Mathematical Biosciences 245 (2013) 137–147

Contents lists available at SciVerse ScienceDirect

Mathematical Biosciences

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

# A new molecular evolution model for limited insertion independent of substitution

Sophie Lèbre, Christian J. Michel *

*Equipe de Bioinformatique Théorique, BFO, ICube, Université de Strasbourg, CNRS, 300 Boulevard Sébastien Brant, 67400 Illkirch, France*

ARTICLE INFO

*Article history:*
Received 11 September 2012
Received in revised form 26 May 2013
Accepted 3 June 2013
Available online 13 June 2013

*Keyword:*
Evolution models
Substitution
Limited insertion
Sequence length
Sequence content
GC content

ABSTRACT

We recently introduced a new molecular evolution model called the *IDIS* model for Insertion Deletion Independent of Substitution [13,14]. In the *IDIS* model, the three independent processes of substitution, insertion and deletion of residues have constant rates. In order to control the genome expansion during evolution, we generalize here the *IDIS* model by introducing an insertion rate which decreases when the sequence grows and tends to 0 for a maximum sequence length $n_{\max}$.

This new model, called *LIIS* for Limited Insertion Independent of Substitution, defines a matrix differential equation satisfied by a vector $P(t)$ describing the sequence content in each residue at evolution time $t$. An analytical solution is obtained for any diagonalizable substitution matrix $M$. Thus, the *LIIS* model gives an expression of the sequence content vector $P(t)$ in each residue under evolution time $t$ as a function of the eigenvalues and the eigenvectors of matrix $M$, the residue insertion rate vector $R$, the total insertion rate $r$, the initial and maximum sequence lengths $n_0$ and $n_{\max}$, respectively, and the sequence content vector $P(t_0)$ at initial time $t_0$. The derivation of the analytical solution is much more technical, compared to the *IDIS* model, as it involves Gauss hypergeometric functions.

Several propositions of the *LIIS* model are derived: proof that the *IDIS* model is a particular case of the *LIIS* model when the maximum sequence length $n_{\max}$ tends to infinity, fixed point, time scale, time step and time inversion. Using a relation between the sequence length $l$ and the evolution time $t$, an expression of the *LIIS* model as a function of the sequence length $l = n(t)$ is obtained. Formulas for 'insertion only', i.e. when the substitution rates are all equal to 0, are derived at evolution time $t$ and sequence length $l$. Analytical solutions of the *LIIS* model are explicitly derived, as a function of either evolution time $t$ or sequence length $l$, for two classical substitution matrices: the 3-parameter symmetric substitution matrix [12] (*LIIS-SYM*3) and the *HKY* asymmetric substitution matrix [9] (*LIIS-HKY*).

An evaluation of the *LIIS* model (precisely, *LIIS-HKY*) based on four statistical analyses of the *GC* content in complete genomes of four prokaryotic taxonomic groups, namely Chlamydiae, Crenarchaeota, Spirochaetes and Thermotogae, shows the expected improvement from the theory of the *LIIS* model compared to the *IDIS* model.

## 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. Stochastic evolution models were initially developed to study the substitution rates of nucleotides (adenine *A*, cytosine *C*, guanine *G*, thymine *T*). The first substitution models were based on symmetric substitution matrices with one formal parameter for all nucleotide substitution types [10], two formal parameters for the nucleotide transitions and transversions [11] and three formal parameters for transitions and the two types of transversions [12]. These substitution models were later generalized to asymmetric substitution matrices [6,30,9,32,31,38,7] with an equilibrium distribution different from 1/4 for all nucleotides.

Over the last 20 years, only very few molecular evolution models were extended to the insertion and the deletion of residues (nucleotides, amino acids) in addition to residue substitution. These substitution insertion deletion (SID) models were designed for statistical alignement of two sequences and can be divided into three classes. A pioneering paper by Thorne et al. [34] proposed a time-reversible Markov model for insertions and deletions (termed the TKF91 model). This SID model represents the sequence evolution in two steps. First, the sequence is subject to an insertion-deletion process which is homogeneous over all sites in the sequence. Second, and conditional on the result of the insertion-deletion process, a substitution process is applied to the two sequences. The process is time-reversible whenever the substitution process is. Some drawbacks of the preliminary TKF91 proposal have first

* Corresponding author. Tel.: +33 368854462.
*E-mail addresses:* slebre@unistra.fr (S. Lèbre), c.michel@unistra.fr (C.J. Michel).

0025-5564/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.mbs.2013.06.001

been improved by the same authors with the TKF92 version of the model [35]. Then, the original SID models have been later refined in many ways, as for instance by Metzler [17] and Miklòs et al. [19] (see e.g. [20] for a review). A second class of SID models was introduced by McGuire et al. [16] who defined a Markov model by extending the F84 substitution matrix [7] 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 [25] with a non-reversible evolution model which extends the model of McGuire et al. [16] 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. An analytical expression of the substitution probabilities $P_t(i,j)$ of residue $i$ by residue $j$ over time $t$ is given in the particular case where the insertion rate is proportional to the substitution matrix equilibrium distribution [26]. However, even if the insertion process is independent of the substitution process, the substitution and deletion processes are not independent. Indeed, the occurrence probability $P_i(t)$ of residue $i$ at time $t$ which can be derived from $P_t(i,j)$ depends on the deletion rate. However, a deletion rate which is identical for all residues (uniform deletion rate) is expected to alter the sequence length but obviously not the residue distribution (detailed in Introduction in [14]).

Inspired by a concept in population dynamics [15], we have developed a dynamic evolution model, called the *IDIS* model, where the three processes of substitution, insertion and deletion of nucleotides are independent of each other [13,14]. The *IDIS* model gives an analytical expression of the sequence content vector $P(t)$ at evolution time $t$ [13] or $P(l)$ at sequence length $l$ [14] for any diagonalizable substitution matrix $M$ as a function of the eigenvalues and the eigenvectors of matrix $M$, the vector $R$ of the residue insertion rates, the total insertion rate $r$, the deletion rate $d$ and the vector of initial sequence content $P(t_0)$ at evolution time $t_0$ or $P(n_0)$ at sequence length $n_0$. It presents several interesting mathematical properties compared to all mathematical models in this research field: (i) it has a uniform deletion rate which does not alter the sequence content as expected from a probabilistic point of view; (ii) it relies 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, it allows a realistic interpretation of the model parameters (evolution time $t$, sequence length $l$ and rates of substitution, insertion and deletion); (iii) it allows 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.; (iv) it provides a description of the sequence content evolution and in particular the evolution of motif content inside the sequence, contrary to the phylogenetic approaches for tree reconstruction; and (v) it extends our previous approaches developed over the last 20 years for substitution models (e.g. [1,2,18,4,5]) which allowed to introduce models of ‘primitive’ genes or ‘primitive’ motifs of nucleotides or amino acids, to study substitution rates, to analyse the residue occurrence probabilities in the natural evolution time direction (from past to present or from present to future) or in the inverse direction (from present to past).

In the *IDIS* model, the growth rate describing the insertion process is constant. We generalize here the *IDIS* model with an insertion process whose rate varies during evolution time. In a concept similar to the limited growth model for population dynamics by Verhulst [36], the insertion rate decreases when the sequence grows and tends to 0 for a maximum sequence length $n_{\max}$. This new model, called *LIIS* for Limited Insertion Independent of Substitution, is defined by a matrix differential equation, for which an analytical solution is obtained for any diagonalizable substitution matrix $M$ and involves Gauss hypergeometric functions. Thus, the *LIIS* model gives an analytical expression of the content vector $P(t)$ in each residue in the sequence at evolution time $t$ as a function of the eigenvalues and the eigenvectors of matrix $M$, the residue insertion rate vector $R$, the total insertion rate $r$, the initial and maximum sequence lengths $n_0$ and $n_{\max}$, respectively, and the initial sequence content vector $P(t_0)$ at initial time $t_0$.

This paper is organized as follows. Section 2 introduces the mathematical model *LIIS*. Section 3 gives several propositions of the *LIIS* model: proof that the *IDIS* model is a particular case of the *LIIS* model when the maximum sequence length $n_{\max}$ tends to infinity, residue equilibrium distribution, time scale, time step and time inversion. Section 4 derives an expression of the *LIIS* model as a function of the sequence length $l = n(t)$. Section 5 gives formulas for ‘insertion only’, i.e. when the substitution rates are all equal to 0, both at evolution time $t$ and sequence length $l$. Section 6 derives the analytical solutions of the *LIIS* model for the two classical substitution matrices both at evolution time $t$ and sequence length $l$: the 3-parameter symmetric substitution matrix [12] (*LIIS-SYM*3) and the *HKY* asymmetric substitution matrix [9] (*LIIS-HKY*). In Section 7, an evaluation of the *LIIS* model (precisely, *LIIS-HKY*) based on four statistical analyses of the *GC* content in complete genomes of four prokaryotic taxonomic groups, namely Chlamydiae, Crenarchaeota, Spirochaetes and Thermotogae, shows the expected improvement from the theory of the *LIIS* model compared to the *IDIS* model.

## 2. Mathematical model

We present here a new molecular evolution model for Limited Insertion Independent of Substitution (*LIIS*). The originality of the *LIIS* model relies on two points: (i) as in the *IDIS* model, the insertion process is independent of the substitution process; and (ii) contrary to the *IDIS* model, the insertion rate is time dependent, decreases when the sequence grows and tends to 0 for a maximum sequence length $n_{\max}$. Hence, the *LIIS* model generalizes the *IDIS* model in the particular case of an insertion-substitution model (Proposition 3 in Section 3).

Before deriving the general *LIIS* model equation, we analyse the limited insertion and the substitution processes separately by building a specific differential equation for each evolution process.

### 2.1. Limited insertion model

Let us consider an alphabet of $K$ residues, e.g. $K = 4$ for nucleotides and $K = 20$ for amino acids. For all $1 \leqslant i \leqslant K$, we denote by $n_i(t)$ the occurrence number of residue $i$ in the sequence at time $t$ and by $n(t) = \sum_{1 \leqslant i \leqslant K} n_i(t)$ the sequence length. In the *IDIS* model, the growth rate of residue $i$ resulting from the insertion-deletion process is assumed to be equal to $n_i'(t) = \frac{\partial n_i(t)}{\partial t} = r_i n(t) - d n_i(t)$, for all $1 \leqslant i \leqslant K$, where $r_i$ is a specific instantaneous insertion rate for each residue $i$ and $d$ is a uniform deletion rate applied to any residue. Thus, the sequence length $n(t)$ at time $t$ is equal to the expected length of a random sequence subject to a linear birth–death process with birth rate equal to $\lambda = \sum_i r_i n(t)$ and a death rate equal to $\mu = dn(t)$, i.e. $n(t) = n_0 e^{(\sum_i r_i - d)t}$ where $n_0$ is the initial sequence length.

