Bull Math Biol (2012) 74:1764–1788
DOI 10.1007/s11538-012-9735-z

ORIGINAL ARTICLE

An Evolution Model for Sequence Length Based on Residue Insertion–Deletion Independent of Substitution: An Application to the *GC* Content in Bacterial Genomes

Sophie Lèbre · Christian J. Michel

Received: 3 July 2011 / Accepted: 1 May 2012 / Published online: 30 May 2012

Abstract We introduce here a gene evolution model which is an extension of the time-continuous stochastic *IDIS* model (Lèbre and Michel in J. Comput. Biol. Chem. 34:259–267, 2010) to sequence length. This new *IDISL* (Insertion Deletion Independent of Substitution based on sequence Length) model gives an analytical expression of the residue occurrence probability $\mathbf{p}(l)$ at sequence length l depending on stochastically independent processes of substitution, insertion, and deletion. Furthermore, in contrast to all mathematical models in this research field, the substitution, insertion, and deletion parameters of the *IDISL* model are independent of each other. For any diagonalizable substitution matrix $\mathbf{M}$, the residue occurrence probability $\mathbf{p}(l)$ is given as a function of the eigenvalues of $\mathbf{M}$, the eigenvector matrix of $\mathbf{M}$, a vector $\mathbf{r}$ of the residue insertion rates, a deletion rate d (unlike our previous *IDIS* model), and a vector of the initial residue occurrence probability $\mathbf{p}(l_0)$ at sequence length l_0.

As another difference with the classical evolution approaches which mainly focus on sequence alignment, the *IDIS* class of models allows a mathematical analysis of the behavior of the residue occurrence probability according to either evolution time or sequence length. The length parameter can be associated with any nucleotide regions: genes, genomes, introns, repeats, $5'$ and $3'$ regions, etc. Three properties of the *IDISL* model are given in relation with the sequence length l: parameter scale, inverse evolution, and residue equilibrium distribution. Nucleotide occurrence probabilities are given in the particular case of the *IDISL-HKY* model, i.e. the *IDISL* model associated with the *HKY* asymmetric substitution matrix (Hasegawa et al. in J. Mol. Evol. 22:160–174, 1985).

S. Lèbre · C.J. Michel (✉)
Equipe de Bioinformatique Théorique, BFO, LSIIT (UMR 7005), Université de Strasbourg,
Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France
e-mail: michel@dpt-info.u-strasbg.fr

S. Lèbre
e-mail: lebre@dpt-info.u-strasbg.fr

An application of the *IDISL* model is developed for a massive statistical analysis of *GC* content in all complete bacterial genomes available to date (894 non-anaerobic and anaerobic genomes). The *IDISL-HKY* model confirms the increase of the *GC* content with the genome length for two non-anaerobic taxonomic groups of bacterial genomes. Moreover, the non-linear modelling proposed by the *IDISL* model outperforms the most recent modelling of *GC* content in these bacterial genomes (Wang et al. in Biochem. Biophys. Res. Commun. 342:681–684, 2006; Musto et al. in Biochem. Biophys. Res. Commun. 347:1–3, 2006).

Keywords Evolution model · Stochastic model · Substitution · Insertion · Deletion · Nucleotides · Sequence length · Bacterial genomes · *GC* content · Comparative genomics

1 Introduction

Substitution, insertion, and deletion of nucleotides are important evolutionary processes. A major challenge for understanding genome evolution is to elucidate the role of each of these mutational events. Evolution models were initially developed to study the substitution rates of nucleotides (adenine A, cytosine C, guanine G, thymine T). The first evolution models were based on symmetric substitution matrices with one formal parameter (probability a for all nucleotide substitution types) (Jukes and Cantor 1969), two formal parameters (probabilities α and β for the nucleotide transitions and transversions, respectively) (Kimura 1980) and three formal parameters (probability α for transitions and probabilities β_1 and β_2 for the two types of transversions) (Kimura 1981). These models were later generalized to asymmetric substitution matrices (Felsenstein 1981; Takahata and Kimura 1981; Hasegawa et al. 1985; Tavaré 1986; Tamura and Nei 1993; Yang 1994; Felsenstein and Churchill 1996). As their equilibrium distribution may differs from $1/4$ for all nucleotides, the latter are less constrained.

Over the last 20 years, only very few evolution models were extended to the insertion and the deletion of residues (nucleotides, amino acids, etc.) in addition to residue substitution. The first model is based on a continuous time birth–death process which is governed by explicit parameters for the insertion and deletion rates (Thorne et al. 1991). An extension of this model enables the insertion or deletion 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). However, for all these insertion–deletion models, the inserted residues are assumed to be distributed according to the equilibrium distribution of the substitution process, which drastically restricts the scope of the model. Moreover, these insertion–deletion models are reversible. This property is useful for inferring un-rooted phylo-genetic trees but requires some theoretical constraints to be imposed on the insertion and deletion rates which prevent the insertion and deletion processes from being independent of each other. For example, in a pairwise alignment, the reversibility constraint imposes that the expected frequencies of insertions and deletions must be identical.

Another class of model uses a Markov model with an extended substitution matrix comprising the four nucleotides and the gap character, i.e. a substitution matrix with one additional line and one additional column (McGuire et al. 2001). Thus, an insertion corresponds to the substitution of a gap by a nucleotide whereas a deletion amounts to the substitution of a nucleotide by a gap. In this first approach introducing an extended substitution matrix, the inserted residues are assumed to be distributed according to the equilibrium distribution of the substitution process. In order to obtain an insertion process independent of the substitution process, the model of McGuire et al. (2001) was extended to a non-reversible model by adding explicit parameters for the insertion rates (Rivas 2005; Rivas and Eddy 2008). In the particular case of a reversible substitution matrix and insertion rates proportional to the nucleotide equilibrium distribution associated with the substitution matrix, Rivas and Eddy (2008) obtain an analytical expression of the substitution probability of residue i into residue j over some time t (Eq. (9) in Rivas and Eddy 2008). However, the residue occurrence probability which can be derived from their Eq. (9), is a function of the common deletion rate of nucleotides (called μ in their paper). This concept is not realistic as a deletion rate which is identical for all residues (common rate) should only alter the sequence length but obviously not the residue distribution. Moreover, in both cases, the definition of the substitution matrix extended to gaps leads to substitution and deletion processes which are stochastically dependent. Indeed, the definition of the stochastic extended matrix implies constraints between the substitution and the deletion rates (the lines of the matrices of Eq. (1) in McGuire et al. (2001) and Eq. (4) in Rivas and Eddy (2008) sum to 1).

