Computational Biology and Chemistry 34 (2010) 259–267

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

A stochastic evolution model for residue Insertion–Deletion Independent from Substitution

Sophie Lèbre, Christian J. Michel*

Equipe de Bioinformatique Théorique, FDBT, LSIIT (UMR UdS-CNRS 7005), Université de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France

ARTICLE INFO

Article history:
Received 31 August 2010
Received in revised form 2 September 2010
Accepted 2 September 2010

Keywords:
Gene evolution
Stochastic model
Analytical solutions
Substitution
Insertion
Deletion
Time and sequence length
Occurrence probability
Nucleotides

ABSTRACT

We develop here a new class of stochastic models of gene evolution based on residue Insertion–Deletion Independent from Substitution (*IDIS*). Indeed, in contrast to all existing evolution models, insertions and deletions are modeled here by a concept in population dynamics. Therefore, they are not only independent from each other, but also independent from the substitution process.

After a separate stochastic analysis of the substitution and the insertion–deletion processes, we obtain a matrix differential equation combining these two processes defining the *IDIS* model. By deriving a general solution, we give an analytical expression of the residue occurrence probability at evolution time t as a function of a substitution rate matrix, an insertion rate vector, a deletion rate and an initial residue probability vector. Various mathematical properties of the *IDIS* model in relation with time t are derived: time scale, time step, time inversion and sequence length. Particular expressions of the nucleotide occurrence probability at time t are given for classical substitution rate matrices in various biological contexts: equal insertion rate, insertion–deletion only and substitution only. All these expressions can be directly used for biological evolutionary applications.

The *IDIS* model shows a strongly different stochastic behavior from the classical substitution only model when compared on a gene dataset. Indeed, by considering three processes of residue insertion, deletion and substitution independently from each other, it allows a more realistic representation of gene evolution and opens new directions and applications in this research field.

1. Introduction

Gene evolution models were initially developed to study the substitution rates of nucleotides (adenine *A*, cytosine *C*, guanine *G*, thymine *T*). The first gene evolution model was proposed by Jukes and Cantor (1969) with 1-parameter substitution (probability α for all nucleotide substitution types). It was generalized to a 2-parameter substitution model (Kimura, 1980) (probability γ for the nucleotide transitions $A \leftrightarrow G$ and $C \leftrightarrow T$, and probability β for the nucleotide transversions $A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$ and $G \leftrightarrow T$) and then, to a 3-parameter substitution model (Kimura, 1981) (probability a for transitions, probability b for transversion type $A \leftrightarrow T$ and $C \leftrightarrow G$, and probability c for transversion type $A \leftrightarrow C$ and $G \leftrightarrow T$). Later, these substitution models were generalized up to nine free parameters, in particular (Felsenstein, 1981; Takahata and Kimura, 1981; Hasegawa et al., 1985; Tavaré, 1986; Tamura and Nei, 1993; Yang, 1994; Felsenstein and Churchill, 1996).

Over the last 20 years, only very few evolution models were extended to the insertions and the deletions of residues (nucleotides, amino acids, etc.) in addition to residue substitutions. Such models are the topic of current research, in particular in the context of probabilistic methods for alignment. The first approach, proposed by Thorne et al. (1991) and commonly called TKF91, models insertion and deletion as a continuous time birth–death process governed by explicit parameters for the insertion and deletion rates. Some extensions of the TKF91 model were proposed for the insertion of fragments of several residues (long indels) (Thorne et al., 1992; Metzler, 2003; Miklós et al., 2004; see e.g. Miklós et al., 2009 for a review).

Another class of evolution models with insertion and deletion was introduced by McGuire et al. (2001) as an extension to the nucleotide substitution model F84 introduced by Felsenstein and Churchill (1996). A fifth residue referring to the gap character is added to the four nucleotides and is incorporated in a Markov model of nucleotide substitution. This model is based on an extended substitution matrix for the extended alphabet comprising the four nucleotides and the gap character, i.e. a substitution matrix with one additional line and one additional column. Thus, an insertion corresponds to the substitution of a gap by a nucleotide whereas a deletion amounts to the substitution of a nucleotide by

* Corresponding author.
E-mail addresses: lebre@dpt-info.u-strasbg.fr (S. Lèbre), michel@dpt-info.u-strasbg.fr (C.J. Michel).

1476-9271/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2010.09.001

a gap. The nucleotide insertion probability is proportional to the nucleotide equilibrium distribution whereas the deletion probability of any nucleotide is associated with an extra parameter for the constant gap frequency.

All the insertion–deletion models mentioned above are reversible, a useful property for inferring unrooted phylogenetic trees. This reversibility property is also classical with some substitution models. However, for a reversible insertion–deletion model, some theoretical constraints must be imposed on the insertion and deletion rates which prevents the insertion and deletion processes to be independent from each other. For example in a pairwise alignment, the reversibility constraint imposes that the expected frequencies of insertions and deletions must be identical.

Rivas (2005) and Rivas and Eddy (2008) later generalized the model of McGuire et al. (2001) to a non-reversible model by adding explicit parameters for the insertion and deletion rates. In the particular case of a reversible substitution matrix and insertion rates proportional to the nucleotide equilibrium distribution associated with the substitution matrix, an analytical expression of the substitution probability over some time t is obtained (Equation (9) in Rivas and Eddy, 2008).

We develop here a new class of non-reversible evolution models for residue Insertion–Deletion Independent from Substitution which we call *IDIS*. Based on a continuous Markov process governed by an instantaneous substitution rate matrix, our *IDIS* model considers that the insertion and deletion processes are independent from each other, as in Rivas (2005) and Rivas and Eddy (2008). However, in contrast to the previous approaches by McGuire et al. (2001), Rivas (2005), and Rivas and Eddy (2008), it is not based on the introduction of a gap character for extending the substitution rate matrix. Indeed, the modeling of the insertion and deletion processes was inspired by a concept in population dynamics (Malthus, 2000). Thus, the insertion and deletion processes are not only independent from each other, but they are also independent from the substitution process. Therefore, the *IDIS* model relies on a real physical process of evolution which is based on substitutions, insertions and deletions in residue sequences (simulated sequence evolution, see Remark 6 for details). After a separate analysis of the substitution and the insertion–deletion processes, we define the Insertion–Deletion Independent from Substitution (*IDIS*) model via a matrix differential equation satisfied by the residue occurrence probability at evolution time t. By deriving the solution of this differential equation, we obtain an analytical expression for residue occurrence probability at time t as a function of the initial residue occurrence probability, the insertion and deletion rates, and the eigenvalues and eigenvectors of the instantaneous substitution rate matrix (Eq. (2.13)).

To our knowledge, our approach opens a new theoretical field in gene evolution, mainly by the fact that the three processes of residue insertion, deletion and substitution are independent from each other, in contrast to all previous evolution models. Furthermore, applications of the *IDIS* model can be various: sequence alignment and phylogeny, but also in the line of our previous evolution models during the last 20 years: models of 'primitive' genes, of 'primitive' genetic and amino acids motifs, study of substitution rates and analysis of residue occurrence probabilities in the direct evolution time direction (past-present) or in the inverse one (present-past), e.g. Arquès and Michel (1990, 1993, 1995), Michel (2007), and Benard and Michel (2009). Indeed, contrary to the classical approaches focusing on substitution probability matrix and sequence alignment, the *IDIS* model allows us to analyse the behavior of the residue occurrence probability along time (in both directions, Section 3). In particular, nucleotide probability curves with local/global maxima or minima, increasing or decreasing curves, crossing curves and asymptotic behavior can be observed and studied in genes.