In order to generalize the *IDIS* model where the sequence growth rate is constant, we now consider in the *LIIS* model that the residue insertion rate depends on the sequence length.

Similarly to the population dynamics model introduced by Verhulst [36], we set the growth rate $n_i'(t)$ of residue $i$ at time $t$ in the limited insertion process equal to, for all $1 \leqslant i \leqslant K$,

$$n_i'(t) = r_i\left(1 - \frac{n(t)}{n_{\max}}\right)n(t) \qquad (2.1)$$

where $n_{\max}$ is the maximal sequence length which modulates the insertion process ($n_{\max} \geqslant 2$, see Condition of Eq. (2.12)). Note that, as in the *IDIS* model, the insertion process is modelled by explicit parameters which are set independently of the substitution parameters: $r_i$, the insertion rate per site of each residue $i, \forall 1 \leqslant i \leqslant K$, $r_i \geqslant 0$.

**Remark 1.** $\lim_{n(t)\to n_{\max}} n_i'(t) = 0$. When the sequence length $n(t)$ increases to $n_{\max}$ then the growth rate $n_i'(t)$ of residue $i$ decreases to 0 and only the substitution process is active.

**Remark 2.** $\lim_{\frac{n(t)}{n_{\max}}\to 0} n_i'(t) = r_i n(t)$. When the sequence length $n(t)$ is much smaller than $n_{\max}$ ($n(t) \ll n_{\max}$) then the growth rate $n_i'(t)$ of residue $i$ in the *LIIS* model is equal to the growth rate of residue $i$ in the *IDIS* model (Eq. (2.5) with $d = 0$ in [13]).

Let $r = \sum_{1\leqslant i\leqslant K} r_i$ be the total residue insertion rate. The total sequence length variation rate $n'(t) = \sum_{1\leqslant i\leqslant K} n_i'(t)$ is equal to

$$n'(t) = r\left(1 - \frac{n(t)}{n_{\max}}\right)n(t). \qquad (2.2)$$

The solution of this differential Eq. (2.2) with initial sequence length $n_0$ at time $t_0$ gives

$$n(t) = \frac{n_0}{\tau + (1-\tau)e^{-r(t-t_0)}} \qquad (2.3)$$

with $\tau = \frac{n_0}{n_{\max}}$.

Let $P_i(t) = \frac{n_i(t)}{n(t)}$ be the sequence content in residue $i$ at time $t \geqslant 0$. The column vector $P(t) = [P_i(t)]_{1\leqslant i\leqslant K}$ of size $K$ is made of the sequence content $P_i(t)$ in residue $i$ for all $1 \leqslant i \leqslant K$. Using Eqs. (2.1) and (2.2), the derivative $P_i'(t)$ of the sequence content in residue $i$ at time $t$ in the limited insertion process reads

$$\begin{aligned}
P_i'(t) &= \frac{\partial}{\partial t}\left(\frac{n_i(t)}{n(t)}\right) = \frac{n_i'(t)n(t) - n_i(t)n'(t)}{n^2(t)} = \frac{n_i'(t)}{n(t)} - \frac{n'(t)}{n(t)}P_i(t) \\
&= \frac{r_i\left(1 - \frac{n(t)}{n_{\max}}\right)n(t)}{n(t)} - r\left(1 - \frac{n(t)}{n_{\max}}\right)P_i(t) \\
&= \left(1 - \frac{n(t)}{n_{\max}}\right)(r_i - rP_i(t)).
\end{aligned}$$

Finally, the derivative $P'(t)$ of the sequence content at time $t$ in the limited insertion process is modelled by

$$P'(t) = \theta(t)(R - rP(t)) \qquad (2.4)$$

where $r = \sum_{1\leqslant i\leqslant K} r_i$ is the total residue insertion rate, $R = [r_i]_{1\leqslant i\leqslant K}$ is the vector of residue insertion rates and $\theta(t) = 1 - \frac{n(t)}{n_{\max}}$ is, using Eq. (2.3), equal to

$$\theta(t) = 1 - \frac{\tau}{\tau + (1-\tau)e^{-r(t-t_0)}}. \qquad (2.5)$$

### 2.2. Substitution model

The sequence content evolution due to the substitution process is defined as in the *IDIS* model, i.e. such that the sequence content vector $P(t) = [P_i(t)]_{1\leqslant i\leqslant K}$ is equal to the expected content of a random sequence subject to a classical substitution process defined by a constant substitution rate matrix, each site in the sequence being independent and identically distributed.

Thus, the substitution process is handled by the following differential equation (e.g. [18]) which determines the sequence content vector $P(t)$ for all time $t \geqslant 0$,

$$P'(t) = M \cdot P(t) - P(t) = (M - I) \cdot P(t) \qquad (2.6)$$

where $M = [m_{ij}]_{1\leqslant i,j\leqslant K}$ is a constant substitution rate matrix, stochastic in column, i.e. with element $m_{ij} = P(j \to i)$ in row $i$ and column $j$ referring to the substitution rate of residue $j$ into residue $i$, matrix $I$ is the identity matrix of size $K$ and the symbol $\cdot$ is the matrix product.

**Remark 3.** The substitution rate matrix $M$ is the transpose matrix of the classical substitution matrix $\pi = [P(i \to j)]_{1\leqslant i,j\leqslant K}$ which is stochastic in line (e.g. [11,12]), i.e. $\pi_{ij} = m_{ji}$.

### 2.3. LIIS model: limited insertion independent of substitution

The substitution and the limited insertion processes are assumed to be independent. From the two differential equations describing the residue substitution (Eq. (2.6)) and the residue limited insertion (Eq. (2.4)), we derive a general matrix differential equation allowing for these two processes to be superimposed. Then, the derivative $P'(t)$ of the sequence content at time $t$ is the result of the instantaneous variation due to substitution and limited insertion, and the sequence content vector $P(t) = [P_i(t)]_{1\leqslant i\leqslant K}$ satisfies

$$P'(t) = \underbrace{(M - I) \cdot P(t)}_{\text{Substitution}} + \underbrace{\theta(t)(R - rP(t))}_{\text{Limited insertion}} = A(t) \cdot P(t) + \theta(t)R \qquad (2.7)$$

where $A(t) = M - (1 + r\theta(t))I, \theta(t)$ is defined in Eq. (2.5), $M$ is the substitution rate matrix defined in Eq. (2.6), $R = [r_i]_{1\leqslant i\leqslant K}$ is the vector of the residue insertion rates per site and $r = \sum_{1\leqslant i\leqslant K} r_i$ is the total residue insertion rate, $\forall 1 \leqslant i \leqslant K$, $r_i \geqslant 0$.

This nonhomogeneous matrix linear differential equation with non-constant coefficients can be easily solved in the particular case where, for all $s, t \geqslant 0$, matrices $A(t)$ and $A(s)$ commute (Section 3.4 of Part I in [33]). This condition is satisfied here as the time dependent term $\theta(t)$ in matrix $A(t) = M - (1 + r\theta(t))I$ is in the diagonal. Then, for all $s, t \geqslant 0, [A(t), A(s)] = A(t)A(s) - A(s)A(t) = 0$ and the solution of Eq. (2.7) is, for all $t \geqslant 0$, for all initial time $t_0 \geqslant 0$,

$$P(t; t_0, P(t_0)) = e^{\left(\int_{t_0}^{t} A(u)du\right)} \cdot P(t_0) + \int_{t_0}^{t} \theta(s)e^{\left(\int_{s}^{t} A(u)du\right)} \cdot Rds. \qquad (2.8)$$

When the substitution rate matrix $M$ is diagonalizable with real eigenvalues $(\lambda_k)_{1\leqslant k\leqslant K}$ then $A(t)$ is also diagonalizable with real eigenvalues equals to $(\lambda_k - 1 - r\theta(t))_{1\leqslant k\leqslant K}$. Indeed, if matrix $M$ decomposes as $M = Q \cdot D \cdot Q^{-1}$ where $D = Diag((\lambda_k)_{1\leqslant k\leqslant K})$ is the eigenvalues diagonal matrix and $Q$ is an associated eigenvectors matrix, the $k$th column of $Q$ being an eigenvector for eigenvalue $\lambda_k$, then $A(t) = M - (1 + r\theta(t))I = Q \cdot D \cdot Q^{-1} - (1 + r\theta(t))I = Q \cdot \widetilde{D}(t) \cdot Q^{-1}$ where matrix $\widetilde{D}(t) = D - (1 + r\theta(t))I = Diag((\lambda_k - 1 - r\theta(t))_{1\leqslant k\leqslant K})$.

We derive an analytical solution of the matrix differential Eq. (2.8) defining the *LIIS* model for all diagonalizable substitution matrix $M$ and for all initial time $t_0$. Proposition 1 gives the sequence content vector $P(t; t_0, P(t_0))$ under evolution time $t$ as a function of the eigenvalues of $M$ and the associated eigenvector matrix $Q$, the residue insertion rate vector $R$, the total residue insertion rate $r$, the initial sequence length $n_0$ at time $t_0$, the maximum sequence length $n_{\max}$ and the initial sequence content vector $P(t_0)$ at initial time $t_0$. Proposition 2 is a particular case of Proposition 1 with $t_0 = 0$. The analytical solutions of the *LIIS* model are more general and advanced than the solutions previously obtained with the *IDIS* model [13]. In particular, they involve several Gauss hypergeometric functions in addition to the classical exponential terms.