Recently, we developed a more general class of non-reversible evolution models for residue Insertion Deletion Independent of Substitution called the *IDIS* model (Lèbre and Michel 2010). The *IDIS* model is based on a continuous Markov process where the substitution, insertion, and deletion processes are stochastically independent of each other. Furthermore, in contrast to all mathematical models in this research field, the substitution, insertion, and deletion parameters of the *IDISL* model are independent of each other; in particular, the distribution of the inserted residues is different from the substitution equilibrium distribution. In contrast to the approaches in the last 10 years (McGuire et al. 2001; Rivas 2005; Rivas and Eddy 2008), the *IDIS* model is not based on the introduction of a gap character for extending the substitution rate matrix. Indeed, the modelling of the insertion and deletion processes was inspired by a concept in population dynamics (Malthus 2000) and used to model a stochastic differential equation satisfied by the residue occurrence probability $\mathbf{p}(t)$ at time t. This mathematical approach which was never considered in the previous evolution models, leads to several important consequences for the study of gene/genome evolution. Thus, the *IDIS* model allows one to represent the real physical process of gene evolution based on the three processes of substitution, insertion, and deletion in residue sequences. It gives exact residue occurrence probabilities according to evolution time t, which could only be approximated by a heavy computer simulation study reproducing these three evolutionary processes. Moreover, as naturally expected from a probabilistic point of view, the residue occurrence probability given by the *IDIS* model does not depend on the deletion rate which is identical for all residues (the deletion variable d is not part of the general

Eq. (2.13) in Lèbre and Michel 2010). Thus, the modelling of the deletion rate constitutes a major difference with the model in Rivas and Eddy (2008).

The *IDIS* model, originally a function of time t, is extended here to a model for residue occurrence probability as a function of the residue sequence length l. We call the extension of the *IDIS* model based on sequence length, the *IDISL* model. An analytical expression is obtained for residue occurrence probability $\mathbf{p}(l)$ at sequence length l for any diagonalizable substitution matrix $\mathbf{M}$, as a function of the eigenvalues λ_k of $\mathbf{M}$, the eigenvector matrix $\mathbf{Q}$ of $\mathbf{M}$, a vector $\mathbf{r}$ of the residue insertion rates, a deletion rate d (in contrast to the *IDIS* model) and a vector of initial residue occurrence probability $\mathbf{p}(l_0)$ at sequence length l_0. An entirely explicit expression of the nucleotide occurrence probabilities is given in the particular case of the *IDISL-HKY* model, i.e. the *IDISL* model associated with the classical *HKY* substitution model (Hasegawa et al. 1985) which is defined by an asymmetric substitution matrix and represents one of the most general substitution models. Explicit formulas are also given for two particular cases: equal nucleotide insertion rates and nucleotide insertion–deletion only.

The applications of the *IDISL* model can be various: sequence alignment, phylogeny, etc. The *IDISL* model can also be used for extending our previous evolution models developed over the last 20 years: 'primitive' gene models, 'primitive' genetic motif models (DNA or amino acids), substitution rate study, analysis of residue occurrence probabilities in the natural evolution time direction (past → present or present → future) or in the inverse direction (present → past), e.g. Arquès and Michel (1993, 1995), Michel (2007), Benard and Michel (2009). In summary, the *IDISL* model allows the mathematical analysis of the residue occurrence probability according to any residue sequence length, e.g. on the genetic alphabet: genes, genomes, introns, repeats, $5'$ and $3'$ regions, etc. In Sect. 3, we give an example of application of the *IDISL* model for the analysis of the GC content in complete bacterial genomes according to their genome lengths. This biological problem has been a matter of debate for several years. The GC content of complex microbial communities may be influenced by several environmental factors (Foerstner et al. 2005): oxygen requirements, nitrogen, utilization, habitats, salinity, alkalinity, etc. However, it has been observed that large bacterial genomes tend to be GC-rich and small bacterial genomes tend to be AT-rich (Moran 1962; Rocha and Danchin 2002; Bastolla et al. 2004). Furthermore, a linear relationship between GC content and genome length within a sub-group of genomes was reported (Wang et al. 2006; Satapathy et al. 2010). It was also suggested that the bacterial GC content variation has arisen as a consequence of mutational bias (Freese 1962; Sueoka 1962; Wang et al. 2006). These biological questions—GC content, genome length, genome evolution—can be modeled by our *IDISL* model. We perform a massive statistical analysis based on the complete bacterial genomes currently available from the NCBI site (January 2011). The genomes are classified according to their taxonomy and anaerobic/non-anaerobic property. The sub-groups containing at least 30 complete genomes are selected and involve the non-anaerobic sub-groups of *Actinobacteria, Alphaproteobacteria, Bacteroidetes/Chlorobi, Betaproteobacteria, Cyanobacteria, Firmicutes, and Gammaproteobacteria* and the anaerobic subgroups of *Euryarchaeota* and *Firmicutes*. Only the *Firmicutes* is selected in both anaerobic and non-anaerobic categories. This data represents a total of 894 complete genomes.

This paper is organized as follows. Section 2 gives the mathematical aspects of the *IDISL* model. After recalling the time-continuous *IDIS* model definition, the new *IDISL* model is derived. An analytical expression of the residue occurrence probability $\mathbf{p}(l)$ as a function of sequence length l is given for any diagonalizable substitution matrix $\mathbf{M}$. Three properties of the *IDISL* model are given in relation with the sequence length l: parameter scale, inverse evolution and residue equilibrium distribution. Analytical nucleotide occurrence probabilities are given in the particular case of the *IDISL-HKY* model, i.e. the *IDISL* model associated with the *HKY* substitution model (Hasegawa et al. 1985), one of the most general substitution models. Two particular formulas are also derived: equal nucleotide insertion rates and nucleotide insertion–deletion only. In Sect. 3, the *IDISL-HKY* model is applied to the *GC* content modelling in bacterial genomes and brings new insights to this biological debate.

2 Methods

2.1 Recall of the *IDIS* Model Function of Evolution Time *t*

We recall here the definition of the *IDIS* model, a time-continuous stochastic evolution model for residue insertion and deletion independent of substitution recently introduced by Lèbre and Michel (2010). In contrast to the classical substitution–insertion–deletion models (see Introduction), the *IDIS* model is defined by explicit parameters for the deletion rate d and the insertion rate r_i of each residue i. The insertion rates and the deletion rate are not only independent of each other, but also independent of the substitution parameters. Let us consider an alphabet of K residues. For example, $K = 4$ for the set of nucleotides $\{A, C, G, T\}$, $K = 20$ for the set of amino acids, $K = 2$ for the set of purine and pyrimidine $\{R, Y\}$. For all $1 \le i \le K$, let $p_i(t)$ be the occurrence probability of residue i at time $t \ge 0$ per 'residue site' in the sequence and $\mathbf{p}(t) = [p_i(t)]_{1\le i\le K}$ the column vector of size K made of the probabilities $p_i(t)$ for all $1 \le i \le K$.

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

$$\begin{aligned}\mathbf{p}'(t) &= \underbrace{(\mathbf{M}-\mathbf{I})\cdot\mathbf{p}(t)}_{\text{Substitution}} + \underbrace{\big(-\tau\mathbf{p}(t)+\mathbf{r}\big)}_{\text{Insertion–Deletion}}\\ &= \mathbf{A}\cdot\mathbf{p}(t)+\mathbf{r} \qquad (1)\end{aligned}$$

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

(i) Substitution term in Eq. (1). The change of the residue occurrence probability due to substitutions is governed by the classical matrix differential equation (Michel 2007)

$$\begin{aligned}\mathbf{p}'(t) &= \mathbf{M} \cdot \mathbf{p}(t) - \mathbf{p}(t) \\ &= (\mathbf{M} - \mathbf{I}) \cdot \mathbf{p}(t) \qquad (2)\end{aligned}$$

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

Remark 1 The substitution probability matrix $\mathbf{M}$ is stochastic in column. Indeed, for all $1 \le j \le K$, the elements of matrix $\mathbf{M}$ satisfy $\sum_{1 \le i \le K} m_{ij} = \sum_{1 \le i \le K} \Pr(j \to i) = 1$. The substitution probability matrix $\mathbf{M}$ is the transpose matrix of the classical substitution matrix $\pi = [\Pr(i \to j)]_{1 \le i,j \le K}$ which is stochastic in line (e.g. Kimura 1980, 1981), i.e. $\pi_{ij} = \Pr(i \to j) = m_{ji}$.

