 0 \\ 0 & 0 & N_2 & 0 \\ 0 & 0 & 0 & N_2 \end{pmatrix} \end{aligned}$$

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

Then,

$$N_1 \oplus N_2 = \begin{pmatrix} d_1 I_1 + N_2 & c_1 I_1 & a_1 I_1 & b_1 I_1 \\ c_1 I_1 & d_1 I_1 + N_2 & b_1 I_1 & a_1 I_1 \\ a_1 I_1 & b_1 I_1 & d_1 I_1 + N_2 & c_1 I_1 \\ b_1 I_1 & a_1 I_1 & c_1 I_1 & d_1 I_1 + N_2 \end{pmatrix}$$

$$= \begin{pmatrix} d & c_2 & a_2 & b_2 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & 0 \\ c_2 & d & b_2 & a_2 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 \\ a_2 & b_2 & d & c_2 & 0 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 \\ b_2 & a_2 & c_2 & d & 0 & 0 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 \\ c_1 & 0 & 0 & 0 & d & c_2 & a_2 & b_2 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 \\ 0 & c_1 & 0 & 0 & c_2 & d & b_2 & a_2 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 \\ 0 & 0 & c_1 & 0 & a_2 & b_2 & d & c_2 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 \\ 0 & 0 & 0 & c_1 & b_2 & a_2 & c_2 & d & 0 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 \\ a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & 0 & d & c_2 & a_2 & b_2 & c_1 & 0 & 0 & 0 \\ 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & c_2 & d & b_2 & a_2 & 0 & c_1 & 0 & 0 \\ 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & a_2 & b_2 & d & c_2 & 0 & 0 & c_1 & 0 \\ 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & b_2 & a_2 & c_2 & d & 0 & 0 & 0 & c_1 \\ b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & 0 & 0 & d & c_2 & a_2 & b_2 \\ 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & 0 & c_2 & d & b_2 & a_2 \\ 0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & a_2 & b_2 & d & c_2 \\ 0 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & b_2 & a_2 & c_2 & d \end{pmatrix} = A_2.$$

with $d = -(a_1 + b_1 + c_1 + a_2 + b_2 + c_2)$.

Appendix B. Occurrence probability of a given genetic motif i with the models *TECI*, *Tt*, *ECT* and *ECl*

Model *TECI* (Transformation Expansion Contraction) at sequence length l . From Eq. (3.3), the occurrence probability $P_i(l)$ of a given genetic motif i at sequence length l is

$$\begin{aligned} P_i(l) &= \left(\sum_{k=1}^{4^n} \frac{1}{r+1-\lambda_k} O_k[i, l] \right) \cdot R \\ &+ \sum_{k=1}^{4^n} O_k[i, l] \cdot \left(P(l_0) - \frac{1}{r+1-\lambda_k} R \right) \left(\frac{l}{l_0} \right)^{-\frac{r+1-\lambda_k}{r-d}}. \end{aligned}$$

Model *Tt* (Transformation) at time *t*. From Eq. (3.4), the occurrence probability $P_i(t)$ of a given genetic motif *i* at time *t* is

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

Model *Ect* (Expansion Contraction) at time *t*. From Eq. (3.5), the occurrence probability $P_i(t)$ of a given genetic motif *i* at time *t* is

$$P_i(t) = \frac{R[i]}{r} + \left(P_i(0) - \frac{R[i]}{r} \right) e^{-rt}.$$

Model *Ecl* (Expansion Contraction) at sequence length *l*. From Eq. (3.6), the occurrence probability $P_i(l)$ of a given motif *i* at sequence length *l* is

$$P_i(l) = \frac{R[i]}{r} + \left(P_i(l_0) - \frac{R[i]}{r} \right) \left(\frac{l}{l_0} \right)^{-\frac{r}{r-d}}.$$

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 2a in Michel, 2015). It is used for defining the initial vectors $P(0) = P(0)$ of codon occurrence probabilities at time $t=0$ in the models *Tt* (Transformation; Eq. (3.4)) and *TEct* (Transformation Expansion Contraction; Eq. (3.2)).