This paper is organized as follows. In Section 2, we define the *IDIS* model for sequence evolution and derive the residue occurrence probability at evolution time t under the substitution and independent insertion–deletion processes. In Section 3, various mathematical properties of the *IDIS* model are given in relation with time t: time scale, time step, time inversion and sequence length. In Section 4, we derive analytical occurrence probabilities of nucleotides for the *IDIS-sym*3 model with the 3-parameter substitution model (Kimura, 1981) and for particular cases: equal nucleotide insertion rates, insertion–deletion only, substitution only, *IDIS* model with 2-parameter (Kimura, 1980) or 1-parameter (Jukes and Cantor, 1969) substitution rate matrix. Finally, Section 5 shows a comparison of the *IDIS* model with a substitution only model on a gene dataset.

2. Mathematical model

We introduce here the *IDIS* model, a time-continuous stochastic evolution model for residue Insertion and Deletion Independent from Substitution. It allows the substitution, the insertion and the deletion of residues in a biological sequence. In contrast to the classical substitution–insertion–deletion models, the insertion and deletion rates of each residue are explicit parameters of the model, i.e. independent from 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 $1 \le i \le K$, we denote by $P_i(t)$, the occurrence probability of residue i at time $t \ge 0$ per 'residue site' in the sequence. The column vector $P(t) = [P_i(t)]_{1 \le i \le K}$ of size K is made of the probabilities $P_i(t)$ for all $1 \le i \le K$. Before deriving the general stochastic *IDIS* model, we analyse the substitution and insertion–deletion processes separately. We first build a specific differential equation for each evolution process.

2.1. Stochastic substitution model

The substitution process is handled by a differential equation which determines the occurrence probability $P(t)$ at time $t \ge 0$ of the K residues mutating according to constant substitution probabilities. Let us consider two residues $1 \le i, j \le K$. We denote by $P_{t,t+T}(j \to i)$, the substitution probability of residue j into residue i between time t and $t+T$, $T>0$, which can be the result of several consecutive substitutions per residue site. The difference $P_i(t+T) - P_i(t)$ of occurrence probability of residue i at time t and $t+T$ is equal to the probability of residue i to appear by substitution ($j \to i$, $\forall j \neq i$) minus the probability of residue i to disappear by substitution ($i \to j$, $\forall j \neq i$) over the time interval $[t, t+T[$, i.e.

$$P_i(t+T) - P_i(t) = \underbrace{\sum_{j \neq i} P_j(t) P_{t,t+T}(j \to i)}_{\text{Probability of residue } i \text{ to appear}} - \underbrace{P_i(t) \sum_{j \neq i} P_{t,t+T}(i \to j)}_{\text{Probability of residue } i \text{ to disappear}} . \quad (2.1)$$

Remark 1. $\sum_{j \neq i} P_{t,t+T}(i \to j) = 1 - P_{t,t+T}(i \to i)$ where $P_{t,t+T}(i \to i)$ represents the probability that residue i does not mutate into a different residue $j \neq i$ between time t and $t+T$.

From Eq. (2.1), the derivative with respect to time $P_i'(t) = \partial P_i(t)/\partial t$ of the occurrence probability of residue i at time t is

$$\begin{aligned} P_i'(t) &= \lim_{T\to 0}\left(\frac{P_i(t+T)-P_i(t)}{T}\right) \\ &= \lim_{T\to 0}\left(\frac{\sum_{j\neq i} P_j(t)P_{t,t+T}(j\to i) - P_i(t)\sum_{j\neq i} P_{t,t+T}(i\to j)}{T}\right). \end{aligned}$$

As the limit of a sum of finite functions is the sum of the function limits, the derivative $P_i'(t)$ is

$$P_i'(t) = \sum_{j\neq i} P_j(t)\lim_{T\to 0}\left(\frac{P_{t,t+T}(j\to i)}{T}\right) - P_i(t)\sum_{j\neq i}\lim_{T\to 0}\left(\frac{P_{t,t+T}(i\to j)}{T}\right).$$

For all residues i, j, the instantaneous substitution probability $P(j\to i)$ of residue j into residue i is assumed to be constant along time. When T is small enough, there is not more than one residue substitution per residue site (the substitution of residue j into residue i cannot be the result of several consecutive substitutions for a given residue site). Then, the following approximation applies

$$P_{t,t+T}(j\to i) \underset{T\to 0}{=} P(j\to i)T.$$

and consequently

$$\lim_{T\to 0}\left(\frac{P_{t,t+T}(j\to i)}{T}\right) = P(j\to i).$$

Finally, for any residue i, the derivative $P_i'(t)$ is

$$\begin{aligned} P_i'(t) &= \sum_{1\le j\le K,\, j\neq i} P_j(t)P(j\to i) - P_i(t)\sum_{1\le j\le K,\, j\neq i} P(i\to j) \\ &= \sum_{1\le j\le K,\, j\neq i} P_j(t)P(j\to i) - P_i(t)(1-P(i\to i)) \\ &= \sum_{1\le j\le K} P_j(t)P(j\to i) - P_i(t). \end{aligned} \tag{2.2}$$

From Eq. (2.2), we derive a matrix differential equation which describes the substitution process

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

where the symbol $\cdot$ is the matrix product, matrix I is the identity matrix of size K and matrix $M = [m_{ij}]_{1\le i,j\le K}$ is the instantaneous substitution probability matrix whose element m_{ij} in row i and column j refers to the substitution probability of residue j into residue i

$$m_{ij} = P(j\to i). \tag{2.4}$$

The instantaneous substitution probability matrix M is stochastic in column. Indeed, for all $1\le j\le K$, the elements of matrix M satisfy $\sum_{1\le i\le K} m_{ij} = \sum_{1\le i\le K} P(j\to i) = 1$. Eq. (2.3) is equal to Equation 2 in Michel (2007) obtained by a similar approach. In Section 4, analytical solutions of the *IDIS* model are derived for various nucleotide substitution matrices stochastic in column.

Remark 2. The instantaneous substitution probability matrix M (2.4) is the transpose matrix of the classical substitution matrix $\pi = [P(i\to j)]_{1\le i,j\le K}$ which is stochastic in line (Kimura, 1981; Rivas, 2005) ($\pi_{ij} = P(i\to j) = m_{ji}$). When π is symmetric, $M = \pi$.

Remark 3. The general solution of Eq. (2.3) describing the substitution process is $P(t) = P(0)e^{t(M-I)}$. It is equivalent to the classical substitution probability matrix $M(t)$ over time t with $M(t) = e^{tQ}$ and $Q = M - I$ (Yang, 2006) (including possible intermediate successive substitutions). The ith row of matrix $M(t)$, denoted by $M[i,\,](t)$, describes the substitution probability of residue i. Alternatively, $M[i,\,](t)$ can be obtained from the *IDIS* model by setting the vector of initial probability for letter i to 1 ($P_i(0) = 1$), i.e. $P(0) = (\delta_{ij})_{1\le j\le K}$, leading to $M[i,\,](t) = (\delta_{ij})_{1\le j\le K}e^{t(M-I)}$.