**Proposition 1.** *Given an initial time $t_0 \geqslant 0$, the sequence content vector $P(t;t_0,P(t_0))$ at time t is, for all $t \geqslant 0$,*

$$P(t;t_0,P(t_0)) = \sum_k O_k \cdot [d_1(t;k,t_0)P(t_0) + d_2(t;k,t_0)R] \tag{2.9}$$

*where $R = [r_i]_{1\leqslant i\leqslant K}$ is the vector of residue insertion rates, $P(t_0)$ is the initial sequence content vector at an initial time $t_0$, for all $1 \leqslant k \leqslant K$, matrix $O_k$ of size $K \times K$ is defined from the eigenvector matrix Q of substitution rate matrix M as follows*

$$O_k[i,j] = Q[i,k]Q^{-1}[k,j] \tag{2.10}$$

*and the two scalar terms are defined by*

$$d_1(t;k,t_0) = \left(\tau + (1-\tau)e^{-r(t-t_0)}\right)e^{-(1-\lambda_k)(t-t_0)} \tag{2.11}$$

*and, with $r = \sum_{1\leqslant i\leqslant K} r_i > 0$ and $1 \leqslant n_0 < n_{\max}$,*

$$\begin{aligned}d_2(t;k,t_0) = \frac{1}{r}\Bigg[1 - \left(\tau + (1-\tau)e^{-r(t-t_0)}\right)\Bigg(e^{-(1-\lambda_k)(t-t_0)} + \frac{1-\lambda_k}{(\tau-1)(1-\lambda_k+r)}\\ \times \left(e^{-(1-\lambda_k)(t-t_0)}{}_2\mathcal{F}_1(k,1) - e^{r(t-t_0)}{}_2\mathcal{F}_1(k,e^{r(t-t_0)})\right)\Bigg)\Bigg]\end{aligned} \tag{2.12}$$

*where $(\lambda_k)_{1\leqslant k\leqslant K}$ are the eigenvalues of matrix M, $\tau = \frac{n_0}{n_{\max}}$ where $n_0$ is the initial sequence length and $n_{\max}$ is the maximum sequence length and, $\forall 1 \leqslant k \leqslant K$ and $\forall x \geqslant 0, {}_2\mathcal{F}_1(k,x)$ is the Gauss hypergeometric function*

$${}_2\mathcal{F}_1(k,x) = H2F1\left[1, 1 + \frac{1-\lambda_k}{r}, 2 + \frac{1-\lambda_k}{r}, \frac{\tau}{\tau-1}x\right].$$

**Proof.** In order to obtain an analytical expression of Eq. (2.8), we first evaluate the term $e^{\left(\int_{t_0}^t A(u)du\right)}$ using successively the diagonalization of matrix $A(t)$, the time independence of matrix Q and the equality $e^{Q\cdot D\cdot Q^{-1}} = Q \cdot e^D \cdot Q^{-1}$,

$$\begin{aligned}e^{\left(\int_s^t A(u)du\right)} &= e^{\left(\int_s^t Q\cdot Diag\left((\lambda_k - 1 - r\theta(u))_{1\leqslant k\leqslant K}\right)\cdot Q^{-1} du\right)}\\ &= e^{Q\cdot\left(\int_s^t Diag\left((\lambda_k - 1 - r\theta(u))_{1\leqslant k\leqslant K}\right)du\right)\cdot Q^{-1}}\\ &= Q \cdot e^{\left(\int_s^t Diag\left((\lambda_k - 1 - r\theta(u))_{1\leqslant k\leqslant K}\right)du\right)} \cdot Q^{-1}\\ &= Q \cdot e^{-r\left(\int_s^t \theta(u)du\right)I + (t-s)Diag\left((\lambda_k - 1)_{1\leqslant k\leqslant K}\right)} \cdot Q^{-1}\\ &= Q \cdot Diag\left(d_1(s,t;k,t_0)_{1\leqslant k\leqslant K}\right) \cdot Q^{-1}\\ &= \left(\sum_k d_1(s,t;k,t_0)Q[i,k]Q^{-1}[k,j]\right)_{1\leqslant i,j\leqslant K}\\ &= \left(\sum_k d_1(s,t;k,t_0)O_k[i,j]\right)_{1\leqslant i,j\leqslant K}\\ &= \sum_k d_1(s,t;k,t_0)O_k\end{aligned} \tag{2.13}$$

where, after some algebraic manipulation,

$$d_1(s,t;k,t_0) = \frac{1-\tau+\tau e^{r(t-t_0)}}{1-\tau+\tau e^{r(s-t_0)}} e^{-(1-\lambda_k+r)(t-s)}.$$

The analytical expression of the first term of Eq. (2.8) is obtained from Eq. (2.13) with $s = t_0$,

$$e^{\left(\int_{t_0}^t A(u)du\right)} = \sum_k d_1(t;k,t_0)O_k$$

where $d_1(t;k,t_0) = d_1(t_0,t;k,t_0)$ is Eq. (2.11).

Using Eq. (2.13) and the time independence of vector R and matrix $O_k$, we now evaluate the second term of Eq. (2.8)

$$\begin{aligned}\int_{t_0}^t \theta(s)e^{\left(\int_s^t A(u)du\right)} \cdot Rds &= \int_{t_0}^t \theta(s)\left(\sum_k d_1(s,t;k,t_0)O_k\right) \cdot Rds\\ &= \left(\sum_k\left(\int_{t_0}^t \theta(s)d_1(s,t;k,t_0)ds\right)O_k\right) \cdot R\\ &= \left(\sum_k d_2(t;k,t_0)O_k\right) \cdot R\end{aligned}$$

where, after some algebraic manipulation, $d_2(t;k,t_0)$ is Eq. (2.12).

Finally, from Eq. (2.8), for all $t_0, t \geqslant 0$, we obtain the following analytical expression of the sequence content vector $P(t;t_0,P(t_0))$ as a function of $t_0$ and $P(t_0)$,

$$\begin{aligned}P(t;t_0,P(t_0)) &= \sum_k d_1(t;k,t_0)O_k \cdot P(t_0) + \left(\sum_k d_2(t;k,t_0)O_k\right) \cdot R\\ &= \sum_k [d_1(t;k,t_0)O_k \cdot P(t_0) + d_2(t;k,t_0)O_k \cdot R]\\ &= \sum_k O_k \cdot [d_1(t;k,t_0)P(t_0) + d_2(t;k,t_0)R]\end{aligned}$$

which is Eq. (2.9). □

**Remark 4.** The eigenvalues of stochastic matrix M satisfy $\forall 1 \leqslant k \leqslant K, 0 < \lambda_k \leqslant 1$ with one eigenvalue equal to 1 (Perron–Frobenius theorem ensures that the largest eigenvalue of a stochastic matrix is always 1). Then, the denominator $(1 - \lambda_k + r)$ in Eq. (2.12) is not null whenever the total residue insertion rate $r > 0$.

**Remark 5.** $\sum_{1\leqslant k\leqslant K} O_k = Q.Q^{-1} = I$. Indeed, using Definition (2.10) of matrix $O_k$ and for all $i,j$, $\sum_{1\leqslant k\leqslant K} O_k[i,j] = \sum_{1\leqslant k\leqslant K} Q[i,k]Q^{-1}[k,j]$ is the term in row i and column j of the matrix product $Q.Q^{-1}$. Thus, the sum of matrices $\{O_k\}_k$ is equal to the identity matrix.

**Proposition 2.** *The sequence content vector $P(t;0,P(0))$ at time t as a function of an initial sequence content vector $P(0)$ is, for all $t \geqslant 0$,*

$$P(t;0,P(0)) = \sum_k O_k \cdot [d_1(t;k,0)P(0) + d_2(t;k,0)R] \tag{2.14}$$

*where*

$$d_1(t;k,0) = \left(\tau + (1-\tau)e^{-rt}\right)e^{-(1-\lambda_k)t} \tag{2.15}$$

*and*

$$\begin{aligned}d_2(t;k,0) = \frac{1}{r}\Bigg[1 - \left(\tau + (1-\tau)e^{-rt}\right)\Bigg(e^{-(1-\lambda_k)t} + \frac{1-\lambda_k}{(\tau-1)(1-\lambda_k+r)}\\ \times \left(e^{-(1-\lambda_k)t}{}_2\mathcal{F}_1(k,1) - e^{rt}{}_2\mathcal{F}_1(k,e^{rt})\right)\Bigg)\Bigg]\end{aligned} \tag{2.16}$$

*where the parameters are defined in Proposition 1.*

**Proof.** Straightforward from Proposition 1 with $t_0 = 0$. □

Proposition 2 will be used to derive the sequence content vector as a function of the sequence length (Section 4), the analytical formulas for classical substitution matrices (Section 6) and the analytical formula of *GC* content for a practical evaluation of the *LIIS* model (Section 7).

## 3. Mathematical properties of the *LIIS* model

We set here five mathematical propositions which relate the evolution time $t$ to the values of the mutation parameters, i.e. the substitution rate matrix $M$, the insertion rate vector $R$ and the maximum sequence length $n_{\max}$. These propositions are important to model gene evolution in practice.

**Proposition 3** (*Generalization of the IDIS model for substitution and insertion*). *The IDIS model [13] is a particular case of the LIIS model when $n_{\max}$ tends to infinity. The effect of the parameter $n_{\max}$ modulating insertion during evolution of the sequence is removed when $n_{\max}$ tends to infinity. This generalization is satisfied in three ways:*

(i) *the insertion Eq. (2.1);*
(ii) *the global differential Eq. (2.7);*
(iii) *the Eq. (2.9) of the sequence content vector $P(t)$ at time t.*

**Proof**

(i) The term $\theta(t)$ satisfies

$$\lim_{n_{\max}\to+\infty}\theta(t)=1-\frac{n(t)}{n_{\max}}=1. \tag{3.1}$$

Consequently, Eq. (2.1) of the growth rate of residue $i$ becomes $n_i'(t)=r_i n(t)$ which is the growth rate of residue $i$ in the *IDIS* model (Eq. (2.5) with $d=0$ in [13]).

(ii) From Limit (3.1), the matrix differential Eq. (2.7) becomes $P'(t)=(M-(1+r)I)\cdot P(t)+R$ which is the matrix differential equation of the *IDIS* model (Eq. (2.8) with $d=0$ in [13]).