Codon <i>i</i>	$P_i(0)$	Codon <i>i</i>	$P_i(0)$	Codon <i>i</i>	$P_i(0)$	Codon <i>i</i>	$P_i(0)$
AAA	2.87	CAA	1.61	GAA	3.47	TAA	0.00
AAC	1.79	CAC	1.05	GAC	2.63	TAC	1.32
AAG	1.97	CAG	2.18	GAG	2.62	TAG	0.00
AAT	1.93	CAT	1.06	GAT	2.80	TAT	1.62
ACA	1.00	CCA	0.77	GCA	1.69	TCA	0.77
ACC	2.12	CCC	1.08	GCC	3.54	TCC	0.99
ACG	1.39	CCG	1.88	GCG	3.06	TCG	1.10
ACT	0.90	CCT	0.81	GCT	1.60	TCT	0.86
AGA	0.54	CGA	0.42	GGA	1.23	TGA	0.01
AGC	1.39	CGC	2.25	GGC	3.35	TGC	0.56
AGG	0.32	CGG	1.08	GGG	1.22	TGG	1.25
AGT	0.81	CGT	1.10	GGT	1.76	TGT	0.38
ATA	0.89	CTA	0.56	GTA	1.08	TTA	1.64
ATC	2.71	CTC	1.73	GTC	2.04	TTC	1.94
ATG	2.34	CTG	3.66	GTG	2.58	TTG	1.44
ATT	2.43	CTT	1.28	GTT	1.52	TTT	2.01

Appendix D. Evolution of glycine in bacterial genes

See Table 2.

Table 2

Occurrence probabilities $P_{Gly}(t)$ and $P_{Gly}(t)$ of glycine in bacterial genes at time *t* in the models *Tt* (Transformation; Eq. (3.4)) and *TEct* (Transformation Expansion Contraction; Eq. (3.2)), respectively, with the six configurations cs_j and cs_j : substitution configurations 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 site 3 and equal to 0 for the codon sites 1 and 2), and substitution–insertion configurations 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).

Model	Configuration	Solution
<i>Tt</i>	cs_1	$P_{Gly}(t) = 0.0441859 + 0.0313821e^{-4t/3}$
<i>Tt</i>	cs_2	$P_{Gly}(t) = 0.0904434 - 0.0148754e^{-4t/3}$
<i>Tt</i>	cs_3	$P_{Gly}(t) = 0.075568 + 1.73472 \times 10^{-18}e^{-4t/3}$
<i>TEct</i>	csi_1	$P_{Gly}(t) = 0.283582 - 0.00219995e^{-67t/48} - 0.205814e^{-t/16}$
<i>TEct</i>	csi_2	$P_{Gly}(t) = 0.283582 - 0.0484575e^{-67t/48} - 0.159557e^{-t/16}$
<i>TEct</i>	csi_3	$P_{Gly}(t) = 1 + 1.73472 \times 10^{-18}e^{-67t/48} - 0.924432e^{-t/16}$

References

Aldous, D., Fill, J.A., 2002. Reversible Markov Chains and Random Walks on Graphs. University of California, Berkeley.

Anisimova, M., Kosiol, C., 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol. Biol. Evol.* 26, 255–271.

Arquès, D.G., Fallot, J.-P., Michel, C.J., 1998. An evolutionary analytical model of a complementary circular code simulating the protein coding genes, the 5' and 3' regions. *Bull. Math. Biol.* 60, 163–194.

Arquès, D.G., Fallot, J.-P., Marsan, L., Michel, C.J., 1999. An evolutionary analytical model of a complementary circular code. *Biosystems* 49, 83–103.

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.

Bahi, J.M., Michel, C.J., 2004. A stochastic gene evolution model with time dependent mutations. *Bull. Math. Biol.* 66, 763–778.

Bahi, J.M., Michel, C.J., 2008. A stochastic model of gene evolution with chaotic mutations. *J. Theor. Biol.* 255, 53–63.