2.2. Stochastic insertion–deletion model

We derive a differential equation modeling the insertion–deletion process, the substitution process being not considered here. For any residue i, the occurrence number of residue i in the biological sequence at time t is denoted by $n_i(t)$. The total number of residues at time t is denoted by $n(t) = \sum_{1\le i\le K} n_i(t)$. Let r_i be the insertion rate per site of each residue i, $\forall 1\le i\le K$, $r_i \ge 0$. In the *IDIS* model, the insertion rates are explicit parameters which are entirely independent from the substitution parameters. Let d be the deletion rate for all residues, $d \ge 0$. For any residue i, we assume that the growth rate $n_i'(t) = \partial n_i(t)/\partial t$ of residue i at time t due to insertions is equal to $r_i \times n(t)$, as in population dynamics (Malthus, 2000). The growth rate $n_i'(t)$ of residue i at time t due to deletions is $d \times n_i(t)$ where $n_i(t)$ is the number of occurrences of residue i in the sequence. Then, the growth rate $n_i'(t)$ resulting from the insertion–deletion process is for all $1\le i\le K$,

$$n_i'(t) = r_i \times n(t) - d \times n_i(t). \tag{2.5}$$

Remark 4. As in all the previous insertion–deletion models (Thorne et al., 1991, 1992; Metzler, 2003; Miklós et al., 2004; McGuire et al., 2001; Rivas, 2005; Rivas and Eddy, 2008), 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 analytical expression is not ensured with specific deletion rate d_i for each residue i.

Basing on the insertion–deletion growth rate $n_i'(t)$ (Eq. (2.5)), the derivative $P_i'(t)$ of the occurrence probability of residue i at time t writes

$$\begin{aligned} P_i'(t) &= \left(\frac{n_i(t)}{n(t)}\right)' = \frac{1}{n^2(t)}\left[n(t)[r_i n(t) - d n_i(t)] - n_i(t)\sum_{1\le j\le K} n_j'(t)\right] \\ &= \frac{1}{n^2(t)}\left[n(t)[r_i n(t) - d n_i(t)] - n_i(t)\sum_{1\le j\le K}[r_j n(t) - d n_j(t)]\right] \\ &= \frac{1}{n^2(t)}\left[n(t)[r_i n(t) - d n_i(t)] - n_i(t)\right. \\ &\quad \left.\times\left[n(t)\sum_{1\le j\le K} r_j - d\sum_{1\le j\le K} n_j(t)\right]\right] \\ &= \frac{1}{n^2(t)}\left[r_i n^2(t) - d n_i(t)n(t) - n_i(t)n(t)\sum_{1\le j\le K} r_j + d n_i(t)n(t)\right] \\ &= r_i - \left(\sum_{1\le j\le K} r_j\right) P_i(t). \end{aligned}$$

Finally,

$$P'(t) = -rP(t) + R, \tag{2.6}$$

where $r = \sum_{1 \le i \le K} r_i$ is the sum of all residue insertion rates, $\forall 1 \le i \le K$, $r_i \ge 0$, and $R = [r_i]_{1 \le i \le K}$ is the vector of residue insertion rates.

Remark 5. Eq. (2.6) does not depend on the deletion rate d. As expected, a deletion rate identical for each residue i does not affect the occurrence probability $P_i(t)$. Thus, the residue distribution is not affected by the deletions, in contrast to the most general and recent non-reversible insertion–deletion model (Rivas and Eddy, 2008) (detailed in Remark 6). Obviously, the sequence length depends on the deletion rate (Eq. (3.3)).

2.3. IDIS model: residue Insertion–Deletion Independent from Substitution

From the previous mathematical results, we derive a general matrix differential equation allowing for the substitution process and the insertion–deletion process to be superimposed. The *IDIS* model allows both substitution and insertion–deletion of residues. These processes are assumed to be independent, i.e. a substitution event does not alter the probability of an insertion–deletion event and reciprocally. Then, the derivative $P'(t)$ of the residue occurrence probability at time t is the result of the instantaneous variation due to the substitution (2.3) and the insertion–deletion (2.6). Thus, the residue occurrence probability vector $P(t) = [P_i(t)]_{1 \le i \le K}$ satisfies

$$\begin{aligned} P'(t) &= \underbrace{(M - I) \cdot P(t)}_{\text{Substitution}} + \underbrace{(-rP(t) + R)}_{\text{Insertion–Deletion}} \\ &= [M - (1 + r)I] \cdot P(t) + R, \end{aligned} \tag{2.7}$$

where M is the substitution probability matrix defined in (2.4), $R = [r_i]_{1 \le i \le K}$ is the vector of the residue insertion rates per site and $r = \sum_{1 \le i \le K} r_i$ is the sum of the residue insertion rates, $\forall 1 \le i \le K$, $r_i \ge 0$. Eq. (2.7) leads to the following nonhomogeneous matrix linear differential equation

$$P'(t) = A \cdot P(t) + R, \tag{2.8}$$

where $A = M - (1 + r)I$.

The general solution for the nonhomogeneous matrix differential Eq. (2.8) defining the *IDIS* model is obtained from the method of variation parameters (Hubbard and West, 1995). For all $s, t \ge 0$,

$$P(t) = e^{A(t-s)} P(s) + e^{At} \left(\int_s^t e^{-Au} du \right) R, \tag{2.9}$$

where $A = M - (1 + r)I$, $r = \sum_{1 \le i \le K} r_i$.

For $s = 0$, the residue occurrence probability $P(t)$ is defined as a function of initial probabilities $P(0)$. For all $t \ge 0$,

$$P(t) = e^{At} P(0) + e^{At} \left(\int_0^t e^{-Au} du \right) R. \tag{2.10}$$

When the substitution probability matrix M can be diagonalized with real eigenvalues $(\lambda_k)_{1 \le k \le K}$, then $\forall 1 \le k \le K - 1$, $\lambda_k \le 1$ and $\lambda_K = 1$ as Perron–Frobenius theorem ensures that the largest eigenvalue associated with a stochastic matrix, like M, is always 1. Let $D = Diag((\lambda_k)_{1 \le k \le K})$ be the eigenvalues diagonal matrix and Q be an associated eigenvectors matrix, the k th column of Q being an eigenvector for eigenvalue λ_k. Then, matrix M decomposes as $M = Q \cdot D \cdot Q^{-1}$. Using $A = Q \cdot D \cdot Q^{-1} - (1 + r)I = Q \cdot \tilde{D} \cdot Q^{-1}$ where matrix $\tilde{D} = D - (1 + r)I = Diag((\mu_k)_{1 \le k \le K})$, matrix A can be diagonalized with real eigenvalues $(\mu_k)_{1 \le k \le K}$ where $\forall 1 \le k \le K - 1$, $\mu_k = \lambda_k - (1 + r)$ and $\mu_K = -r$. Then, $e^{At} = Qe^{\tilde{D}t} Q^{-1}$ (Lange, 2005), and for all $t \ge 0$,

$$\begin{aligned} P(t) &= Qe^{\tilde{D}t} Q^{-1} P(0) + Qe^{\tilde{D}t} Q^{-1} \left(\int_0^t Qe^{-\tilde{D}u} Q^{-1} du \right) R \\ &= QD_1 Q^{-1} P(0) + QD_1 Diag \left(\left(\int_0^t e^{-\mu_k u} du \right)_{1 \le k \le K} \right) Q^{-1} R \\ &= QD_1 Q^{-1} P(0) + QD_2 Q^{-1} R, \end{aligned} \tag{2.11}$$

where $D_1 = Diag((e^{\mu_k t})_{1 \le k \le K})$ and $D_2 = Diag(((1/\mu_k)(e^{\mu_k t} - 1))_{1 \le k \le K})$.