(ii) Insertion–deletion term in Eq. (1). The insertion–deletion process is modelled by explicit parameters which are set independently from the substitution parameters: r_i, the insertion rate per site of each residue i, $\forall 1 \le i \le K$, $r_i \ge 0$, and d, the deletion rate for all residues, $d \ge 0$. Let $n_i(t)$ be the occurrence number of residue i in the biological sequence at time t and $n(t) = \sum_{1 \le i \le K} n_i(t)$ be the total number of residues at time t. By definition, a sequence has at least one residue, i.e. $n(t) \ge 1$. Using a common model in population dynamics (Malthus 2000), the growth rate $n'_i(t) = \frac{\partial n_i(t)}{\partial t}$ of residue i at time t due to insertion is equal to $r_i \times n(t)$. Similarly, the growth rate $n'_i(t)$ of residue i at time t due to deletion is $d \times n_i(t)$. Thus, the growth rate $n'_i(t)$ resulting from the insertion–deletion process is, for all $1 \le i \le K$,

$$n'_i(t) = r_i \times n(t) - d \times n_i(t). \qquad (3)$$

The derivative $\mathbf{p}'(t)$ of the occurrence probability of residue i at time t can be written

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

By replacing $n'_j(t)$ using Eq. (3), we obtain (see the detail in Lèbre and Michel 2010)

$$p'_i(t) = r_i - \left(\sum_{1 \le j \le K} r_j\right) p_i(t).$$

Then the change of the residue occurrence probability due to insertion–deletion is explained by the matrix differential equation

$$\mathbf{p}'(t) = -\tau \mathbf{p}(t) + \mathbf{r}. \tag{4}$$

An analytical solution of the *IDIS* model defined by Eq. (1) is derived when the substitution probability matrix $\mathbf{M}$ can be diagonalized with real eigenvalues $(\lambda_k)_{1\le k\le K}$. Let $\mathbf{Q}$ be an associated eigenvector matrix of $\mathbf{M}$, the kth column of $\mathbf{Q}$ being an eigenvector for eigenvalue λ_k. Then, for any non-zero residue insertion rates vector $\mathbf{r} = [r_i]_{1\le i\le K}$ (such that $\tau = \sum_{1\le i\le K} r_i > 0$), using the method of variation parameters, the residue occurrence probability $\mathbf{p}(t)$ is

$$\mathbf{p}(t) = \mathbf{Q}\cdot \mathbf{D}_1(t)\cdot \mathbf{Q}^{-1}\cdot \mathbf{p}(0) + \mathbf{Q}\cdot \mathbf{D}_2(t)\cdot \mathbf{Q}^{-1}\cdot \mathbf{r} \tag{5}$$

where $\mathbf{D}_1(t) = \mathrm{Diag}((e^{-(\tau+1-\lambda_k)t})_{1\le k\le K})$, $\mathbf{D}_2(t) = \mathrm{Diag}((\frac{1}{\tau+1-\lambda_k}(1 - e^{-(\tau+1-\lambda_k)t}))_{1\le k\le K})$ and $\mathbf{p}(0) = [p_i(0)]_{1\le i\le K}$ is the initial residue occurrence probability at time 0. Let us define matrix $\mathbf{O}_k$ of size $K\times K$ by using the eigenvector matrix $\mathbf{Q}$ of $\mathbf{M}$,

$$\mathbf{O}_k[i,j] = \mathbf{Q}[i,k]\mathbf{Q}^{-1}[k,j]. \tag{6}$$

After some algebraic manipulation, an analytical expression of the residue occurrence probability $\mathbf{p}(t)$ is obtained

$$\mathbf{p}(t) = \left(\sum_{k=1}^{K}\frac{1}{\tau+1-\lambda_k}\mathbf{O}_k\right)\cdot \mathbf{r} + \sum_{k=1}^{K}\mathbf{O}_k\cdot\left(\mathbf{p}(0) - \frac{1}{\tau+1-\lambda_k}\mathbf{r}\right)e^{-(\tau+1-\lambda_k)t} \tag{7}$$

where $(\lambda_k)_{1\le k\le K}$ are the eigenvalues of the substitution probability matrix $\mathbf{M}$, matrices $(\mathbf{O}_k)_{1\le k\le K}$ are defined in (6) from the eigenvector matrix $\mathbf{Q}$ of $\mathbf{M}$, $\mathbf{r} = [r_i]_{1\le i\le K}$ is the vector of the residue insertion rates per site, $\tau = \sum_{1\le i\le K} r_i$ is the total insertion rate and $\mathbf{p}(0) = [p_i(0)]_{1\le i\le K}$ is the initial residue occurrence probability at time 0 (see the detail in Lèbre and Michel 2010). It is well known that the substitution matrices of the reversible models are diagonalizable with real eigenvalues (Aldous and Fill 2002) but this is not an exclusive condition as substitution matrices of non-reversible models can also be diagonalized with eigenvalues (e.g., Exercises of Chap. 1 in Kelly 1979).

Remark 2 The sum of the matrices $\{\mathbf{O}_k\}_k$ is $\sum_{k=1}^{K}\mathbf{O}_k = \mathbf{Q}\cdot\mathbf{Q}^{-1} = \mathbf{I}$. Indeed, for all i, j, $\sum_{k=1}^{K}\mathbf{O}_k[i,j] = \sum_{k=1}^{K}\mathbf{Q}[i,k]\mathbf{Q}^{-1}[k,j]$ is the term in row i and column j of the matrix product $\mathbf{Q}\cdot\mathbf{Q}^{-1}$.

Remark 3 The non-zero condition for the insertion rates vector $\mathbf{r}$ ensures that $\tau = \sum_{1\le i\le K} r_i > 0$. Thus, the denominator of the ratio $\frac{1}{\tau+1-\lambda_k}$ is different from zero as the eigenvalues of the stochastic matrix $\mathbf{M}$ satisfies $\lambda_k \le 1$, $\forall 1\le k\le K$. If the insertion rates vector $\mathbf{r}$ is null, then the residue occurrence probability $\mathbf{p}(t)$ satisfies $\mathbf{p}(t) = \mathbf{Q}\cdot\mathbf{D}_1(t)\cdot\mathbf{Q}^{-1}\cdot\mathbf{p}(0)$ with $\tau = 0$ as in the substitution only model (Michel 2007).

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 an analytical expression is not ensured with specific deletion rate d_i for each residue i.