(iii) The term $\tau$ satisfies

$$\lim_{n_{\max}\to+\infty}\tau=\frac{n_0}{n_{\max}}=0.$$

Then, $\forall 1\leqslant k\leqslant K$ and $\forall x\geqslant 0$,

$$\begin{aligned}\lim_{\tau\to0}{}_2\mathcal{F}_1(k,x)&=\lim_{\tau\to0}H2F1\left[1,1+\frac{1-\lambda_k}{r},2+\frac{1-\lambda_k}{r},\frac{\tau}{\tau-1}x\right]\\&={}_2\mathcal{F}_1(k,0)=1.\end{aligned}$$

Thus, the terms $d_1(t;k,t_0)$ and $d_2(t;k,t_0)$ satisfy

$$\lim_{n_{\max}\to+\infty}d_1(t;k,t_0)=e^{-(1-\lambda_k+r)(t-t_0)}$$

and

$$\lim_{n_{\max}\to+\infty}d_2(t;k,t_0)=\frac{1-e^{-(1-\lambda_k+r)(t-t_0)}}{1-\lambda_k+r}.$$

Finally, the limit of Eq. (2.9) is

$$\begin{aligned}\lim_{n_{\max}\to+\infty}P(t;t_0,P(t_0))&=\sum_k O_k\cdot\left[e^{-(1-\lambda_k+r)(t-t_0)}P(t_0)+\frac{1-e^{-(1-\lambda_k+r)(t-t_0)}}{1-\lambda_k+r}R\right]\\&=\sum_k O_k\cdot\left[\frac{R}{1-\lambda_k+r}+\left(P(t_0)-\frac{R}{1-\lambda_k+r}\right)e^{-(1-\lambda_k+r)(t-t_0)}\right]\\&=\left(\sum_k\frac{1}{1-\lambda_k+r}O_k\right)\cdot R\\&\quad+\sum_k O_k\cdot\left(P(t_0)-\frac{1}{1-\lambda_k+r}R\right)e^{-(1-\lambda_k+r)(t-t_0)}.\end{aligned}$$

With $t_0=0$, i.e. with an initial sequence content vector $P(0)$, the sequence content $P(t;0,P(0))$ satisfies $\lim_{n_{\max}\to+\infty}P(t;0,P(0))=\left(\sum_k\frac{1}{1-\lambda_k+r}O_k\right)\cdot R+\sum_k O_k\cdot\left(P(0)-\frac{1}{1-\lambda_k+r}R\right)e^{-(1-\lambda_k+r)t}$ which is the analytical expression of the sequence content $P(t)$ at time $t$ of the *IDIS* model (Eq. (2.13) in [13]). □

**Proposition 4** (Fixed point). *The LIIS model admits a fixed point equal to the equilibrium distribution $\pi_M^{\star}$ of the stochastic substitution model defined by matrix M. This fixed point is reached by the sequence content vector after an infinite amount of time,*

$$\lim_{t\to\infty}P(t;t_0,P(t_0))=\pi_M^{\star}. \tag{3.2}$$

**Proof.** When evolution time $t$ tends to infinity, the sequence length reaches its maximum $n_{\max}$ and the insertion rate tends to 0. Then, only the substitution process is active. Indeed, when time tends to infinity, the fixed point of the sequence content vector $P(t;t_0,P(t_0))$ at time $t$ (Eq. (2.9)) simplifies as follows. From Remark 4, one eigenvalue of stochastic matrix $M$ denoted $\lambda_K$ is equal to 1 with a multiplicity equal to 1 (see also e.g. in Section 6 for the 3-parameter symmetric and *HKY* substitution matrices).

For $\lambda_K=1$, from Eq. (2.11), $d_1(t;K,t_0)=\tau+(1-\tau)e^{-r(t-t_0)}$ and thus,

$$\lim_{t\to\infty}d_1(t;K,t_0)=\tau$$

and from Eq. (2.12), $d_2(t;K,t_0)=\frac{1-\tau-(1-\tau)e^{-r(t-t_0)}}{r}$ and thus,

$$\lim_{t\to\infty}d_2(t;K,t_0)=\frac{1-\tau}{r}.$$

For all other eigenvalues, i.e. for all $k\neq K$, then $\lambda_k<1$ and from Eq. (2.11),

$$\lim_{t\to\infty}d_1(t;k,t_0)=0$$

as for all $\lambda_k<1,\lim_{t\to\infty}e^{-(1-\lambda_k+r)(t-t_0)}=0$ and from Eq. (2.12),

$$\lim_{t\to\infty}d_2(t;k,t_0)=0$$

as for all $\lambda_k<1$, $\lim_{t\to\infty}e^{-(1-\lambda_k)(t-t_0)}=0$ and $\lim_{t\to\infty}-e^{r(t-t_0)}\,{}_2\mathcal{F}_1(k,e^{r(t-t_0)})=\frac{(\tau-1)(1-\lambda_k+r)}{\tau(1-\lambda_k)}$.

Then, the limit of the sequence content vector $P(t;t_0,P(t_0))$ defined in Eq. (2.9) when time $t$ tends to infinity satisfies

$$\begin{aligned}\lim_{t\to\infty}P(t;t_0,P(t_0))&=O_K\cdot\left[\lim_{t\to\infty}d_1(t;K,t_0)P(t_0)+\lim_{t\to\infty}d_2(t;K,t_0)R\right]\\&=O_K\cdot\left(\tau P(t_0)+(1-\tau)\frac{R}{r}\right)\\&=\tau O_K\cdot P(t_0)+(1-\tau)O_K\cdot\frac{R}{r}.\end{aligned} \tag{3.3}$$

The eigenvector associated to $\lambda_K=1$ is the equilibrium distribution $\pi_M^{\star}$ of the substitution model. The columns of matrix $O_K$ are all equal to $\pi_M^{\star}$. For example, for the classical substitution matrices ($K=4$), $O_4=\left(\frac{1}{4}\right)_{1\leqslant i,j\leqslant K}$ for the 3-parameter and each column of matrix $O_4$ is $(\pi_A,\pi_C,\pi_G,\pi_T)$ for the *HKY* matrix (see also Section 6). Then, in Eq. (3.3), vectors $P(t_0)$ and $\frac{R}{r}$ sum to 1, then $O_K\cdot P(t_0)=\pi_M^{\star}$ and $O_K\cdot\frac{R}{r}=\pi_M^{\star}$ leading to $\lim_{t\to\infty}P(t;t_0,P(t_0))=\pi_M^{\star}$. □

**Proposition 5** (*Time scale*). *When multiplying all the substitution-insertion parameters, i.e. the non-diagonal elements $[m_{ij}]_{i\neq j}$ of the substitution rate matrix M and the insertion rates $[r_i]$, by a scalar $\alpha$, the sequence content vector $P(t;t_0,P(t_0))$ at time t given an initial time $t_0$ (Eq. (2.9)) is equal to the sequence content vector obtained at time $\alpha t$ and at initial time $\alpha t_0$ with the substitution-insertion parameters $([m_{ij}]_{i\neq j},[r_i])$*

$$P(t;t_0,P(t_0);[\alpha m_{ij}]_{i\neq j},[\alpha r_i])=P(\alpha t;\alpha t_0,P(\alpha t_0);[m_{ij}]_{i\neq j},[r_i]). \tag{3.4}$$

**Proof.** The multiplication of the *LIIS* model parameters by a scalar $\alpha$ leads to residue insertion rate vector $\widetilde{R}=\alpha R$, thus total insertion rate $\widetilde{r}=\alpha r$, and substitution rate matrix $\widetilde{M}=\alpha M+(1-\alpha)I$. The substitution rate matrix $M$ decomposes as $M=Q\cdot D\cdot Q^{-1}$ where

$D = Diag((\lambda_k)_{1\leqslant k\leqslant K})$ is the eigenvalue diagonal matrix and $Q$ is an associated eigenvector matrix, the $k$th column of $Q$ being an eigenvector of eigenvalue $\lambda_k$. Then, matrix $\widetilde{M}$ decomposes as $\widetilde{M} = \alpha Q \cdot D \cdot Q^{-1} + (1-\alpha)I = Q \cdot \widetilde{D} \cdot Q^{-1}$ where matrix $\widetilde{D} = \alpha D + (1-\alpha)I = Diag((\widetilde{\lambda}_k)_{1\leqslant k\leqslant K})$. Matrix $\widetilde{M}$ can be diagonalized with real eigenvalues $(\widetilde{\lambda}_k)_{1\leqslant k\leqslant K}$ where $\widetilde{\lambda}_k = 1 + \alpha(\lambda_k - 1)$ for all $1 \leqslant k \leqslant K$. Then, $1 - \widetilde{\lambda}_k + \widetilde{r} = \alpha(1 - \lambda_k + r)$. Matrix $Q$, and consequently the matrices $O_k$, remain unchanged. Then, for all $1 \leqslant k \leqslant K$, $\left(1-\widetilde{\lambda}_k\right) = \alpha(1-\lambda_k)$ and $\left(1 - \widetilde{\lambda}_k + \widetilde{r}\right) = \alpha(1-\lambda_k+r)$. Finally, from Eq. (2.11),

$$\begin{aligned} d_1(t;k,t_0;[\alpha m_{ij}]_{i\neq j},[\alpha r_i]) &= (1-\tau+\tau e^{\widetilde{r}(t-t_0)})e^{-\left(1-\widetilde{\lambda}_k+\widetilde{r}\right)(t-t_0)} \\ &= (1-\tau+\tau e^{\alpha r(t-t_0)})e^{-\alpha(1-\lambda_k+r)(t-t_0)} \\ &= d_1(\alpha t;k,\alpha t_0;[m_{ij}]_{i\neq j},[r_i]) \end{aligned}$$

and from Eq. (2.12),

$$\begin{aligned} d_2(t;k,t_0;[\alpha m_{ij}]_{i\neq j},[\alpha r_i]) =& \frac{1}{\widetilde{r}}\Bigg[1-\left(\tau+(1-\tau)e^{-\widetilde{r}(t-t_0)}\right) \\ &\times\left(e^{-\left(1-\widetilde{\lambda}_k\right)(t-t_0)}+\frac{1-\widetilde{\lambda}_k}{(\tau-1)\left(1-\widetilde{\lambda}_k+\widetilde{r}\right)}\right. \\ &\times\left(e^{-\left(1-\widetilde{\lambda}_k\right)(t-t_0)}{}_2\mathcal{F}_1(k,1)\right. \\ &\left.\left.-e^{\widetilde{r}(t-t_0)}{}_2\mathcal{F}_1(k,e^{\widetilde{r}(t-t_0)})\right)\right)\Bigg] \\ =& \frac{1}{\alpha r}[1-(\tau+(1-\tau)e^{-\alpha r(t-t_0)}) \\ &\times\left(e^{-\alpha(1-\lambda_k)(t-t_0)}+\frac{\alpha(1-\lambda_k)}{(\tau-1)\alpha(1-\lambda_k+r)}\right. \\ &\times(e^{-\alpha(1-\lambda_k)(t-t_0)}{}_2\mathcal{F}_1(k,1) \\ &-e^{\alpha r(t-t_0)}{}_2\mathcal{F}_1(k,e^{\alpha r(t-t_0)})))] \\ =& \frac{1}{\alpha}d_2(\alpha t;k,\alpha t_0;[m_{ij}]_{i\neq j},[r_i]). \end{aligned}$$