Bahi, J.M., Michel, C.J., 2009. A stochastic model of gene evolution with time dependent pseudo-chaotic mutations. *Bull. Math. Biol.* 71, 681–700.

Bailey, T.L., Bodén, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. *Nucl. Acids Res.* 37, W202–W208.

Benard, E., Michel, C.J., 2009. Computation of direct and inverse mutations with the SEGM web server (Stochastic Evolution of Genetic Motifs): an application to splice sites of human genome introns. *Comput. Biol. Chem.* 33, 245–252.

Benard, E., Michel, C.J., 2011. A generalization of substitution evolution models of nucleotides to genetic motifs. *J. Theor. Biol.* 288, 73–83.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.

Felsenstein, J., Churchill, G.A., 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* 13, 93–104.

Frey, G., Michel, C.J., 2006. An analytical model of gene evolution with 6 mutation parameters: an application to archaeal circular codes. *Comput. Biol. Chem.* 30, 1–11.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160–174.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, N. (Ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–132.

Kelly, F.P., 1979. *Reversibility and Stochastic Networks*. Wiley, Chichester.

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

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

Laub, A.J., 2005. *Matrix Analysis for Scientists and Engineers*. SIAM Publications, Philadelphia, PA.

Lèbre, S., Michel, C.J., 2010. A stochastic evolution model for residue insertion–deletion independent from substitution. *Comput. Biol. Chem.* 34, 259–267.

Lèbre, S., Michel, C.J., 2012. An evolution model for sequence length based on residue insertion–deletion independent of substitution: an application to the GC content in bacterial genomes. *Bull. Math. Biol.* 74, 1764–1788.

Lèbre, S., Michel, C.J., 2013. A new molecular evolution model for limited insertion independent of substitution. *Math. Biosci.* 245, 137–147.

Malthus, T.R., 1798. *An Essay on the Principle of Population*. Penguin, Harmondsworth, England.

McGuire, G., Denham, M.C., Balding, D.J., 2001. Models of sequence evolution for DNA sequences containing gaps. *Mol. Biol. Evol.* 18, 481–490.

Metzler, D., 2003. Statistical alignment based on fragment insertion and deletion models. *Bioinformatics* 19, 490–499.

Michel, C.J., 2007a. An analytical model of gene evolution with 9 mutation parameters: an application to the amino acids coded by the common circular code. *Bull. Math. Biol.* 69, 677–698.

Michel, C.J., 2007b. Codon phylogenetic distance. *Comput. Biol. Chem.* 31, 36–43.

Michel, C.J., 2007c. Evolution probabilities and phylogenetic distance of dinucleotides. *J. Theor. Biol.* 249, 271–277.

Michel, C.J., 2014. A genetic scale of reading frame coding. *J. Theor. Biol.* 355, 83–94.

Michel, C.J., 2015. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. *J. Theor. Biol.* 380, 156–177.

Miklós, I., Lunter, G.A., Holmes, I., 2004. A “long indel” model for evolutionary sequence alignment. *Mol. Biol. Evol.* 21, 529–540.

Miklós, I., Novák, A., Satija, R., Lyngsø, R., Hein, J., 2009. Stochastic models of sequence evolution including insertion–deletion events. *Stat. Methods Med. Res.* 18, 453–485.

Rivas, E., 2005. Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC Bioinform.* 6, 63.

Rivas, E., Eddy, S.R., 2008. Probabilistic phylogenetic inference with insertions and deletions. *PLoS Comput. Biol.* 4 (9), e1000172.

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.

- Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526.
- Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* 17, 57–86.
- Thorne, J.L., Kishino, H., Felsenstein, J., 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* 33, 114–124.
- Thorne, J.L., Kishino, H., Felsenstein, J., 1992. Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.* 34, 3–16.
- Tian, Y., Styan, G.P.H., 2001. How to establish universal block-matrix factorizations. *Electron. J. Linear Algebra* 8, 115–127.
- Yang, Z., 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39, 105–111.
- Yang, Z., 2006. Computational molecular evolution. In: Harvey, P.H., May, R.M. (Eds.), *Oxford Series in Ecology and Evolution*. Oxford University Press.