Remark 5 In contrast to the model of Rivas and Eddy (2008) (see Introduction), the residue distribution of the *IDIS* model (Eq. (4)) is independent of the deletion rate d. Indeed, deletions occur with a constant rate for any residue of the sequence, thus leaving the residue distribution globally unchanged.

2.2 *IDISL* Model as a Function of Sequence Length l

From the growth rate $n_i'(t)$ for each residue i defined in Eq. (3), the derivative sequence length is $n'(t) = \sum_{1\le i\le K} n_i'(t) = (\tau - d)n(t)$. We denote by $\Delta = \tau - d$, the difference between the insertion and the deletion rates of residues. Then the number $n(t)$ of residues in the sequence at time t is

$$\forall t > 0, \quad n(t) = n(0)e^{\Delta t} \tag{8}$$

where $n(0)$ is the sequence length at time $t = 0$. From Eq. (8) and with $\Delta \neq 0$, the evolution time t between the sequence length $n(0)$ and $n(t)$ is

$$\forall t > 0, \quad t = \frac{1}{\Delta}\ln\left(\frac{n(t)}{n(0)}\right). \tag{9}$$

From Eq. (7), we derive the residue occurrence probability $\mathbf{p}(l)$ as a function of the sequence length $l = n(t)$ at evolution time t and the sequence length $l_0 = n(0)$. Then, for any non-zero residue insertion rates vector $\mathbf{r} = [r_i]_{1\le i\le K}$ (such that $\tau = \sum_{1\le i\le K} r_i > 0$) and for all $l, l_0 \ge 1$,

$$\mathbf{p}(l) = \left(\sum_{k=1}^{K}\frac{1}{\tau+1-\lambda_k}\mathbf{O}_k\right)\cdot\mathbf{r} + \sum_{k=1}^{K}\mathbf{O}_k\cdot\left(\mathbf{p}(l_0) - \frac{1}{\tau+1-\lambda_k}\mathbf{r}\right)\left(\frac{l}{l_0}\right)^{-\frac{\tau+1-\lambda_k}{\Delta}} \tag{10}$$

where $(\lambda_k)_{1\le k\le K}$ are the eigenvalues of the substitution probability matrix $\mathbf{M}$, matrices $(\mathbf{O}_k)_{1\le k\le K}$ are defined in (6) from the eigenvector matrix $\mathbf{Q}$ of $\mathbf{M}$, $\mathbf{r} = [r_i]_{1\le i\le K}$ is the vector of the residue insertion rates per site, $\tau = \sum_{1\le i\le K} r_i$ is the total insertion rate, $\Delta = \tau - d$ is the difference between the total insertion rate τ and the deletion rate d, and $\mathbf{p}(l_0)$ is the initial residue occurrence probability at length l_0.

2.3 Properties

Proposition 1 (Parameter scale) *When multiplying all the substitution–insertion–deletion parameters, i.e. the non-diagonal elements* $[m_{ij}]_{i\neq j}$ *of the substitution probability matrix* $\mathbf{M}$, *the insertion* $[r_i]$ *and the deletion* d *rates, by a scalar* α, *the residue occurrence probability as a function of the sequence length* l *remains unchanged*

$$\mathbf{p}\big(l; [\alpha m_{ij}]_{i\neq j}, [\alpha r_i], \alpha d\big) = \mathbf{p}\big(l; [m_{ij}]_{i\neq j}, [r_i], d\big). \tag{11}$$

Proof Multiplying the model parameters by a scalar α leads to residue insertion rates vector $\widetilde{\mathbf{r}} = \alpha \mathbf{r}$, then total insertion rate $\widetilde{\tau} = \alpha\tau$, deletion rate $\widetilde{d} = \alpha d$, difference $\widetilde{\Delta} = \widetilde{\tau} - \widetilde{d} = \alpha(\tau - d) = \alpha\Delta$ and finally substitution probability matrix $\widetilde{\mathbf{M}} = \alpha\mathbf{M} + (1-\alpha)\mathbf{I}$. Matrix $\mathbf{M}$ decomposes as $\mathbf{M} = \mathbf{Q}\cdot\mathbf{D}\cdot\mathbf{Q}^{-1}$ where $\mathbf{D} = \mathrm{Diag}((\lambda_k)_{1\le k\le K})$ is the eigenvalues diagonal matrix and $\mathbf{Q}$ is an associated eigenvector matrix, the kth column of $\mathbf{Q}$ being an eigenvector for eigenvalue λ_k. Then matrix $\widetilde{\mathbf{M}}$ decomposes as $\widetilde{\mathbf{M}} = \alpha\mathbf{Q}\cdot\mathbf{D}\cdot\mathbf{Q}^{-1} + (1-\alpha)\mathbf{I} = \mathbf{Q}\cdot\widetilde{\mathbf{D}}\cdot\mathbf{Q}^{-1}$ where matrix $\widetilde{\mathbf{D}} = \alpha\mathbf{D} + (1-\alpha)\mathbf{I} = \mathrm{Diag}\ ((\widetilde{\lambda}_k)_{1\le k\le K})$. Matrix $\widetilde{\mathbf{M}}$ can be diagonalized with real eigenvalues $(\widetilde{\lambda}_k)_{1\le k\le K}$ where $\widetilde{\lambda}_k = \alpha\lambda_k + (1-\alpha)$ for all $1 \le k \le K$. Then $\widetilde{\tau}+1-\widetilde{\lambda}_k = \alpha(\tau+1-\lambda_k)$ and in Eq. (10), $\forall 1 \le k \le K$, $(\widetilde{\tau}+1-\widetilde{\lambda}_k) = \alpha(\tau+1-\lambda_k)$, $\frac{1}{\widetilde{\tau}+1-\widetilde{\lambda}_k}\widetilde{\mathbf{r}} = \frac{1}{\tau+1-\lambda_k}\mathbf{r}$ and $\frac{\widetilde{\lambda}_k-1-\widetilde{\tau}}{\widetilde{\Delta}} = \frac{\lambda_k-1-\tau}{\Delta}$. □

Proposition 2 (Inverse evolution) *The model can be inverted* (*from present to past*), *i.e. the residue occurrence probability* $\mathbf{p}(l_0)$ *of an initial sequence of length* $l_0 \ge 1$ *observed at time* t_0 $(l_0 = n(t_0))$ *is given as a function of the residue occurrence probability* $\mathbf{p}(l)$ *of a sequence of length* $l \ge 1$ *observed at time* $t > t_0$ $(l = n(t))$

$$\mathbf{p}(l_0) = -\left(\sum_{k=1}^{K}\frac{1}{\tau+1-\lambda_k}\mathbf{O}_k\right)\cdot\mathbf{r} + \sum_{k=1}^{K}\mathbf{O}_k\cdot\left(\mathbf{p}(l) + \frac{1}{\tau+1-\lambda_k}\mathbf{r}\right)\left(\frac{l}{l_0}\right)^{\frac{\tau+1-\lambda_k}{\Delta}}. \tag{12}$$

Proof Equation (12) is obtained after algebraic manipulation of the general solution (7) and the expression of the sequence length as a function of time t (8). □