Consequently, in Eq. (2.9),

$$\begin{aligned} P(t;t_0,P(t_0);[\alpha m_{ij}]_{i\neq j},[\alpha r_i]) =& \sum_k O_k \cdot \Big[d_1(\alpha t;k,\alpha t_0;[m_{ij}]_{i\neq j},[r_i])P(t_0) \\ &+\left(\frac{1}{\alpha}d_2(\alpha t;k,\alpha t_0;[m_{ij}]_{i\neq j},[r_i])\right)\alpha R\Big] \\ =& P(\alpha t;\alpha t_0,P(\alpha t_0);[m_{ij}]_{i\neq j},[r_i]). \quad \square \end{aligned}$$

**Proposition 6** (*Time step*). *For all $t_0, t_1, t_2 > 0$, the sequence content vector satisfies*

$$P(t_2) = P(t_2;t_0,P(t_0)) = P(t_2;t_1,P(t_1)) = P(t_2;t_1,P(t_1;t_0,P(t_0))). \tag{3.5}$$

**Proof.** Straightforward from Eq. (2.8) with $P(t_1) = P(t_1;t_0, P(t_0))$. □

**Proposition 7** (*Time inversion*). *For all $t_0, t > 0$ with $t_0 < t$,*

$$\begin{aligned} P(t_0;t,P(t)) &= \sum_k O_k \cdot [d_1(t_0;k,t)P(t) + d_2(t_0;k,t)R] \\ &= \sum_k O_k \cdot [d_1(-t;k,-t_0)P(t) + d_2(-t;k,-t_0)R]. \end{aligned} \tag{3.6}$$

**Proof.** First line is straightforward from Eq. (2.9). Then, we have directly $d_1(t_0;k,t) = d_1(-t;k,-t_0)$ from Eq. (2.11) and $d_2(t_0;k,t) = d_2(-t;k,-t_0)$ from Eq. (2.12). □

**Remark 6.** Time inversion can be simply derived by replacing $P(t_0)$ by $P(t)$ and $(t - t_0)$ by $(t_0 - t)$ in all analytical formulas.

## 4. Time and sequence length

We derive here the *LIIS* model for sequence length where the sequence content vector $P(l;n_0,P(n_0))$ is expressed as a function of the sequence length $l$ observed at evolution time $t$ ($l = n(t)$).

**Proposition 8** (*Time and sequence length*). *Given an initial sequence length $n_0 = n(0) \geqslant 1$ observed at initial time $t_0$ chosen equal to 0 for convenience, the sequence content vector $P(l;n_0,P(n_0))$ at sequence length $l = n(t)$ at time $t$ is, for all $l \geqslant 1$,*

$$P(l;n_0,P(n_0)) = \sum_k O_k \cdot [d_1(l;k,n_0)P(n_0) + d_2(l;k,n_0)R] \tag{4.1}$$

*where matrices $(O_k)_{1\leqslant k\leqslant K}$ are defined in (2.10) from the eigenvector matrix $Q$ of substitution rate matrix $M$, $R = [r_i]_{1\leqslant i\leqslant K}$ is the vector of the residue insertion rates per site, and $P(n_0)$ is the initial sequence content vector at length $n_0$ and*

$$d_1(l;k,n_0) = \frac{n_0}{l}h(l)^{-\frac{1-\lambda_k}{r}} \tag{4.2}$$

$$\begin{aligned} d_2(l;k,n_0) = \frac{1}{r}\Bigg[1-\frac{n_0}{l}\Bigg(h(l)^{-\frac{1-\lambda_k}{r}}+\frac{1-\lambda_k}{(\tau-1)(1-\lambda_k+r)} \\ \times\left(h(l)^{-\frac{1-\lambda_k}{r}}{}_2\mathcal{F}_1(k,1)-h(l){}_2\mathcal{F}_1(k,h(l))\right)\Bigg)\Bigg] \end{aligned} \tag{4.3}$$

where $(\lambda_k)_{1\leqslant k\leqslant K}$ *are the eigenvalues of $M$, $r = \sum_{1\leqslant i\leqslant K} r_i$ is the total residue insertion rate, $\tau = \frac{n_0}{n_{\max}}$ where $n_{\max}$ is the maximum sequence length and $\forall x \geqslant 0, \forall 1 \leqslant k \leqslant K, {}_2\mathcal{F}_1(k,x) = H2F1\left[1, 1+\frac{1-\lambda_k}{r}, 2+\frac{1-\lambda_k}{r}, \frac{\tau}{\tau-1}x\right]$ is the Gauss hypergeometric function and*

$$h(l) = \frac{\tau-1}{\tau-\frac{n_0}{l}} = e^{rt}. \tag{4.4}$$

**Proof.** $P(l;n_0,P(n_0))$ is obtained from Eq. (2.14). From Eq. (2.3), the evolution time $t$ between the sequence lengths $n(0) = n_0$ at time $t_0 = 0$ and $n(t) = l$ at time $t$ is $l = \frac{n_0}{\tau+(1-\tau)e^{-rt}}$ leading to Eq. (4.4). Then, $d_1(l;k,n_0)$ and $d_2(l;k,n_0)$ are obtained from Eq. (2.15) and (2.16), respectively. Note that the denominator in Eq. (4.3) is not null according to both Remark 4 and the conditions of Eq. (2.12). □

**Remark 7.** Eq. (4.1) is true for the very particular case $l = n_0$. Indeed, for all $1 \leqslant k \leqslant K$, $d_1(l;k,n_0) = 1$ and $d_2(l;k,n_0) = 0$. Then, from Eq. (4.1), $P(l;n_0,P(n_0)) = \sum_k O_k \cdot P(n_0) = P(n_0)$ as $\sum_k O_k = I$ (see Remark 5).

**Remark 8.** The *IDIS* model as a function of the sequence length [14] is a particular case of the *LIIS* model when $n_{\max}$ tends to infinity. We have the following limit: as $\tau = \frac{n_0}{n_{\max}}$, then

$$\lim_{n_{\max}\to\infty} \tau = 0.$$

As

$$\lim_{\tau\to 0} h(l) = \frac{l}{n_0}$$

and $\forall 1 \leqslant k \leqslant K, \forall x \geqslant 0$,

$$\begin{aligned} \lim_{\tau\to 0} {}_2\mathcal{F}_1(k,x) &= \lim_{\tau\to 0} H2F1\left[1, 1+\frac{1-\lambda_k}{r}, 2+\frac{1-\lambda_k}{r}, \frac{\tau}{\tau-1}x\right] \\ &= {}_2\mathcal{F}_1(k,0) = 1 \end{aligned}$$

then,

$$\lim_{\tau\to 0} \ d_1(l;k,n_0) = \left(\frac{l}{n_0}\right)^{-\left(1+\frac{1-\lambda_k}{r}\right)}$$

$$\lim_{\tau\to 0} \ d_2(l;k,n_0) = \frac{1}{1-\lambda_k+r}\left(1-\left(\frac{l}{n_0}\right)^{-\left(1+\frac{1-\lambda_k}{r}\right)}\right).$$

Thus, from Eq. (4.1),

$$\begin{aligned}\lim_{n_{\max}\to\infty} P(l;n_0,P(n_0)) &= \sum_k O_k\cdot\left[\left(\frac{l}{n_0}\right)^{-\left(1+\frac{1-\lambda_k}{r}\right)} P(n_0)+\frac{1}{1-\lambda_k+r}\right.\\ &\quad\left.\times\left(1-\left(\frac{l}{n_0}\right)^{-\left(1+\frac{1-\lambda_k}{r}\right)}\right)R\right]\\ &= \left(\sum_k \frac{1}{1-\lambda_k+r}O_k\right)\cdot R\\ &\quad+\sum_k O_k\cdot\left(P(n_0)-\frac{1}{1-\lambda_k+r}R\right)\left(\frac{l}{n_0}\right)^{-\left(1+\frac{1-\lambda_k}{r}\right)}\end{aligned}$$

which is the *IDIS* model as a function of the sequence length (Eq. (10) with $d = 0$ in [14]).

## 5. Insertion only model

**Proposition 9** (*Insertion only – evolution time*). *In the insertion only model, i.e. the substitution rates are all equal to 0 ($M = I$), and given an initial time $t_0 \geqslant 0$, then the sequence content vector $P(t;t_0,P(t_0))$ at time t is*

$$P(t;t_0,P(t_0)) = \tau P(t_0) + (1-\tau)\left[\frac{R}{r}+\left(P(t_0)-\frac{R}{r}\right)e^{-r(t-t_0)}\right] \quad (5.1)$$

*and given an initial time $t_0 = 0, P(t;t_0,P(0))$ at time t is*

$$P(t;0,P(0)) = \tau P(0) + (1-\tau)\left[\frac{R}{r}+\left(P(0)-\frac{R}{r}\right)e^{-rt}\right]. \quad (5.2)$$

**Proof.** In the insertion only model, substitutions are not allowed, i.e. $M = I$. Then, all the eigenvalues are equal to 1. Thus, for all $1 \leqslant k \leqslant K$, $\lambda_k = 1$ and, from Eq. (2.11),

$$d_1(t;k,t_0) = \tau + (1-\tau)e^{-r(t-t_0)}$$

and from Eq. (2.12), given that $\forall x \geqslant 0, H2F1\left[1,1,2,\frac{\tau}{\tau-1}x\right] = \frac{(1-\tau)\ln\left(1+\frac{\tau x}{1-\tau}\right)}{\tau x}$,

$$d_2(t;k,t_0) = \frac{1}{r}(1-\tau)\left(1-e^{-r(t-t_0)}\right).$$

Using Eq. (2.9) and $\sum_k O_k = I$ (Remark 5), then $P(t;t_0,P(t_0))$ is equal to

$$\begin{aligned}P(t;t_0,P(t_0)) &= \sum_k O_k\cdot\left[\left(\tau+(1-\tau)e^{-r(t-t_0)}\right)P(t_0)+(1-\tau)\left(1-e^{-r(t-t_0)}\right)\frac{R}{r}\right]\\ &= \left(\sum_k O_k\right)\left(\tau P(t_0)+(1-\tau)\frac{R}{r}+(1-\tau)\left(P(t_0)-\frac{R}{r}\right)e^{-r(t-t_0)}\right)\\ &= \tau P(t_0)+(1-\tau)\frac{R}{r}+(1-\tau)\left(P(t_0)-\frac{R}{r}\right)e^{-r(t-t_0)}.\end{aligned}$$

Eq. (5.2) is obtained straightforward from Eq. (5.1) with $t_0 = 0$. □

**Proposition 10** (*Insertion only – sequence length*). *In the insertion only model, the sequence content vector $P(l;n_0,P(n_0))$ at sequence length l is, for all $0 \leqslant l \leqslant n_{max}$,*

$$P(l;n_0,P(n_0)) = \frac{R}{r}+\left(P(l_0)-\frac{R}{r}\right)\frac{n_0}{l}. \quad (5.3)$$

**Proof.** Straightforward from Eqs. (5.2) and (4.4),

$$\begin{aligned}P(l;n_0,P(n_0)) &= \tau P(n_0)+(1-\tau)\left[\frac{R}{r}+\left(P(n_0)-\frac{R}{r}\right)h(l)^{-1}\right]\\ &= \frac{R}{r}+\left(P(n_0)-\frac{R}{r}\right)\frac{n_0}{l}. \qquad \square\end{aligned}$$

**Remark 9.** The sequence content vector $P(l;n_0,P(n_0))$ at sequence length $l$ is independent of the maximum sequence length $n_{\max}$ and from the ratio $\tau = \frac{n_0}{n_{\max}}$. This a priori surprising observation is explained by the fact that $P(l;n_0,P(n_0))$ is equal to the sequence content vector obtained with the non-limited insertion model *IDIS* at the same length $l$ (Eq. (18) with $d = 0$ in [14]). In the *LIIS* model, the growth rate $n_i'(t)$ decreases in time due to parameter $n_{\max}$ (Eq. 2.1). However, from a sequence length point of view, the sequence content for a given length is the same as the one obtained with a non-limited insertion process. The only difference between the two models is that the sequence spends more time in each length when insertion is limited, i.e. in the *LIIS* model.