For all $1 \le k \le K$, we build a matrix O_k of size $K \times K$ such that

$$O_k[i, j] = Q(i, k) Q^{-1}(k, j). \tag{2.12}$$

After some algebraic manipulation of Eq. (2.11), we obtain an expression of the residue occurrence probability $P(t)$ as a function of the insertion rates $R = [r_i]_{1 \le i \le K}$, the eigenvalues $(\lambda_k)_{1 \le k \le K}$ of the substitution probability matrix M and the matrices $(O_k)_{1 \le k \le K}$ defined in Eq. (2.12) using eigenvectors matrix Q of M

$$\begin{aligned} P(t) = &\left(\sum_{k=1}^{K} \frac{1}{r + 1 - \lambda_k} O_k \right) \cdot R \\ &+ \sum_{k=1}^{K} O_k \left(P(0) - \frac{1}{r + 1 - \lambda_k} R \right) e^{-(r+1-\lambda_k)t}. \end{aligned} \tag{2.13}$$

As the total insertion rate r is positive and the eigenvalues $(\lambda_k)_{1 \le k \le K}$ are smaller than 1, then $\forall 1 \le k \le K$, $-(r + 1 - \lambda_k) \le 0$ and the exponential terms are bounded: $\forall t \ge 0$, $\forall 1 \le k \le K$, $0 \le e^{-(r+1-\lambda_k)t} \le 1$. In Section 4, Eq. (2.13) will be used to derive nucleotide analytical probabilities for various substitution rate matrices.

Remark 6. As already mentioned when modeling the insertion–deletion process (Eq. (2.6) and Remark 5), the residue occurrence probability $P(t)$ is not function of the deletion rate d when d is identical for all residues. This property is in agreement with the physical model of gene evolution. Indeed, the probability $P(t)$ of the *IDIS* model given by Eq. (2.13) can be retrieved by computer simulation (by generating simulated sequences and applying evolution by substitutions, insertions and deletions). It is also a major difference with the most general and recent non-reversible insertion–deletion model (Rivas and Eddy, 2008). Indeed, the residue occurrence probability $P(t)$ which can be derived from the transition probability from residue i into residue j over time t using Equation (9) in Rivas and Eddy (2008), is a function of the deletion rate (called μ in their paper) and in contradiction with the physical model of gene evolution.

3. IDIS model properties

We set here four mathematical properties which relate the evolution time t to the values of the mutation parameters (M, R). These properties are important to model sequence evolution in practice.

3.1. Time scale

By multiplying all the substitution–insertion–deletion parameters, i.e. the non-diagonal elements $[m_{ij}]_{1 \le i,j \le K, i \ne j}$ of the substitution probability matrix M and the insertion rates $[r_i]_{1 \le i \le K}$, by a scalar α, then the occurrence probability $P(t/\alpha; [\alpha m_{ij}], [\alpha r_i])$ at time t/α with the mutation parameters $([\alpha m_{ij}], [\alpha r_i])$ is equal to

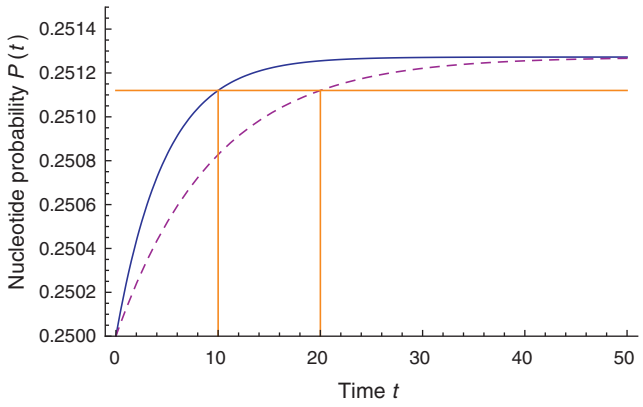

Fig. 1. Time scale property (Eq. (3.1)). Occurrence probability $P_k(t;[m_{ij}],[r_i])$ of a given residue k as a function of time t for two sets of substitution–insertion–deletion parameters: a given set $([m_{ij}],[r_i])$ (solid line) and a set $([\alpha m_{ij}],[\alpha r_i])$ of parameters $([m_{ij}],[r_i])$ multiplied by $\alpha = 0.5$ (dashed line). In particular, when $t = 10$, $P_k(10;[m_{ij}],[r_i]) = P_k(20;[0.5m_{ij}],[0.5r_i]) = 0.2511$.

the occurrence probability $P(t;[m_{ij}],[r_i])$ at time t with the original mutation parameters $([m_{ij}],[r_i])$

$$P\left(\frac{t}{\alpha};[\alpha m_{ij}],[\alpha r_i]\right) = P(t;[m_{ij}],[r_i]). \tag{3.1}$$

Indeed, multiplying the model parameters by a scalar α leads to insertion rates $\tilde{R} = \alpha R$ and substitution probability matrix $\tilde{M} = \alpha M + (1-\alpha)I$ with eigenvalues $\tilde{\lambda}_k = \alpha\lambda_k + 1 - \alpha$, $\forall 1 \le k \le K$. Then, in Eq. (2.13), $\forall 1 \le k \le K$, $(\tilde{r} + 1 - \tilde{\lambda}_k) = \alpha(r + 1 - \lambda_k)$, $\tilde{R}/\left(\tilde{r} + 1 - \tilde{\lambda}_k\right) = R/(r + 1 - \lambda_k)$ and $(\tilde{r} + 1 - \tilde{\lambda}_k)t = (r + 1 - \lambda_k)\alpha t$. This time scale property is illustrated in Fig. 1. As a consequence, the order of magnitude of the mutation parameters (substitution parameters M and insertion rates R) is directly related to time. The larger the parameters are, the faster evolution goes. This time scale property which is classical in substitution models (Jukes and Cantor, 1969; Kimura, 1980, 1981), is also verified in the *IDIS* model. Then, without loss of generality, the total insertion rate r can be set to 1 (only a time shifting).

3.2. Time step

We derive here a time step formula. From Eq. (2.9), the residue occurrence probability $P(t)$ at time t can be written as a function of the residue occurrence probability $P(s)$ at any time s and the time difference $t - s$. For all times $t, s \ge 0$, the probability $P(t)$ at time t is

$$\begin{aligned} P(t) &= e^{A(t-s)}P(s) + e^{At}\left(\int_0^{t-s} e^{-A(v+s)}dv\right)R \\ &= e^{A(t-s)}P(s) + e^{A(t-s)}\left(\int_0^{t-s} e^{-Av}dv\right)R. \end{aligned}$$

This result is obtained by a variable change $v = u - s$ and using the property $e^{-A(v+s)} = e^{-Av}e^{-As}$ as matrices Av and As commute. Let us denote by $F : (y, P(x)) \to e^{Ay}P(x) + e^{Ay}\left(\int_0^{y} e^{-Au}du\right)R$, then for all $s, t \ge 0$, the residue occurrence probability $P(t)$ satisfies

$$P(t) = F(t - s, P(s)). \tag{3.2}$$

3.3. Time inversion

From Eq. (3.2), the time inverse model giving the residue occurrence probability $P(0)$ as a function of $P(t)$ is, for all $t \ge 0$,

$$\begin{aligned} P(0) &= F(-t, P(t)) \\ &= e^{-At}P(t) + e^{-At}\left(\int_0^{-t} e^{-Av}dv\right)R. \end{aligned}$$

This property allows the evolution time direction to be inverted. 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) Eq. (2.10) by replacing t by $-t$.