Proposition 3 (Residue equilibrium distribution) *The residue occurrence probability has the same equilibrium distribution for sequence growth or sequence shrinkage. Let* $\mathbf{p}_K = (\sum_{k=1}^{K}\frac{1}{\tau+1-\lambda_k}\mathbf{O}_k)\cdot\mathbf{r}$ *be the constant term in the residue occurrence probability* $\mathbf{p}(l)$ *defined in Eq.* (10). *When considering either a sequence growth* $(\Delta > 0)$ *or a sequence shrinkage* $(\Delta < 0)$ *process*, *then the residue equilibrium distribution* $\mathbf{p}_K$ *is equal to*

$$\mathbf{p}_K = \lim_{l\to\infty,\,\Delta>0}\mathbf{p}(l) = \lim_{l\to 0,\,\Delta<0}\mathbf{p}(l). \tag{13}$$

Proof The sign of the difference Δ determines either sequence growth $(\Delta > 0)$ or sequence shrinkage $(\Delta < 0)$. When considering a sequence growth process, then $\frac{l}{l_0} > 1$ for all sequence length $l \neq l_0$. For all $1 \le k \le K$, the eigenvalue satisfies $\lambda_k \le 1$ and the total insertion rate τ is strictly positive. Thus, $\lambda_k - 1 - \tau < 0$. Then, the exponent term $\frac{\lambda_k-1-\tau}{\Delta}$ in Eq. (10) is strictly negative. Then, for all $1 \le k \le K$, the term $(\frac{l}{l_0})^{\frac{\lambda_k-1-\tau}{\Delta}}$ tends to 0 when $l \to \infty$ and the residue equilibrium distribution satisfies $\lim_{l\to\infty,\Delta>0}\mathbf{p}(l) = \mathbf{p}_K$. Respectively, when considering a sequence shrinkage process, i.e. $\Delta < 0$ and $\frac{l}{l_0} < 1$ for all sequence length $l \neq l_0$, then the exponent term $\frac{\lambda_k-1-\tau}{\Delta}$ is strictly positive and for all $1 \le k \le K$, the term $(\frac{l}{l_0})^{\frac{\lambda_k-1-\tau}{\Delta}}$ tends to 0 when $l \to 0$. Then the residue equilibrium distribution also satisfies $\lim_{l\to 0,\Delta<0}\mathbf{p}(l) = \mathbf{p}_K$. □

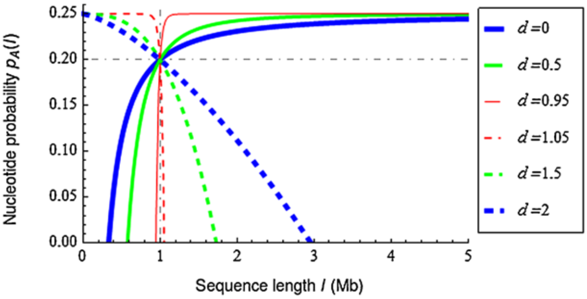

Fig. 1 Illustration of the propositions 'Inverse evolution' and 'Residue equilibrium' distribution. Occurrence probability $p_A(l)$ of nucleotide A at sequence length l with insertion rate $\tau = 1$ ($r_A = r_C = r_G = r_T = 0.25$) and deletion rate varying from $d = 0$ to $d = 2$. The initial conditions are set to $l_0 = 1$ Mb (*vertical dot-dashed line*) and $p_A(l_0) = 0.2$ (*horizontal dot-dashed line*). A particular HKY substitution matrix **M** used in Sect. 2.4 (Eq. (16)) was chosen with parameters: $\alpha = 0.2$, $\beta = 0.1$, $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$

Starting from an initial sequence length l_0 and an initial residue occurrence probability $\mathbf{p}(l_0)$, the *IDISL* model can be applied in four distinct frameworks determined by (i) the sign of the difference Δ; and (ii) the direction of evolution which can be natural (from present to future) or inverse (from present to past). Figure 1 gives an illustration for nucleotide A with $l_0 = 1$ Mb, $p_A(l_0) = 0.2$, $\tau = 1$ and deletion rate d varying from 0 to 2. Thus, the difference Δ is strictly positive for $d = 0$, 0.5 or 0.95 (solid curves) and negative for $d = 1.05$, 1.5 or 2 (dashed curves).

Evolution dominated by residue insertion ($\Delta > 0$) increases the sequence length in natural evolution time direction (solid curves starting from the initial sequence length 1 Mb up to larger sequences in the upper right corner of Fig. 1) and the occurrence probability of nucleotide A toward 0.25 (due to the choice of the evolution parameters, see Fig. 1 legend). On the contrary, inverting this evolutionary process amounts to sequence shrinkage (solid curves from initial sequence length 1 Mb to smaller sequences in the lower left corner of Fig. 1) and decreases the occurrence probability of A toward 0.

Evolution dominated by residue deletion ($\Delta < 0$) decreases the sequence length in natural evolution time direction (dashed curves from initial sequence length 1 Mb to smaller sequences in the upper left corner of Fig. 1) and increases the occurrence probability of nucleotide A toward 0.25. On the contrary, inverting this evolutionary process amounts to sequence increase (dashed curves from initial sequence length 1 Mb to larger sequences in the lower right corner of Fig. 1) and decreases the occurrence probability of A toward 0.

As expected from Proposition 3, occurrence probability $p_A(l)$ of nucleotide A tends to 0.25 when evolution goes, i.e. either $l \to_{t\to\infty,\Delta>0} \infty$ when $\Delta > 0$ or $l \to_{t\to\infty,\Delta<0} 0$ when $\Delta < 0$. In terms of quantitative behavior, the smaller the difference Δ between insertion and deletion rates is the greater the variation of the residue occurrence probability is. Indeed, when $\Delta \simeq 0$ as plotted with thin curves in Fig. 1 where $\tau = 1$ ($d = 0.95$ with thin solid curve or $d = 1.05$ with thin dashed curve), a large change in occurrence probability of nucleotide A is observed for a small change in sequence length. Indeed, with $\Delta \simeq 0$, the number of insertions mainly counterbalances the deletions. Thus, the sequence is subject to many evolution events while the sequence length changes only slightly. Thus, for example in Fig. 1, the changes in occurrence probability of nucleotide A due to residue insertion are greater when $d = 0.95$ (thin solid curve) than when $d = 0.5$ (medium solid curve).

2.4 Analytical Probabilities of Nucleotides with the *IDISL-HKY* Model

Analytical expressions of the occurrence probabilities $\mathbf{p}(l)$ of nucleotides A, C, G, and T are derived for the *IDISL* model with the classical substitution matrix $\mathbf{M}_{HKY}$ (Hasegawa et al. 1985) defined by six formal parameters: the transition and transversion rates, α and β, respectively, and the equilibrium nucleotide frequencies π_A, π_C, π_G, π_T

$$\mathbf{M}_{HKY} = \begin{pmatrix} n_A & \beta\pi_A & \alpha\pi_A & \beta\pi_A \\ \beta\pi_C & n_C & \beta\pi_C & \alpha\pi_C \\ \alpha\pi_G & \beta\pi_G & n_G & \beta\pi_G \\ \beta\pi_T & \alpha\pi_T & \beta\pi_T & n_T \end{pmatrix} \tag{14}$$

where for all j in $\{A, C, G, T\}$, $n_j = 1 - \Sigma_{i \neq j} \mathbf{M}_{HKY}[i, j]$ such that matrix $\mathbf{M}_{HKY}$ is stochastic in column. This asymmetric matrix $\mathbf{M}_{HKY}$ is the basis of one of the most general substitution models whose equilibrium distribution differs from 1/4 for all nucleotides. This version of the *IDISL* model is called *IDISL-HKY* model. The *IDISL-HKY* model is used in Sect. 3 for modelling the GC content in complete bacterial genomes.