## 6. Analytical formulas for classical substitution matrices

### 6.1. LIIS-SYM3 analytical formula

The *LIIS-SYM*3 model gives the expression of nucleotide sequence content vector $P(t) = P(t;0,P(0))$ by deriving Eq. (2.14) with the classical 3-parameter symmetric substitution matrix $M_{SYM3}$ [12]. This matrix $M_{SYM3}$ 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 $M_{SYM3}$ is defined as follows

$$M_{SYM3} = \begin{pmatrix} n & c & a & b \\ c & n & b & a \\ a & b & n & c \\ b & a & c & n \end{pmatrix}$$

where $n = 1-(a+b+c)$. The four eigenvalues of matrix $M_{SYM3}$ are

$$\{\lambda_1 = 1-2(a+b), \quad \lambda_2 = 1-2(a+c), \quad \lambda_3 = 1-2(b+c), \quad \lambda_4 = 1\} \quad (6.1)$$

and their associated eigenvectors are

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

After some algebraic manipulation of Eq. (2.14), we obtain the sequence content vector $P(x)$ in each nucleotide $A, C, G$ and $T$ as a function of a variable $x$ representing either evolution time $x = t$ or sequence length $x = l$. Then, the variable $x_0$ represents the initial condition of $x$ which is $x_0 = 0$ for evolution time $t$ and $x_0 = n_0$ for sequence length $l$. Finally, a function $h(x)$ is introduced which is equal to $h(x) = e^{rt}$ for evolution time $t$ and to $h(x) = h(l) = \frac{\tau-1}{\tau-\frac{n_0}{l}}$ for sequence length $l$. Thus, with the following convention $(x, x_0, h(x)) = (t, 0, e^{rt})$ for an evolution time process and $(x, x_0, h(x)) = (l, n_0, h(l))$ for a sequence length process, the sequence content vector reads

$$P(x) = \frac{1}{r}\begin{pmatrix} r_A \\ r_C \\ r_G \\ r_T \end{pmatrix} + \frac{1}{4}(1-\tau+\tau h(x))\begin{pmatrix} f_1(x)+f_2(x)+f_3(x) \\ f_1(x)-f_2(x)-f_3(x) \\ -f_1(x)-f_2(x)+f_3(x) \\ -f_1(x)+f_2(x)-f_3(x) \end{pmatrix} \quad (6.2)$$

where $r = r_A + r_C + r_G + r_T, \tau = \frac{n_0}{n_{\max}}$ and, for all $k = 1,2,3$,

$$f_k(x) = \left(p_k(x_0) - \frac{r_k}{r}\right)h(x)^{-\frac{1-\lambda_k+r}{r}} + \left(\frac{1}{1-\tau}\right)\left(\frac{r_k}{r}\right)\left(\frac{1-\lambda_k}{1-\lambda_k+r}\right)\left[h(x)^{-\frac{1-\lambda_k+r}{r}}{}_2F_1(k,1) - {}_2F_1(k,h(x))\right]$$

with, for $y = 1, h(x),\ {}_2F_1(k,y) = H2F1\left[1, 1+\frac{1-\lambda_k}{r}, 2+\frac{1-\lambda_k}{r}, \frac{\tau}{\tau-1}y\right]$, $\lambda_k$ defined in (6.1), and $r_1 = r_A + r_C - r_G - r_T,\ r_2 = r_A - r_C - r_G + r_T,\ r_3 = r_A - r_C + r_G - r_T;\quad p_1(x_0) = p_A(x_0) + p_C(x_0) - p_G(x_0) - p_T(x_0)$, $p_2(x_0) = p_A(x_0) - p_C(x_0) - p_G(x_0) + p_T(x_0)$ and $p_3(x_0) = p_A(x_0) - p_C(x_0) + p_G(x_0) - p_T(x_0)$.

### 6.2. LIIS-HKY analytical formula

The *LIIS-HKY* model gives the expression of nucleotide sequence content vector $P(t) = P(t;0,P(0))$ by deriving Eq. (2.14) with the classical substitution matrix $M_{HKY}$ [9]. This matrix $M_{HKY}$ is defined by six formal parameters: the transition and transversion rates, $\alpha$ and $\beta$, respectively, and the equilibrium nucleotide frequencies $\pi_A, \pi_C, \pi_G$ and $\pi_T$,

$$M_{HKY} = \begin{pmatrix} n_A & \beta\pi_A & \alpha\pi_A & \beta\pi_A \\ \beta\pi_C & n_C & \beta\pi_C & \alpha\pi_C \\ \alpha\pi_G & \beta\pi_G & n_G & \beta\pi_G \\ \beta\pi_T & \alpha\pi_T & \beta\pi_T & n_T \end{pmatrix}$$

where for all $j$ in $\{A,C,G,T\}, n_j = 1 - \Sigma_{i\neq j}M_{HKY}[i,j]$ such that matrix $M_{HKY}$ is stochastic in column. This matrix $M_{HKY}$ defines one of the most general substitution models whose equilibrium distribution differs from $1/4$ for all nucleotides. Let us denote by $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$ the equilibrium frequencies for purine $(A,G)$ and pyrimidine $(C,T)$, the four eigenvalues of matrix $M_{HKY}$ are

$$\{\lambda_1 = 1-\beta,\ \lambda_2 = 1 - \alpha\pi_R - \beta\pi_Y,\ \lambda_3 = 1 - \alpha\pi_Y - \beta\pi_R,\ \lambda_4 = 1\} \tag{6.3}$$

and their associated eigenvectors are

$$\left\{v_1 = \left\{-\frac{\pi_Y\pi_A}{\pi_R\pi_T}, \frac{\pi_C}{\pi_T}, -\frac{\pi_Y\pi_G}{\pi_R\pi_T}, 1\right\},\ v_2 = \{-1,0,1,0\},\ v_3 = \{0,-1,0,1\},\ v_4 = \left\{\frac{\pi_A}{\pi_T}, \frac{\pi_C}{\pi_T}, \frac{\pi_G}{\pi_T}, 1\right\}\right\}.$$

After some algebraic manipulation of Eq. (2.14), we obtain the sequence content vector $P(x)$ in each nucleotide $A, C, G$ and $T$ with the following convention $(x, x_0, h(x)) = (t, 0, h(t) = e^{rt})$ for an evolution time process and $(x, x_0, h(x)) = \left(l, n_0, h(l) = \frac{\tau-1}{\tau-\frac{n_0}{l}}\right)$ for a sequence length process then,

$$P(x) = \frac{1}{r}\left[\begin{pmatrix} r_A \\ r_C \\ r_G \\ r_T \end{pmatrix} + (1-\tau+\tau h(x))\begin{pmatrix} \frac{-\pi_A g_1(x) - g_2(x)}{\pi_R} \\ \frac{\pi_C g_1(x) - g_3(x)}{\pi_Y} \\ \frac{-\pi_G g_1(x) + g_2(x)}{\pi_R} \\ \frac{\pi_T g_1(x) + g_3(x)}{\pi_Y} \end{pmatrix}\right] \tag{6.4}$$

where $r = r_A + r_C + r_G + r_T, \tau = \frac{n_0}{n_{\max}}$ and, for all $k = 1,2,3$,

$$g_k(x) = (C_k - D_k r)h(x)^{-\frac{1-\lambda_k+r}{r}} - \frac{C_k(1-\lambda_k)}{(1-\tau)(1-\lambda_k+r)}\left[h(x)^{-\frac{1-\lambda_k+r}{r}}{}_2F_1(k,1) - {}_2F_1(k,h(x))\right] \tag{6.5}$$

with for $y = 1, h(x),\ {}_2F_1(k,y) = H2F1\left[1, 1+\frac{1-\lambda_k}{r}, 2+\frac{1-\lambda_k}{r}, \frac{\tau}{\tau-1}y\right]$, $\lambda_k$ defined in (6.3), and $C_1 = \pi_Y r_R - \pi_R r_Y,\ C_2 = \pi_G r_A - \pi_A r_G,\ C_3 = \pi_T r_C - \pi_C r_T,\quad D_1 = \pi_Y p_R(x_0) - \pi_R p_Y(x_0),\quad D_2 = \pi_G p_A(x_0) - \pi_A p_G(x_0)$, $D_3 = \pi_T p_C(x_0) - \pi_C p_T(x_0),\ r_R = r_A + r_G,\ r_Y = r_C + r_T,\ p_R(x_0) = p_A(x_0) + p_G(x_0)$ and $p_Y(x_0) = p_C(x_0) + p_T(x_0)$.

The *LIIS-HKY* model (Eq. 6.4) as a function of the sequence length (with $x = l$) is used in Section 7 for modelling the *GC* content in complete genomes of four prokaryotic taxonomic groups.