3.4. Time and sequence length

From the insertion growth rate $n_i'(t)$ defined in Eq. (2.5), the derivative sequence length is $n'(t) = \sum_{1 \le i \le K} n_i'(t) = (r - d)n(t)$ where $r = \sum_{1 \le i \le K} r_i$ is the sum of all residue insertion rates, $\forall 1 \le i \le K$, $r_i \ge 0$, and d is the deletion rate. Then, the number $n(t)$ of residues in the sequence at time t is, for all $t \ge 0$,

$$n(t) = n_0 e^{(r-d)t}.$$

The sequence length $L(t)$ at time t which is equal to the number $n(t)$ of residues in the sequence at time t (all residue lengths are equal to 1), can be written as a function of the sequence length at time s ($s < t$ or $s > t$) and the sum r of insertion rates. For all $s, t \ge 0$,

$$L(t) = L(s)e^{(r-d)(t-s)}.$$

In particular with $s = 0$, this formula yields, for all $t \ge 0$,

$$L(t) = L(0)e^{(r-d)t}. \tag{3.3}$$

4. Analytical occurrence probabilities of nucleotides with the *IDIS-sym*3 model

Genetic sequences are series of residues in the set of nucleotides $\{A, C, G, T\}$ of size $K = 4$. Eq. (2.13) of the *IDIS* model allows to derive analytical expressions of nucleotide occurrence probability along time for various substitution rate matrices. In the continuation of our evolution work, e.g. Arquès and Michel (1990, 1993, 1995), Michel (2007), and Benard and Michel (2009), and as an illustration of the general Eq. (2.13), we derive here expressions of the *IDIS* model for the classical substitution rate matrix with three parameters (Kimura, 1981) and for various particular cases: equal insertion rate, insertion–deletion only, substitution only, stationary distribution and *IDIS* model with 2-parameter (Kimura, 1980) or 1-parameter (Jukes and Cantor, 1969) substitution rate matrix. These expressions of nucleotide occurrence probability are entirely explicit and they can be directly used for biological evolutionary applications without mathematical computation. Some of them will also be used for a comparison between two evolution models in Section 5. Other potential applications are proposed in Section 6.

The *IDIS-sym*3 model gives the expression of nucleotide occurrence probability $P(t)$ using Eq. (2.13) with the classical 3-parameter substitution matrix $M(a, b, c)$ (Kimura, 1981). This matrix $M(a, b, c)$ is defined by three formal parameters a, b, c: a is the rate of transitions $A \leftrightarrow G$ and $C \leftrightarrow T$, b is the rate of transversion type $A \leftrightarrow T$ and $C \leftrightarrow G$, and c is the rate of transversion type $A \leftrightarrow C$ and $G \leftrightarrow T$.

Thus, the substitution matrix $M(a, b, c)$ is defined as follows

$$M(a,b,c) = \begin{pmatrix} n & c & a & b \\ c & n & b & a \\ a & b & n & c \\ b & a & c & n \end{pmatrix},$$

where $n = 1 - (a + b + c)$.

4.1. General formula

We derive the nucleotide occurrence probability $P(t)$ with the *IDIS-sym*3 model from Eq. (2.13), the four eigenvalues of matrix $M(a, b, c)$

$$\{\lambda_1 = 1 - 2(a+b), \lambda_2 = 1 - 2(a+c), \lambda_3 = 1 - 2(b+c), \lambda_4 = 1\}$$

and their associated eigenvectors

$$\{v_1 = \{-1, -1, 1, 1\}, v_2 = \{1, -1, -1, 1\}, v_3 = \{-1, 1, -1, 1\},$$
$$v_4 = \{1, 1, 1, 1\}\}.$$

After some algebraic manipulations, we obtain the occurrence probability $P(t)$ of each nucleotide A, C, G and T at time t

$$P(t) = \frac{1}{4}\begin{pmatrix} 1 + \frac{r_1}{z_1} + \frac{r_2}{z_2} + \frac{r_3}{z_3} + \left(p_1 - \frac{r_1}{z_1}\right)e^{-z_1 t} + \left(p_2 - \frac{r_2}{z_2}\right)e^{-z_2 t} + \left(p_3 - \frac{r_3}{z_3}\right)e^{-z_3 t} \\ 1 + \frac{r_1}{z_1} - \frac{r_2}{z_2} - \frac{r_3}{z_3} + \left(p_1 - \frac{r_1}{z_1}\right)e^{-z_1 t} - \left(p_2 - \frac{r_2}{z_2}\right)e^{-z_2 t} - \left(p_3 - \frac{r_3}{z_3}\right)e^{-z_3 t} \\ 1 - \frac{r_1}{z_1} - \frac{r_2}{z_2} + \frac{r_3}{z_3} - \left(p_1 - \frac{r_1}{z_1}\right)e^{-z_1 t} - \left(p_2 - \frac{r_2}{z_2}\right)e^{-z_2 t} + \left(p_3 - \frac{r_3}{z_3}\right)e^{-z_3 t} \\ 1 - \frac{r_1}{z_1} + \frac{r_2}{z_2} - \frac{r_3}{z_3} - \left(p_1 - \frac{r_1}{z_1}\right)e^{-z_1 t} + \left(p_2 - \frac{r_2}{z_2}\right)e^{-z_2 t} - \left(p_3 - \frac{r_3}{z_3}\right)e^{-z_3 t} \end{pmatrix}, \quad (4.1)$$

where $z_k = r + 1 - \lambda_k$ for $k = 1, 2, 3$ and $r = r_A + r_C + r_G + r_T$; $r_1 = r_A + r_C - r_G - r_T$, $r_2 = r_A - r_C - r_G + r_T$, $r_3 = r_A - r_C + r_G - r_T$; $p_1 = P_A(0) + P_C(0) - P_G(0) - P_T(0)$, $p_2 = P_A(0) - P_C(0) - P_G(0) + P_T(0)$, $p_3 = P_A(0) - P_C(0) + P_G(0) - P_T(0)$.

As parameters a, b, c and r are positive, then constants z_1, z_2 and z_3 are positive and the exponential terms tend to 0 when $t \to \infty$. Thus, the nucleotide equilibrium distribution $P_\infty = \lim_{t\to\infty} P(t)$ is easily deduced

$$P_\infty = \frac{1}{4}\begin{pmatrix} 1 + \frac{r_1}{z_1} + \frac{r_2}{z_2} + \frac{r_3}{z_3} \\ 1 + \frac{r_1}{z_1} - \frac{r_2}{z_2} - \frac{r_3}{z_3} \\ 1 - \frac{r_1}{z_1} - \frac{r_2}{z_2} + \frac{r_3}{z_3} \\ 1 - \frac{r_1}{z_1} + \frac{r_2}{z_2} - \frac{r_3}{z_3} \end{pmatrix}. \quad (4.2)$$