2.4.1 General IDISL-HKY Formula

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). We derive the nucleotide occurrence probability $\mathbf{p}(l)$ with the *IDISL-HKY* model from Eq. (10), the four eigenvalues of matrix $\mathbf{M}_{HKY}$

$$\{\lambda_1 = 1 - \beta, \lambda_2 = 1 - \alpha\pi_R - \beta\pi_Y, \lambda_3 = 1 - \alpha\pi_Y - \beta\pi_R, \lambda_4 = 1\} \tag{15}$$

and their associated eigenvectors

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

After some algebraic manipulation, we derive the occurrence probability $\mathbf{p}(l)$ of each nucleotide A, C, G, and T at sequence length l

$$\mathbf{p}(l) = \mathbf{p}_K + k_{1,R,Y} \begin{pmatrix} \frac{\pi_A}{\pi_R} \\ -\frac{\pi_C}{\pi_Y} \\ \frac{\pi_G}{\pi_R} \\ -\frac{\pi_T}{\pi_Y} \end{pmatrix} \left(\frac{l}{l_0}\right)^{-\frac{\mu_1}{\Delta}} + \begin{pmatrix} k_{2,A,G}(\frac{l}{l_0})^{-\frac{\mu_2}{\Delta}} \\ k_{3,C,T}(\frac{l}{l_0})^{-\frac{\mu_3}{\Delta}} \\ -k_{2,A,G}(\frac{l}{l_0})^{-\frac{\mu_2}{\Delta}} \\ -k_{3,C,T}(\frac{l}{l_0})^{-\frac{\mu_3}{\Delta}} \end{pmatrix} \tag{16}$$

where $\pi_A, \pi_C, \pi_G, \pi_T$ are the equilibrium nucleotide frequencies, l_0 is the initial sequence length, and for all $1 \le i \le 3$, $j \in \{A, C, R\}$, $k \in \{G, T, Y\}$,

$$\begin{aligned} k_{i,j,k} &= \frac{\pi_j(r_k - \mu_i p_k(l_0)) - \pi_k(r_j - \mu_i p_j(l_0))}{(\pi_j + \pi_k)\mu_i} \\ \mu_i &= \tau + 1 - \lambda_i \end{aligned}$$

$\lambda_1, \lambda_2, \lambda_3$ are the eigenvalues of matrix $\mathbf{M}_{HKY}$ defined in (15), $\mu_1 = \tau + \beta$, $\mu_2 = \tau + \alpha\pi_R + \beta\pi_Y$, $\mu_3 = \tau + \alpha\pi_Y + \beta\pi_R$ with α and β, the transition and transversion rates, respectively,

$$\begin{aligned} p_R(l_0) &= p_A(l_0) + p_G(l_0), \qquad p_Y(l_0) = p_C(l_0) + p_T(l_0) \\ \pi_R &= \pi_A + \pi_G, \qquad \pi_Y = \pi_C + \pi_T \\ r_R &= r_A + r_G, \qquad r_Y = r_C + r_T \\ \Delta &= \tau - d \\ \tau &= \sum_{1 \le i \le 4} r_i \end{aligned}$$

and $\mathbf{p}_K$ is the equilibrium distribution for predominant insertion ($\Delta > 0$)

$$\mathbf{p}_K = \lim_{l \to \infty, \Delta > 0} \mathbf{p}(l) = \begin{pmatrix} \pi_A(1 - \frac{l_{1,R,Y}}{\pi_R}) - l_{2,A,G} \\ \pi_C(1 + \frac{l_{1,R,Y}}{\pi_Y}) - l_{3,C,T} \\ \pi_G(1 - \frac{l_{1,R,Y}}{\pi_R}) + l_{2,A,G} \\ \pi_T(1 + \frac{l_{1,R,Y}}{\pi_Y}) + l_{3,C,T} \end{pmatrix}$$

with, for all $1 \le i \le 3$, $j \in \{A, C, R\}$, $k \in \{G, T, Y\}$,

$$l_{i,j,k} = \frac{\pi_j r_k - \pi_k r_j}{(\pi_j + \pi_k)\mu_i}.$$

2.4.2 Equal Insertion Rate Formula

When assuming that the nucleotide insertion rates are all equal ($r_A = r_C = r_G = r_T \neq 0$), the occurrence probability $\mathbf{p}(l)$ of each nucleotide A, C, G, and T at sequence length l simplifies to

$$\mathbf{p}(l) = \mathbf{p}_K + m_{1,R,Y} \begin{pmatrix} \frac{\pi_A}{\pi_R} \\ -\frac{\pi_C}{\pi_Y} \\ \frac{\pi_G}{\pi_R} \\ -\frac{\pi_T}{\pi_Y} \end{pmatrix} \left(\frac{l}{l_0}\right)^{-\frac{\mu_1}{\Delta}} + \begin{pmatrix} m_{2,A,G}(\frac{l}{l_0})^{-\frac{\mu_2}{\Delta}} \\ m_{3,C,T}(\frac{l}{l_0})^{-\frac{\mu_3}{\Delta}} \\ -m_{2,A,G}(\frac{l}{l_0})^{-\frac{\mu_2}{\Delta}} \\ -m_{3,C,T}(\frac{l}{l_0})^{-\frac{\mu_3}{\Delta}} \end{pmatrix} \tag{17}$$

where the parameters are defined as in (16) and

$$\mathbf{p}_K = \lim_{l\to\infty,\Delta>0} \mathbf{p}(l) = \begin{pmatrix} \pi_A(1-\frac{n_{1,R,Y}}{\pi_R}) - n_{2,A,G} \\ \pi_C(1+\frac{n_{1,R,Y}}{\pi_Y}) - n_{3,C,T} \\ \pi_G(1-\frac{n_{1,R,Y}}{\pi_R}) + n_{2,A,G} \\ \pi_T(1+\frac{n_{1,R,Y}}{\pi_Y}) + n_{3,C,T} \end{pmatrix}$$

with, for all $1 \le i \le 3$, $j \in \{A, C, R\}$, $k \in \{G, T, Y\}$,

$$m_{i,j,k} = \frac{\pi_j(\frac{r_j+r_k}{2} - \mu_i p_k(l_0)) - \pi_k(\frac{r_j+r_k}{2} - \mu_i p_j(l_0))}{(\pi_j+\pi_k)\mu_i}$$

$$n_{i,j,k} = \frac{(r_j+r_k)(\pi_j-\pi_k)}{2(\pi_j+\pi_k)\mu_i}.$$

2.4.3 Insertion–Deletion Only Formula

When the substitution probabilities are all equal to 0 ($\alpha = \beta = 0$), then the occurrence probability $\mathbf{p}(l)$ of each nucleotide A, C, G, and T at sequence length l becomes

$$\mathbf{p}(l) = \mathbf{p}_K + \begin{pmatrix} (p_A(l_0) - \frac{r_A}{\tau}) \\ (p_C(l_0) - \frac{r_C}{\tau}) \\ (p_G(l_0) - \frac{r_G}{\tau}) \\ (p_T(l_0) - \frac{r_T}{\tau}) \end{pmatrix} \left(\frac{l}{l_0}\right)^{-\frac{\tau}{\Delta}} \tag{18}$$

where the parameters are defined as in (16) and

$$\mathbf{p}_K = \lim_{l\to\infty,\Delta>0} \mathbf{p}(l) = \begin{pmatrix} \frac{r_A}{\tau} \\ \frac{r_C}{\tau} \\ \frac{r_G}{\tau} \\ \frac{r_T}{\tau} \end{pmatrix}.$$

Equation (18) will be used in the analysis of the GC content in complete bacterial genomes in Sect. 3.