## 7. A statistical evaluation of the *LIIS* model with a *GC* content analysis in complete genomes

The *LIIS* model is a generalization of the *IDIS* model [14] where an additional formal parameter $n_{\max}$ modulates the insertion process according to the sequence length. Indeed, Proposition 3 proves that the *IDIS* model is a particular case of the *LIIS* model when the maximum sequence length $n_{\max}$ tends to infinity. In this section, we will show that this theoretical generalization has direct consequences in biological applications. The *IDIS* model was applied to the analysis of the *GC* content in bacterial genomes [14]. The analysis of the *GC* content has been a matter of debate for several years as no mathematical model has been proposed to describe the relationship observed between the *GC* content, the genome length [22,27,3,37,23,21,28] and the mutation events [8,29,37] in bacterial genomes. The *IDIS* model outperforms the most recent modelling of *GC* content which is based on an empirical linear relationship [37,23,21] in bacterial genomes (see the coefficients $R^2$ in Fig. 4 in [14]). From a theoretical point of view, this result was explained, in particular, by the two following facts: (i) the linear model is a particular case of the non-linear *IDIS* model with one more degree of freedom (the parameter $c = -1$ of the *IDIS* model being associated to the linear case); and (ii) the *IDIS* model relies on evolution assumptions for the processes of substitution, insertion and deletion, in contrast to an empirical relationship. A statistical analysis of the *GC* content in complete genomes of four prokaryotic taxonomic groups, namely Chlamydiae, Crenarchaeota, Spirochaetes and Thermotogae, will show the expected improvement from the theory of the *LIIS* model compared to the *IDIS* model.

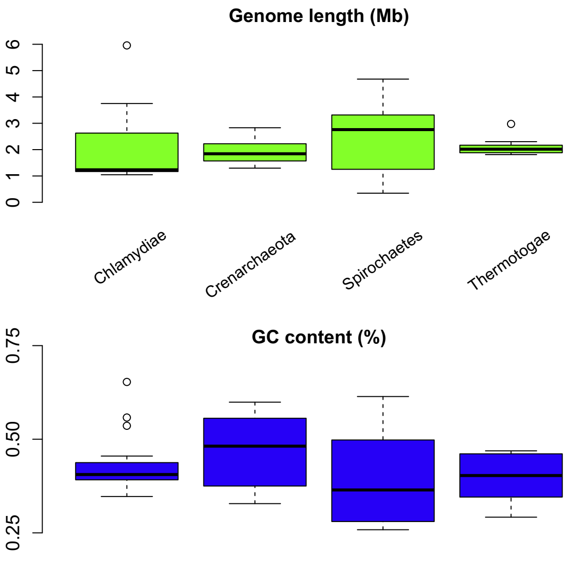

**Fig. 1.** Statistical features of the genome length (Mb) and *GC* content of complete genomes of four taxonomic groups Chlamydiae, Crenarchaeota, Spirochaetes and Thermotogae. The boxplots show for each taxonomic group the distribution of the genome length (top boxplot) and *GC* content (bottom boxplot). The horizontal bar shows the median, the box margins represent the 25th and 75th percentiles, the whiskers indicate data within 2 times the interquartile range and the circles are outliers.

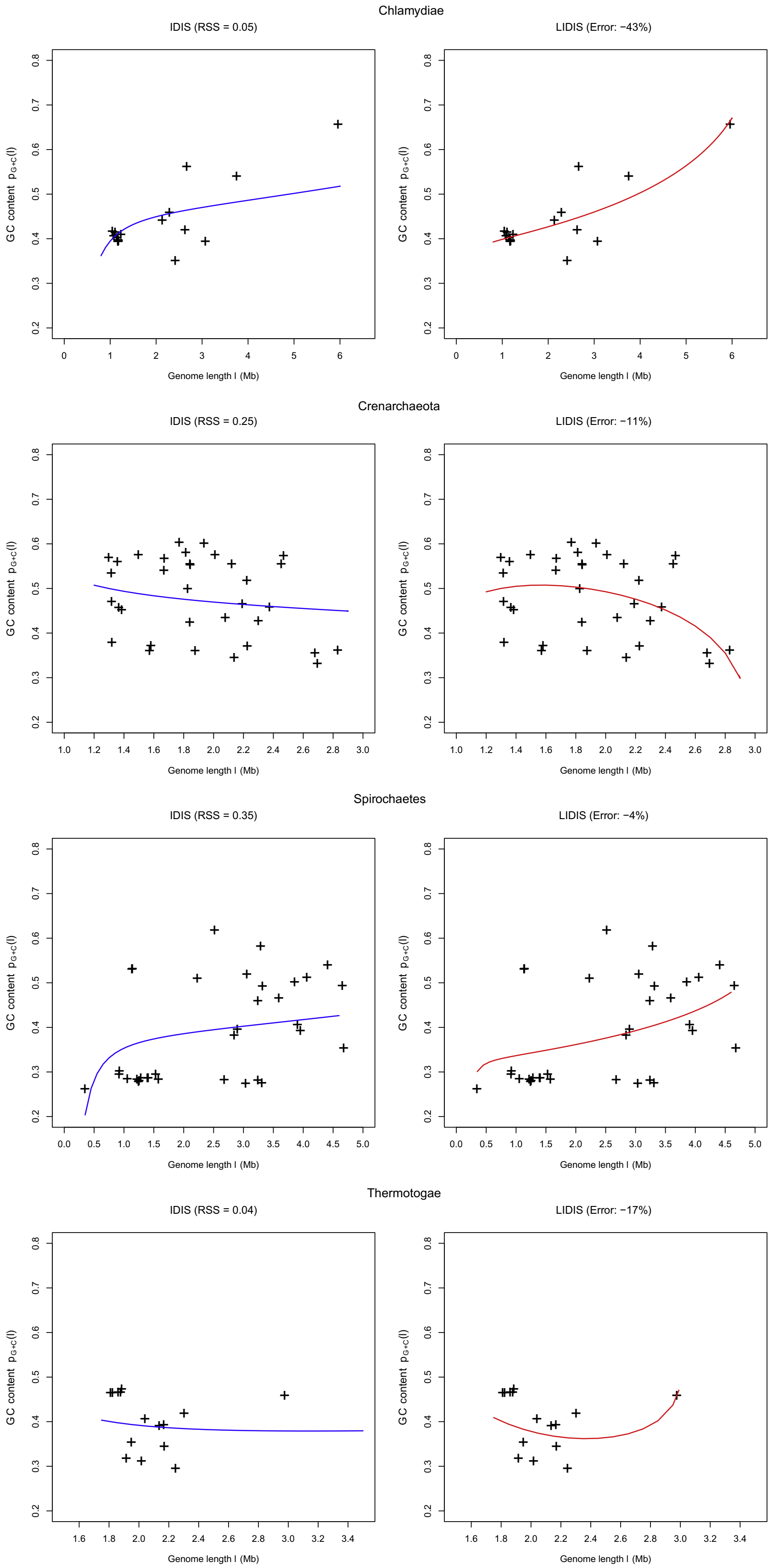

**Fig. 2.** Best fit curves $p_{GC}(l)$ minimizing the error *RSS* (Residual Sum of Squares) for the *IDIS* model (left panel) and the novel *LIIS* model (right panel) of complete genomes of four taxonomic groups Chlamydiae, Crenarchaeota, Spirochaetes and Thermotogae. In each plot, the *x*-axis represents the genome length *l* (Mb) and the *y*-axis, the *GC* content $p_{GC}(l)$. For each taxonomic group, the left panel shows the best fit curve $p_{GC}(l)$ obtained with the *IDIS-HKY* model and its associated error *RSS* and the right panel shows the best fit curves $p_{GC}(l)$ obtained with the *LIIS-HKY* model and its associated error decrease (%) in comparison with the *IDIS-HKY* model.

### 7.1. GC content formula

The analysis of *GC* content in complete genomes leads to the following assumptions with the *LIIS-HKY* model. As the DNA double helix is antiparallel and complementary (*A* bonds *T* and *C* bonds *G*), the number of *C* is equal to the number of *G*, and similarly for *A* and *T*. Thus, the initial sequence content, the equilibrium distribution and the nucleotide insertion rates satisfy the following conditions

$$\begin{cases} p_C(n_0) = p_G(n_0), \quad p_A(n_0) = p_T(n_0), \quad p_R(n_0) = p_Y(n_0) = \frac{1}{2} \\ \pi_C = \pi_G, \quad \pi_A = \pi_T, \quad \pi_R = \pi_Y = \frac{1}{2} \\ r_C = r_G, \quad r_A = r_T, \quad r_R = r_Y = \frac{r}{2}. \end{cases} \tag{7.1}$$

The *GC* content $p_{GC}(l) = p_G(l) + p_C(l)$, i.e. the sum of the sequence content in nucleotides *C* and *G*, is after some algebraic manipulation of Eq. (6.4) with $x = l$, equal to

$$\begin{aligned} p_{GC}(l) = {} & 2\frac{r_C}{r} + 2(1-\tau+\tau h(l))\Big[\Big(p_C(n_0) - \frac{r_C}{r}\Big)h(l)^{-\left(1+\frac{\alpha+\beta}{2r}\right)} \\ & - \frac{\pi_C - \frac{r_C}{r}}{(1-\tau)\left(1+\frac{2r}{\alpha+\beta}\right)}\Big(h(l)^{-\left(1+\frac{\alpha+\beta}{2r}\right)}{}_2F_1(2,1) - {}_2F_1(2,h(l))\Big)\Big] \end{aligned} \tag{7.2}$$

where $r = 2(r_A + r_C), \tau = \frac{n_0}{n_{\max}}, h(l) = \frac{\tau-1}{\tau-\frac{n_0}{l}}$ and, for $x = 1, h(l)$, ${}_2F_1(2,x) = H2F1\left[1, 1+\frac{1-\lambda_2}{r}, 2+\frac{1-\lambda_2}{r}, \frac{\tau}{\tau-1}x\right]$ with $\lambda_2 = 1 - \frac{\alpha+\beta}{2}$.