From Eq. (3.3), the evolution time t can be expressed as a function of the sequence length L. By replacing t by $t = (1/(r-d))\ln(L/L_0)$ in Eq. (4.1), we derive the nucleotide occurrence probability $P(L)$ as a function of the sequence length L at evolution time t and the sequence length L_0 at some time $t_0 = 0$

$$P(L) = \frac{1}{4}\begin{pmatrix} 1 + \frac{r_1}{z_1} + \frac{r_2}{z_2} + \frac{r_3}{z_3} + \left(p_1 - \frac{r_1}{z_1}\right)\left(\frac{L}{L_0}\right)^{-z_1/(r-d)} + \left(p_2 - \frac{r_2}{z_2}\right)\left(\frac{L}{L_0}\right)^{-z_2/(r-d)} + \left(p_3 - \frac{r_3}{z_3}\right)\left(\frac{L}{L_0}\right)^{-z_3/(r-d)} \\ 1 + \frac{r_1}{z_1} - \frac{r_2}{z_2} - \frac{r_3}{z_3} + \left(p_1 - \frac{r_1}{z_1}\right)\left(\frac{L}{L_0}\right)^{-z_1/(r-d)} - \left(p_2 - \frac{r_2}{z_2}\right)\left(\frac{L}{L_0}\right)^{-z_2/(r-d)} - \left(p_3 - \frac{r_3}{z_3}\right)\left(\frac{L}{L_0}\right)^{-z_3/(r-d)} \\ 1 - \frac{r_1}{z_1} - \frac{r_2}{z_2} + \frac{r_3}{z_3} - \left(p_1 - \frac{r_1}{z_1}\right)\left(\frac{L}{L_0}\right)^{-z_1/(r-d)} - \left(p_2 - \frac{r_2}{z_2}\right)\left(\frac{L}{L_0}\right)^{-z_2/(r-d)} + \left(p_3 - \frac{r_3}{z_3}\right)\left(\frac{L}{L_0}\right)^{-z_3/(r-d)} \\ 1 - \frac{r_1}{z_1} + \frac{r_2}{z_2} - \frac{r_3}{z_3} - \left(p_1 - \frac{r_1}{z_1}\right)\left(\frac{L}{L_0}\right)^{-z_1/(r-d)} + \left(p_2 - \frac{r_2}{z_2}\right)\left(\frac{L}{L_0}\right)^{-z_2/(r-d)} - \left(p_3 - \frac{r_3}{z_3}\right)\left(\frac{L}{L_0}\right)^{-z_3/(r-d)} \end{pmatrix}, \quad (4.3)$$

where d is the deletion rate, $p_1 = P_A(L_0) + P_C(L_0) - P_G(L_0) - P_T(L_0)$, $p_2 = P_A(L_0) - P_C(L_0) - P_G(L_0) + P_T(L_0)$, $p_3 = P_A(L_0) - P_C(L_0) + P_G(L_0) - P_T(L_0)$ and the remaining parameters z_1, z_2, z_3, r, r_1, r_2, r_3 as in Eq. (4.1). Eq. (4.3) will be used for an evolution model comparison in Section 5.

4.2. Equal insertion rate formula

When the nucleotide insertion rates are all equal ($r_A = r_C = r_G = r_T \neq 0$), then the occurrence probability $P(t)$ of each nucleotide A, C, G and T at time t simplifies

$$P(t) = \frac{1}{4}\begin{pmatrix} 1 + p_1 e^{-z_1 t} + p_2 e^{-z_2 t} + p_3 e^{-z_3 t} \\ 1 + p_1 e^{-z_1 t} - p_2 e^{-z_2 t} - p_3 e^{-z_3 t} \\ 1 - p_1 e^{-z_1 t} - p_2 e^{-z_2 t} + p_3 e^{-z_3 t} \\ 1 - p_1 e^{-z_1 t} + p_2 e^{-z_2 t} - p_3 e^{-z_3 t} \end{pmatrix}, \quad (4.4)$$

where $z_1, z_2, z_3, p_1, p_2, p_3$ are defined as in Eq. (4.1). The nucleotide equilibrium distribution P_∞ is

$$P_\infty = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)^t. \quad (4.5)$$

The nucleotide equilibrium distribution for equal insertion rates is independent from the insertion rates. It is equal to the nucleotide equilibrium distribution (4.9) obtained with the substitution only model.

4.3. Insertion–deletion formula

When the substitution probabilities are all equal to 0 ($a = b = c = 0$), then the occurrence probability $P(t)$ of each nucleotide A, C, G and T at time t becomes

$$P(t) = \begin{pmatrix} \frac{r_A}{r} + \left(P_A(0) - \frac{r_A}{r}\right)e^{-rt} \\ \frac{r_C}{r} + \left(P_C(0) - \frac{r_C}{r}\right)e^{-rt} \\ \frac{r_G}{r} + \left(P_G(0) - \frac{r_G}{r}\right)e^{-rt} \\ \frac{r_T}{r} + \left(P_T(0) - \frac{r_T}{r}\right)e^{-rt} \end{pmatrix}, \quad (4.6)$$

where $r = r_A + r_C + r_G + r_T$. The nucleotide equilibrium distribution P_∞ is

$$P_\infty = \left(\frac{r_A}{r}, \frac{r_C}{r}, \frac{r_G}{r}, \frac{r_T}{r}\right)^t. \quad (4.7)$$

4.4. Substitution formula

When the insertion rates are all equal to 0 ($r_A = r_C = r_G = r_T = 0$), then the occurrence probability $P(t)$ of each nucleotide A, C, G and T at time t becomes

$$P(t) = \frac{1}{4}\begin{pmatrix} 1 + p_1 e^{-2(a+b)t} + p_2 e^{-2(a+c)t} + p_3 e^{-2(b+c)t} \\ 1 + p_1 e^{-2(a+b)t} - p_2 e^{-2(a+c)t} - p_3 e^{-2(b+c)t} \\ 1 - p_1 e^{-2(a+b)t} - p_2 e^{-2(a+c)t} + p_3 e^{-2(b+c)t} \\ 1 - p_1 e^{-2(a+b)t} + p_2 e^{-2(a+c)t} - p_3 e^{-2(b+c)t} \end{pmatrix}, \quad (4.8)$$

where a, b, c are the substitution rates of $M(a, b, c)$ and p_1, p_2, p_3 are defined as in Eq. (4.1). The nucleotide equilibrium distribution P_∞ is straightforward

$$P_\infty = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)^t \quad (4.9)$$

and is equal to P_∞ for equal insertion rates (4.5).

4.5. Stationary distribution

The stationary distribution varies between two asymptotes, the stationary distribution for substitution only which is equal to 1/4 for each nucleotide whatever the substitution parameters a, b and c (Eq. (4.9)) and the stationary distribution for insertion–deletion only which depends on the insertion rates r_A, r_C, r_G and r_T (Eq. (4.7)). Fig. 2 illustrates this property for nucleotide A. The stationary distribution depends on the order of magnitude of the insertion rates with respect to the substitution rates.

4.6. Particular cases

We now derive the expressions of $P(t)$ from Eq. (2.13) for a 2-parameter substitution matrix $M_2 = M(\gamma, \beta/2, \beta/2)$ (*IDIS-sym2* model) and for a 1-parameter substitution matrix $M_1 = M(\alpha/3, \alpha/3, \alpha/3)$ (*IDIS-sym1* model). Matrices M_2 and M_1 are also used for phylogenetic inference.

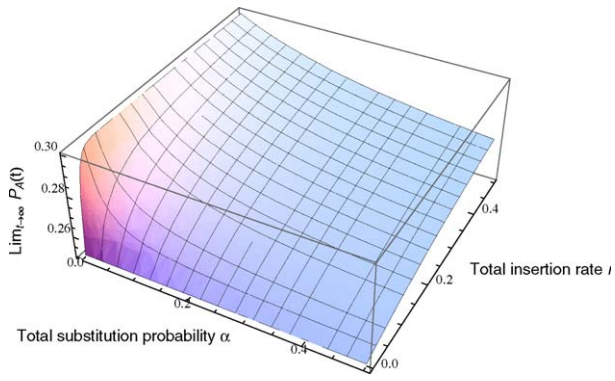

Fig. 2. Stationary nucleotide occurrence probability. Stationary occurrence probability $\lim_{t\to\infty} P_A(t)$ for nucleotide A (z-axe) as a function of the total substitution probability $\alpha = a + b + c$ (x-axe) and the total nucleotide insertion rate $r = r_A + r_C + r_G + r_T$ (y-axe). In this example, $a = b = c = \alpha/3$ and $r_A = (3/10)r$, $r_C = (4/10)r$, $r_G = (2/10)r$ and $r_T = (1/10)r$. As expected, when $r = 0$, then $\lim_{t\to\infty} P_A(t)$ is equal to 1/4 (Eq. (4.9)) and when $\alpha = 0$, $\lim_{t\to\infty} P_A(t)$ is equal to $(r_A/r) = 3/10$ (Eq. (4.7)).