2.5 *GC* Content Analysed with the *IDISL-HKY* Model

In order to analyse the GC content in complete bacterial genomes according to their genome lengths (Sect. 3) and to bring new insights to this biological debate, we derive the GC content formula with the *IDISL-HKY* model and its parameter estimation.

2.5.1 *GC Content Formula*

The analysis of GC content in genomes leads to the following assumptions with the *IDISL-HKY* model. As the DNA double helix is anti-parallel 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 nucleotide occurrence probability, the equilibrium distribution, and the nucleotide insertion rates satisfy

$$\begin{cases} p_C(l_0) = p_G(l_0), \qquad p_A(l_0) = p_T(l_0), \\ \pi_C = \pi_G, \qquad \pi_A = \pi_T, \\ r_C = r_G, \qquad r_A = r_T. \end{cases} \tag{19}$$

From Eq. (16), the GC probability $p_{G+C}(l)$, i.e. the sum of the occurrence probabilities of C and G, is after some algebraic manipulation

$$\begin{aligned} p_{G+C}(l) &= p_C(l) + p_G(l) \\ &= 2\left(\frac{\frac{r_C}{\tau} + \kappa\pi_C}{1+\kappa}\right) + 2\left(p_C(l_0) - \frac{\frac{r_C}{\tau} + \kappa\pi_C}{1+\kappa}\right)\left(\frac{l}{l_0}\right)^{-\frac{1+\kappa}{1-\frac{d}{\tau}}} \end{aligned}$$

where $\kappa = \frac{\alpha+\beta}{2\tau}$, α and β are the transition and transversion rates of matrix $\mathbf{M}_{HKY}$, respectively, π_C the equilibrium frequency of C, τ is the total insertion rate, r_C is the insertion rate of C, d is the deletion rate and $p_C(l_0)$ is the occurrence probability of C at sequence length l_0.

2.5.2 *Domain of Definition*

Under the constraints (19), the *IDISL-HKY* model can be written as the polynomial expression

$$p_{G+C}(l) = a + b\left(\frac{l}{l_0}\right)^{-c} \tag{20}$$

where $a = 2(\frac{\frac{r_C}{\tau}+\kappa\pi_C}{1+\kappa})$, $b = 2(p_C(l_0) - \frac{\frac{r_C}{\tau}+\kappa\pi_C}{1+\kappa})$, $c = \frac{1+\kappa}{1-\frac{d}{\tau}}$. Then

$$\begin{cases} d = -\frac{2\tau(1-c)+\alpha+\beta}{2c}, \\ r_C = r_G = \frac{a}{2}\tau + \frac{\alpha+\beta}{2}(\frac{a}{2} - \pi_C), \\ p_C(l_0) = p_G(l_0) = \frac{1}{2}(a+b). \end{cases} \tag{21}$$

From Eqs. (19) and (21), we derive the following domain of definition:

$$\begin{cases} 0 \le a \le 1, \\ -a \le b \le 1-a, \\ c \ge 1 \quad \text{if } \tau > d, \\ c < 0 \quad \text{if } \tau < d. \end{cases} \tag{22}$$

Indeed, from Eq. (21) line 2 and $\alpha, \beta, \pi_C, r_C \ge 0$, we derive $a \ge 0$. From Eq. (19) and by definition of $r_C + r_G \le \tau$, then $r_C = r_G \le \frac{\tau}{2}$. Similarly, we have $\pi_C = \pi_G \le \frac{1}{2}$ and $0 \le p_C(l_0) \le \frac{1}{2}$. From Eq. (21) line 2, $r_C \le \frac{\tau}{2}$ and $\pi_C \le \frac{1}{2}$, we deduce $a \le 1$.

From Eq. (21) line 3 and $0 \leq p_C(l_0) \leq \frac{1}{2}$, we obtain $-a \leq b \leq 1-a$. From Eq. (21) line 1

$$c = \frac{2\tau + \alpha + \beta}{2(\tau - d)}. \tag{23}$$

Then, when $\tau > d$, i.e. evolution dominated by nucleotide insertion ($\Delta > 0$), using $\alpha, \beta \geq 0$, we derive $c \geq 1$. If $c = 1$, Eq. (23) necessarily leads to $\alpha + \beta = -2d$. Thus, as $\alpha, \beta, d \geq 0$ then

$$\alpha = \beta = d = 0. \tag{24}$$

In contrast, when $\tau < d$, i.e. evolution dominated by nucleotide deletion ($\Delta < 0$), using $\tau \geq 0$, then $c < 0$.

3 Application: Analysis of the *GC* Content in Complete Bacterial Genomes

An example of application of the *IDISL* model is the analysis of the *GC* content in complete bacterial genomes. In 1956, it was found that bacterial guanine and cytosine content (*GC* content) varies between 25 % and 75 % (Lee et al. 1956). It was observed that large bacterial genomes tend to be *GC*-rich and small bacterial genomes tend to be *AT*-rich (Moran 1962; Rocha and Danchin 2002; Bastolla et al. 2004). This relationship was statistically studied. A linear relationship between *GC* content and genome length within a sub-group of bacterial genomes was reported (Wang et al. 2006). This linear relationship was pointed out as valid for aerobic, facultative, and microaerophilic species, but not for anaerobic prokaryotes (Musto et al. 2006). Thus, in our application, we distinguish the anaerobic bacterial genomes from the other ones (see data acquisition in Sect. 3.1). Recently, a positive linear correlation of *GC* content with genome size was observed in bacteria belonging to 433 species (Satapathy et al. 2010). Otherwise, mutational events may also explain the variation of bacterial *GC* content (Freese 1962; Sueoka 1962; Wang et al. 2006).

3.1 Data Acquisition and Description

In order to have significant statistical results with this debated relationship, we perform a massive statistical analysis based on all complete bacterial genomes currently available. Statistical features of complete bacterial genomes were obtained from the NCBI site: www.ncbi.nlm.nih.gov/genomes/lproks.cgi (January 2011). The length (number of nucleotides) and the *GC* content of each bacterial genome were extracted. We removed the duplicated data, i.e. genomes with the same length and *GC* content. The bacterial genomes are classified according to their taxonomic group $\mathcal{G}$ and their anaerobic property as mentioned before. All groups with a number of genomes greater than 30 are included in the statistical analysis. Table 1 gives the number of bacterial genomes for each taxonomic group included in the study and the abbreviations used in Figs. 2–5. The final data set consists in 894 genomes of length starting from 0.24 Mb to 10.47 Mb and represents in total 3460.15 Mb.