**Proof.** From Eq. (6.4) with $x = l$,

$$\begin{aligned} p_{GC}(l) &= p_G(l) + p_C(l) \\ &= \frac{r_C + r_G}{r} + \frac{2}{r}(1-\tau+\tau h(l))[\pi_C g_1(l) - g_3(l) - \pi_G g_1(l) + g_2(l)] \\ &= \frac{2r_C}{r} + \frac{2}{r}(1-\tau+\tau h(l))[\pi_C g_1(l) - g_3(l) - \pi_C g_1(l) + g_2(l)] \\ &= \frac{2r_C}{r} + \frac{2}{r}(1-\tau+\tau h(l))[g_2(l) - g_3(l)] \\ &= \frac{2}{r}[r_C + (1-\tau+\tau h(l))(g_2(l) - g_3(l))] \\ &= \frac{2}{r}[r_C + 2(1-\tau+\tau h(l))g_2(l)]. \end{aligned}$$

Indeed, from Eq. (6.5) and Condition (7.1), the following relations between the terms in $k = 2$ and $k = 3$ are obtained: $\lambda_2 = \lambda_3, C_3 = \pi_T r_C - \pi_C r_T = \pi_A r_C - \pi_C r_A = -C_2$ and $D_3 = -D_2$ lead to $g_3(l) = -g_2(l)$. From Condition (7.1), the following relations are deduced: $p_A(n_0) = \frac{1}{2} - p_C(n_0), \pi_A = \frac{1}{2} - \pi_C$ and $r_A = \frac{r}{2} - r_C$. They lead to $C_2 = \frac{1}{2}(\pi_C r - r_C)$ and $D_2 = \frac{1}{2}(\pi_C - p_C(n_0))$ and finally to $C_2 - D_2 r = \frac{1}{2}(r p_C(n_0) - r_C)$ which allows the analytical Eq. (7.2) to be retrieved. □

### 7.2. GC content estimation

Complete genomes of four prokaryotic taxonomic groups, namely Chlamydiae (17 genomes), Crenarchaeota (34 genomes), Spirochaetes (34 genomes) and Thermotogae (15 genomes), are chosen as an example in order to show the improvement of the *LIIS* model (precisely, *LIIS-HKY*) compared to the previous *IDIS* model (precisely, *IDIS-HKY*; [14]) for modelling the *GC* content according to the genome length. The genome length (Mb) and *GC* content of complete genomes of each taxonomic group are collected from the NCBI site (www.ncbi.nlm.nih.gov/genomes/lproks.cgi, January 2013). Figure 1 gives the main statistical features of their genome length (Mb) and *GC* content. In order to be "general" in the statistical evaluation of the *LIIS* model, we have chosen taxonomic samples with a large variability of observation parameters concerning simultaneously the genome number, from 15 genomes for the Thermotogae to 34 genomes for the Crenarchaeota and the Spirochaetes (i.e. a variation of 127%), the minimum genome length $n_0$, from 0.34 Mb for the Spirochaetes to 1.81 Mb for the Thermotogae (i.e. a variation of 432%), the maximum genome length $n_{max}$, from 2.83 Mb for the Crenarchaeota to 5.96 Mb for the Chlamydiae (i.e. a variation of 111%), the minimum *CG* content, from 25.8% for the Spirochaetes to 34.7% for the Chlamydiae (i.e. a variation of 34%), and the maximum *CG* content, from 46.9% for the Thermotogae to 65.3% for the Chlamydiae (i.e. a variation of 39%).

For each taxonomic group, we set the minimum genome length $n_0$ to the minimum observed among the genomes of each group. In order to obtain the best estimation of parameters of the formula $p_{GC}(l)$ (Eq. (7.2)) for the two models *LIIS* and *IDIS*, all parameters are scanned as follows: the initial content $p_C(n_0)$ in nucleotide *C*, from 0.1 to 0.3 by step 0.01 (i.e. the *GC* content $p_{GC}(n_0)$ varies from 0.2 to 0.6 and thus includes the *GC* content interval observed in the data, Fig. 1), the equilibrium frequency $\pi_C$ of *C*, from 0.00 to 0.5 by step 0.05, the ratio $\frac{\alpha+\beta}{r}$, from 0.01 to 4 by step 0.01, the ratio $\frac{r_C}{r}$, from 0 to 0.5 by step 0.05 and, for the *LIIS* model only, the maximal genome length $n_{max}$ varies from the observed maximum length (Mb) of each taxonomic group to 10 Mb by step 0.02 Mb. The *IDIS* and *LIIS* models are evaluated with the classical statistical parameter *RSS* (Residual Sum of Squares). The best fit curves are plotted in Figure 2.

As expected by the theory (see Introduction in Section 7), the *LIIS* model has *RSS* values significantly smaller than the *IDIS* model with the four taxonomic groups, in particular an error decrease up to 43% with Chlamydiae. Furthermore, the best fit curves $p_{GC}(l)$ in the *LIIS* and *IDIS* models may be totally different with a change of concavity/convexity, in particular with the Chlamydiae and the Crenarchaeota. Otherwise, the application of the *LIIS* model for *GC* content analysis in bacterial genomes may be improved in future by incorporating some additional biological factors such as the effect of selection force, e.g. the variation of base composition at synonymous sites since bacterial genomes have high gene content [24].

## 8. Conclusion

We have developed a new molecular evolution model based on Limited Insertion Independent of Substitution (*LIIS* model). This *LIIS* model is more general than the *IDIS* model [13,14], both from a theoretical point of view as the *IDIS* model is a particular case of the *LIIS* model when the maximum sequence length $n_{\max}$ tends to infinity, and from a practical point of view for the *GC* content analysis in complete genomes of four prokaryotic taxonomic groups. This research work is a theoretical contribution to the very few classes of mathematical models of gene evolution based on substitution and insertion. To our knowledge, there is no mathematical molecular evolution model including a limited insertion process.

## References

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

[2] D.G. Arquès, C.J. Michel, Analytical solutions of the dinucleotide probability after and before random mutations, J. Theor. Biol. 175 (1995) 533.

[3] U. Bastolla, A. Moya, E. Viguera, R.C. van Ham, Genomic determinants of protein folding thermodynamics in prokaryotic organisms, J. Mol. Biol. 343 (2004) 1451.

[4] E. Benard, C.J. Michel, 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 (2009) 245.

[5] E. Benard, C.J. Michel, A generalization of substitution evolution models of nucleotides to genetic motifs, J. Theor. Biol. 288 (2011) 73.

[6] J. Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach, J. Mol. Evol. 17 (1981) 368.
[7] J. Felsenstein, G.A. Churchill, A hidden Markov model approach to variation among sites in rate of evolution, Mol. Biol. Evol. 13 (1996) 93.
[8] E. Freese, On the evolution of base composition of DNA, J. Theor. Biol. 3 (1962) 82.
[9] M. Hasegawa, H. Kishino, T. Yano, Dating of the human–ape splitting by a molecular clock of mitochondrial DNA, J. Mol. Evol. 22 (1985) 160.
[10] T.H. Jukes, C.R. Cantor, Evolution of protein molecules, in: H.N. Munro (Ed.), Mammalian Protein Metabolism, Academic Press, New York, 1969, p. 21.
[11] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, J. Mol. Evol. 16 (1980) 111.
[12] M. Kimura, Estimation of evolutionary distances between homologous nucleotide sequences, Proc. Nat. Acad. Sci. USA 78 (1981) 454.
[13] S. Lèbre, C.J. Michel, A stochastic evolution model for residue insertion-deletion independent from substitution, Comput. Biol. Chem. 34 (2010) 259.
[14] S. Lèbre, C.J. Michel, 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 (2012) 1764.
[15] T.R. Malthus, An Essay on the Principle of Population, Penguin, Harmondsworth, England, 1798.
[16] G. McGuire, M.C. Denham, D.J. Balding, Models of sequence evolution for DNA sequences containing gaps, Mol. Biol. Evol. 18 (2001) 481.
[17] D. Metzler, Statistical alignment based on fragment insertion and deletion models, Bioinformatics 19 (2003) 490.
[18] C.J. Michel, 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 (2007) 677.
[19] I. Miklós, G.A. Lunter, I. Holmes, A long indel model for evolutionary sequence alignment, Mol. Biol. Evol. 21 (2004) 529.
[20] I. Miklós, A. Novák, R. Satija, R. Lyngsø, J. Hein, Stochastic models of sequence evolution including insertion–deletion events, Stat. Methods Med. Res. 18 (2009) 453.
[21] D. Mitchell, GC content and genome length in Chargaff compliant genomes, Biochem. Biophys. Res. Commun. 353 (2007) 207.
[22] N.A. Moran, Microbial minimalism: genome reduction in bacterial pathogens, Cell 108 (1962) 583.
[23] H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarez-Valín, G. Bernardi, Genomic GC level, optimal growth temperature, and genome size in prokaryotes, Biochem. Biophys. Res. Commun. 347 (2006) 1.
[24] R. Raghavan, Y.D. Kelkar, H. Ochman, A selective force favoring increased G+C content in bacterial genes, Proc. Nat. Acad. Sci. USA 109 (2012) 14504.
[25] E. Rivas, Evolutionary models for insertions and deletions in a probabilistic modeling framework, BMC Bioinf. 6 (2005) 63.
[26] E. Rivas, S.R. Eddy, Probabilistic phylogenetic inference with insertions and deletions, PLoS Comput. Biol. 4 (9) (2008) e1000172.
[27] E.P. Rocha, A. Danchin, Base composition bias might result from competition for metabolic resources, Trends Genetics 18 (2002) 291.
[28] S.S. Satapathy, M. Dutta, S.K. Ray, Variable correlation of genome GC% with transfer RNA number as well as with transfer RNA diversity among bacterial groups: a-Proteobacteria and Tenericutes exhibit strong positive correlation, Microbiol. Res. 165 (2010) 232.
[29] N. Sueoka, On the genetic basis of variation and heterogeneity of DNA base composition, Proc. Nat. Acad. Sci. USA 48 (1962) 582.
[30] N. Takahata, M. Kimura, A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes, Genetics 98 (1981) 641.
[31] K. Tamura, M. Nei, Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees, Mol. Biol. Evol. 10 (1993) 512.
[32] S. Tavaré, Some probabilistic and statistical problems in the analysis of DNA sequences, Lect. Math. Life Sci. vol. 17, 1986, pp. 57.
[33] G. Teschl, Ordinary differential equations and dynamical systems. Graduate Studies in Mathematics, vol. 140, Amer. Math. Soc., Providence, 2012.
[34] J.L. Thorne, H. Kishino, J. Felsenstein, An evolutionary model for maximum likelihood alignment of DNA sequences, J. Mol. Evol. 33 (1991) 114.
[35] J.L. Thorne, H. Kishino, J. Felsenstein, Inching toward reality: an improved likelihood model of sequence evolution, J. Mol. Evol. 34 (1992) 3.
[36] P.-F. Verhulst, Notice sur la loi que la population poursuit dans son accroissement, Correspondance mathématique et physique 10 (1838) 113.
[37] H.C. Wang, E. Susko, A.J. Roger, On the correlation between genomic G + C content and optimal growth temperature in prokaryotes: data quality and confounding factors, Biochem. Biophys. Res. Commun. 342 (2006) 681.
[38] Z. Yang, Estimating the pattern of nucleotide substitution, J. Mol. Evol. 39 (1994) 105.