4.6.1. Analytical occurrence probabilities of nucleotides with the IDIS-sym2 model

The classical substitution matrix M_2 (Kimura, 1980) is defined by two formal parameters γ and β: γ is the rate of transitions $A \leftrightarrow G$ and $C \leftrightarrow T$, and β is the rate of transversions $A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$ and $G \leftrightarrow T$

$$M_2 = \begin{pmatrix} n & \frac{\beta}{2} & \gamma & \frac{\beta}{2} \\ \frac{\beta}{2} & n & \frac{\beta}{2} & \gamma \\ \gamma & \frac{\beta}{2} & n & \frac{\beta}{2} \\ \frac{\beta}{2} & \gamma & \frac{\beta}{2} & n \end{pmatrix},$$

where $n = 1 - (\beta + \gamma)$. Matrix M_2 is a particular case of matrix $M(a, b, c)$ where $a = \gamma$ and $b = c = \beta/2$. The occurrence probability $P(t)$ of each nucleotide A, C, G and T at time t is

$$P(t) = \frac{1}{4}\begin{pmatrix} 1 + 2\frac{r_A - r_G}{z_1} + \frac{r_3}{z_3} + 2\left(P_A(0) - P_G(0) - \frac{r_A - r_G}{z_1}\right)e^{-z_1 t} + \left(p_3 - \frac{r_3}{z_3}\right)e^{-z_3 t} \\ 1 + 2\frac{r_C - r_T}{z_1} - \frac{r_3}{z_3} + 2\left(P_C(0) - P_T(0) - \frac{r_C - r_T}{z_1}\right)e^{-z_1 t} - \left(p_3 - \frac{r_3}{z_3}\right)e^{-z_3 t} \\ 1 - 2\frac{r_A - r_G}{z_1} + \frac{r_3}{z_3} - 2\left(P_A(0) - P_G(0) - \frac{r_A - r_G}{z_1}\right)e^{-z_1 t} + \left(p_3 - \frac{r_3}{z_3}\right)e^{-z_3 t} \\ 1 - 2\frac{r_C - r_T}{z_1} - \frac{r_3}{z_3} - 2\left(P_C(0) - P_T(0) - \frac{r_C - r_T}{z_1}\right)e^{-z_1 t} - \left(p_3 - \frac{r_3}{z_3}\right)e^{-z_3 t} \end{pmatrix},$$

where the constants z_1 and z_3 become $z_1 = 2\gamma + \beta + r$ and $z_3 = 2\beta + r$ with $r = r_A + r_C + r_G + r_T$, and the remaining parameters are defined as in Eq. (4.1): $p_3 = P_A(0) - P_C(0) + P_G(0) - P_T(0)$ and $r_3 = r_A - r_C + r_G - r_T$.

4.6.2. Analytical occurrence probabilities of nucleotides with the IDIS-sym1 model

The classical substitution matrix M_1 (Jukes and Cantor, 1969) is defined by one formal parameter α where α is the substitution probability per site

$$M_1 = \begin{pmatrix} n & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & n & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & n & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & n \end{pmatrix},$$

where $n = 1 - \alpha$. Matrix M_1 is a particular case of matrix $M(a, b, c)$ where $a = b = c = \alpha/3$. The occurrence probability $P(t)$ of each nucleotide A, C, G and T at time t is

$$P(t) = \frac{1}{4}\begin{pmatrix} 1 + \frac{4r_A - r}{z_1} + 4\left(P_A(0) - \frac{3r_A + \alpha}{3z_1}\right)e^{-z_1 t} \\ 1 + \frac{4r_C - r}{z_1} + 4\left(P_C(0) - \frac{3r_C + \alpha}{3z_1}\right)e^{-z_1 t} \\ 1 + \frac{4r_G - r}{z_1} + 4\left(P_G(0) - \frac{3r_G + \alpha}{3z_1}\right)e^{-z_1 t} \\ 1 + \frac{4r_T - r}{z_1} + 4\left(P_T(0) - \frac{3r_T + \alpha}{3z_1}\right)e^{-z_1 t} \end{pmatrix},$$

where $z_1 = (4/3)\alpha + r$ and $r = r_A + r_C + r_G + r_T$.

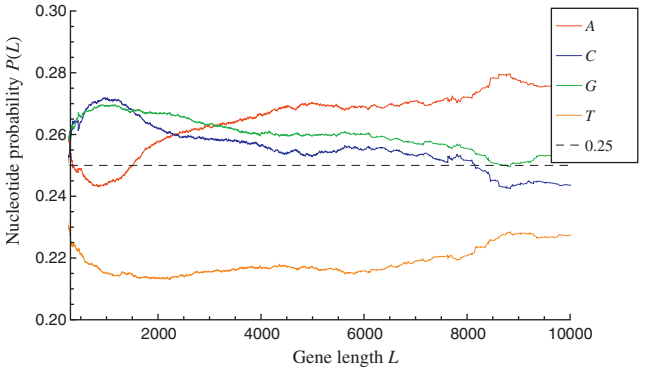

Fig. 3. Averaged occurrence probabilities $P(L)$ of each nucleotide A, C, G and T observed in human genes as a function of their nucleotide length L. The curves are smoothed with a moving average window of 250 nucleotides.

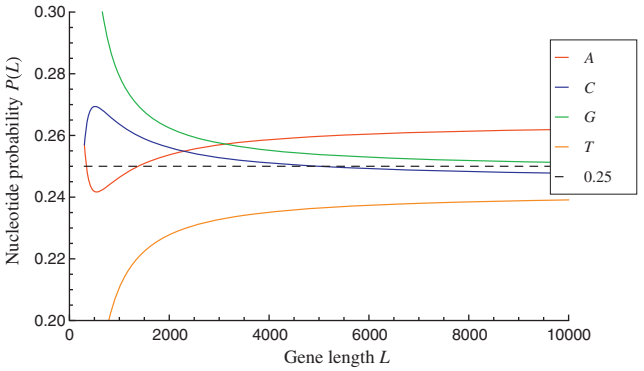

Fig. 4. *IDIS* model. Example of nucleotide occurrence probability $P(L)$ obtained with the *IDIS-sym*3 model using parameters: $L_0 = 300$, $P_A(L_0) = P_C(L_0) = 0.257$, $P_G(L_0) = 0.4$, $P_T(L_0) = 0.086$, $a = 0$, $b = 0.511$, $c = 0$, $r_A = 0.276$, $r_C = 0.244$, $r_G = 0.250$, $r_T = 0.230$, $d = 0$. These curves reproduce several statistical features observed with nucleotides in human genes (Fig. 3 and (5.1)).

5. Comparison of the *IDIS* model with a substitution only model on a gene dataset

To our knowledge, the *IDIS* model is the unique analytical approach to study nucleotide occurrence probabilities according to a comprehensive mutation process (residue insertion, deletion and substitution according to independent processes). In order to show the important difference in stochastic behavior between the *IDIS* model and a classical substitution only model, we have taken an example of nucleotide occurrence probabilities in a human gene dataset (NCBI web site, built 36 version 3 also known as hg18). After excluding the rare cases of extreme gene lengths, Fig. 3 shows the occurrence probabilities $P(L)$ of each nucleotide A, C, G and T in human genes as a function of their nucleotide length L varying from 300 to 10,000.

The main statistical features of these data can be described as simple approximations according to inequalities which range the nucleotide probabilities as a function of their values and their limit probability 1/4. Six intervals I of gene lengths can be described as follows

$$\begin{array}{ll}
P_A(L) \approx P_G(L) \approx P_C(L) > 0.25 > P_T(L) & \text{when } L \in I_1 = [300, 400[\\
P_G(L) \approx P_C(L) > 0.25 > P_A(L) > P_T(L) & \text{when } L \in I_2 = [400, 1500[\\
P_G(L) > P_C(L) > P_A(L) > 0.25 > P_T(L) & \text{when } L \in I_3 = [1500, 2350[\\
P_G(L) > P_A(L) > P_C(L) > 0.25 > P_T(L) & \text{when } L \in I_4 = [2350, 3100[\\
P_A(L) > P_G(L) > P_C(L) > 0.25 > P_T(L) & \text{when } L \in I_5 = [3100, 8100[\\
P_A(L) > P_G(L) > 0.25 > P_C(L) > P_T(L) & \text{when } L \in I_6 = [8100, 10000]
\end{array} \quad . \tag{5.1}$$

The *IDIS* model can describe a complete evolution process (insertion, deletion and substitution) of genes of short lengths. We use the *IDIS-sym*3 model giving the nucleotide occurrence probability as a function of the sequence length (Eq. (4.3)). The initial length L_0 is set to the lower limit of the studied data, i.e. $L_0 = 300$. For sake of simplicity, we set the deletion rate d to 0 as the residue occurrence probability is not affected by deletion (Remark 5) and the total insertion r to 1 (Section 3.1). No scan within the definition set of parameters could simultaneously satisfy the inequalities (5.1). However, there exists a set of parameters (Fig. 4) which simulates the observed statistical features (5.1) for nucleotides A and C with gene lengths varying from 300 to 10,000 nucleotides (intervals I_1 to I_6) and for nucleotides G and T with gene lengths varying from 1500 to 10,000 nucleotides (intervals I_3 to I_6). The *IDIS* model allows us to reproduce the global nucleotide probability behavior of human genes of lengths varying from 1500 to 10,000 nucleotides shown in Fig. 3 (except for $P_G(L_0) = 0.4$ which is higher than the observed probability ≈0.26 and consequently for the dependent probability $P_T(L_0)$ with an opposite situation).

In order to evaluate the impact of the nucleotide insertion on the nucleotide occurrence probabilities, all nucleotide insertion rates are set to 0 ($r_A = r_C = r_G = r_T = 0$), i.e. Eq. (4.8). Keeping the other model parameters identical to the *IDIS-sym*3 model (legend of Fig. 4), the substitution only model gives the occurrence probabilities $P(t)$ of each nucleotide A, C, G and T as a function of t (Fig. 5). At the beginning of the substitution process, the stochastic curves of A, C, G and T have a behavior similar to the curves of the *IDIS* model (Figs. 4 and 5). By increasing the number of substitutions, the nucleotide probabilities converge as expected to 0.25 (Eq. (4.9)) and they become completely different from the real curves observed in large human genes (Fig. 3).

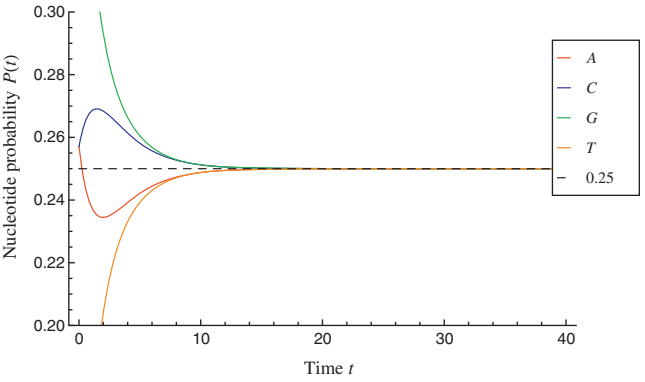

Fig. 5. Substitution only model. Nucleotide occurrence probabilities $P(t)$ with the *IDIS-sym*3 model when all nucleotide insertion rates are set to 0 ($r_A = r_C = r_G = r_T = 0$) and the other model parameters are identical to Fig. 4. These curves can only simulate the observed statistical features for nucleotides A and C in short human genes (intervals I_1 and I_2 in (5.1)).

6. Conclusions

We developed here a new class of stochastic evolution models for residue Insertion–Deletion Independent from Substitution called *IDIS*. The *IDIS* model was inspired by a concept in population dynamics and has the original property that the insertion and deletion processes are not only independent from each other, but they are also independent from the substitution process.

We give a general analytical expression of the residue occurrence probability at time t which can be used for any diagonalizable substitution rate matrix (Eq. (2.13)). Thus, the *IDIS* model gives the residue occurrence probability at time t as a function of a substitution rate matrix M, an insertion rate vector R, a deletion rate d and an initial residue probability vector $P(0)$. The classical substitution only models (Jukes and Cantor, 1969; Kimura, 1980, 1981, and their extensions) become particular cases of the *IDIS* model. Several mathematical properties are also derived: time scale, time step, time inversion and a relation between time and sequence length.

In the continuation of our evolution work and as an illustration of the general expression (Eq. (2.13)), we give an explicit formula for the *IDIS-sym*3 model with the classical substitution rate matrix with three parameters (Kimura, 1981). Expressions for various particular biological contexts are also given: equal insertion rate, insertion–deletion only, substitution only, and also the *IDIS-sym*2 and *IDIS-sym*1 models associated with 2-parameter and 1-parameter substitution matrices (Kimura, 1980; Jukes and Cantor, 1969), respectively. All these expressions can be directly used for biological evolutionary applications.

The *IDIS* model showed a strongly different stochastic behavior from a classical substitution only model with an example of human genes. It can also be used for deriving phylogenetic distances or inferring phylogenetic trees from sequence alignment. Finally, by considering three independent processes for insertion, deletion and substitution, the *IDIS* model allows a more realistic representation of gene evolution and opens new directions in this research field.

References

Arquès, D.G., Michel, C.J., 1990. A model of DNA sequence evolution. Part 1: Statistical features and classification of gene populations, 743–753. Part 2: Simulation model, 753–766. Part 3: Return of the model to the reality, 766–770. Bull. Math. Biol. 52, 741–772.

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.

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. J. Comput. Biol. Chem. 33, 245–252.

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

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

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

Hubbard, J.H., West, B.H., 1995. Differential Equations: A Dynamical Systems Approach. Springer.

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

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

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

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

Malthus, T.R., 2000. An Essay on the Principle of Population. Library of Economics. Liberty, Fund, Inc.

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

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

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

Miklós, I., Lunter, G.A., Holmes, I., 2004. A "long indel" model for evolutionary sequence alignment. Mol. Biol. E 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 Bioinformatics 6, 63.

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

Thorne, J.L., Kishino, H., Felsenstein, J., 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. E 33, 114–124.

Thorne, J.L., Kishino, H., Felsenstein, J., 1992. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. E 34, 3–16.

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. E 10, 512–526.

Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17, 57–86.

Yang, Z., 1994. Estimating the pattern of nucleotide substitution. J. Mol. E 39, 105–111.

Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, New York.