Table 1 Total number of bacterial genomes and abbreviation for each taxonomic group $\mathcal{G}$ studied in the statistical analysis

Taxonomic group $\mathcal{G}$		Abbreviation	Number $Card(\mathcal{G})$ of genomes $\mathcal{G}$
Non-anaerobic	*Actinobacteria*	*Actino N*	82
	Alphaproteobacteria	*Alpha N*	129
	Bacteroidetes/Chlorobi	*Bacteroidetes N*	32
	Betaproteobacteria	*Beta N*	76
	Cyanobacteria	*Cyano N*	39
	Firmicutes	*Firmicutes N*	183
	Gammaproteobacteria	*Gamma N*	252
Anaerobic	*Euryarchaeota*	*Euryarchaeota A*	41
	Firmicutes	*Firmicutes A*	60

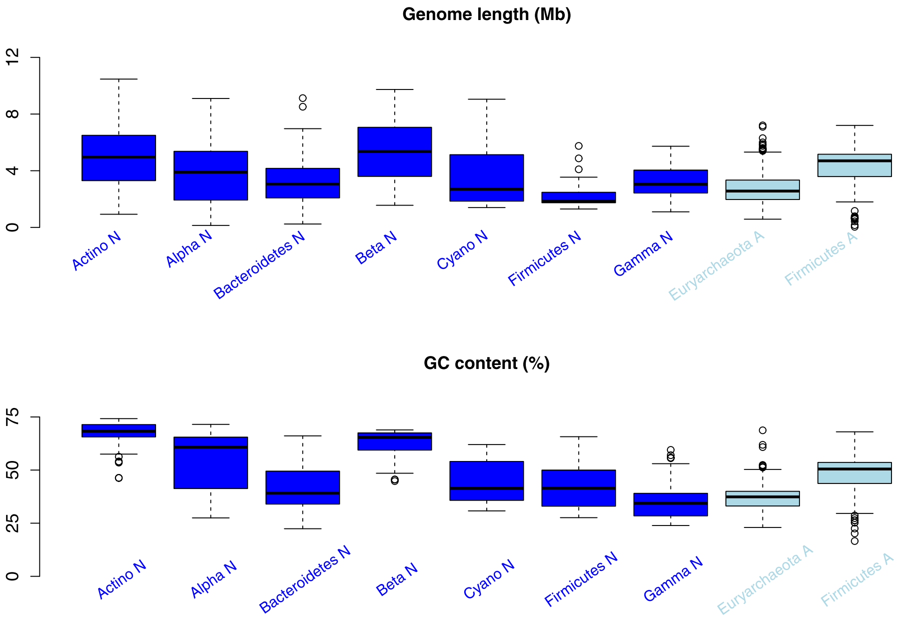

Fig. 2 Statistical features of the genome length (Mb) and GC content of the nine studied taxonomic groups of complete bacterial genomes. The *boxplots* show for each taxonomic group the distribution of the genome length (Fig. 2a) and GC content (Fig. 2b). The *horizontal bar* shows the median, the *box margins* represent the 25th and 75th percentiles, the *whiskers* indicate data within 2 times the interquartile range and the *circles* are outliers. The *dark boxes* represent the non-anaerobic bacteria (N), the *light ones*, the anaerobic bacteria (A). As expected, a large variability of CG content is observed in the 894 complete bacterial genomes with values varying from 16.6 % to 74.2 %

The boxplots of Fig. 2 show the distribution of the genome length (Fig. 2a) and GC content (Fig. 2b) for each taxonomic group of bacteria. The dark boxes represent the non-anaerobic bacteria (N), the light ones, the anaerobic bacteria (A).

The genome length median varies from 1.85 Mb (*Firmicutes N*) to 5.35 Mb (*Beta N*), i.e. almost a factor of 3. Strong variations are observed in CG content: from 16.6 % to 74.2 %. This massive dataset of complete bacterial genomes confirms the large variability of GC content which was already observed 50 years ago (Lee et al. 1956). Three taxonomic groups of bacteria have GC content median greater than 60 % (*Actino N*, *Alpha N*, *Beta N*) and two groups, lower than 40 % (*Gamma N*, *Euryarchae A*), *Actino N* having the highest GC content, and *Gamma N*, the lowest one.

3.2 Application Assumptions

The *IDISL* model allows the study of the relationship between GC content and genome length subject to substitution, insertion, and deletion of nucleotides. This application is based on two main assumptions.

Assumption 1 The residue frequencies in bacterial genomes are still transient.

Assumption 2 The species in a taxonomic group are subject to a common evolution process governed by the mutation parameters $([m_{ij}]_{i\neq j}, [r_i], d)$ up to a multiplicative constant, each species s having a specific evolution speed υ_s.

We assume that the currently observed nucleotide distributions and genome sizes are not stationary but are still varying (Assumption 1), i.e. the equilibrium distribution is not reached. Indeed, even if the currently observed genomes result in hundreds of millions of years old lineages, we consider that they are still subject to mutation (substitution, insertion, deletion) via vertical gene transfer (ortholog genes evolving from ancestral genes) and horizontal gene transfer in agreement with a dynamic view of the prokaryotic world, e.g. (Koonin and Wolf 2008). This active mutation process changes permanently the genome residues content, in particular the observed GC content which differs from the model equilibrium distribution.

Besides, we consider that the species in a taxonomic group are subject to a common mutation process (substitution, insertion, deletion) with different speeds (Assumption 2). Precisely, we assume that the ratios of the mutation parameters remain constant inside a taxonomic group but the overall rate of evolution varies within different species due to physical constraints (environment, specific mutations, cell size, etc.). Let us denote by υ_s the overall speed of evolution for a given species s. Then the associated mutation parameters are $([\upsilon_s m_{ij}]_{i\neq j}, [\upsilon_s r_i], \upsilon_s d)$ and the mutation parameters of all species inside a taxonomic group are all equals up to a multiplicative constant. For instance, for two species s and s', the mutation parameters satisfy $([\upsilon_s m_{ij}]_{i\neq j}, [\upsilon_s r_i], \upsilon_s d) = \frac{\upsilon_s}{\upsilon_{s'}}([\upsilon_{s'} m_{ij}]_{i\neq j}, [\upsilon_{s'} r_i], \upsilon_{s'} d)$. Using Proposition 1 for parameter scale (Eq. (11)), all species in a taxonomic group follow the same *IDISL* model. The currently observed genome length $n_s(t)$ of each species s depends on its overall evolution speed υ_s via parameter $\Delta = \upsilon_s(\tau - d)$ as defined in Eq. (8). Thus, the genomes of species in a taxonomic group, each one with a specific genome length $n_s(t)$, can be interpreted as being different evolution levels in the transient phase of a common mutation process defined by parameters $([m_{ij}]_{i\neq j}, [r_i], d)$. This

Table 2 Best fit parameters with the *IDISL-HKY* model $\widehat{p}_{G+C}(l) = \widehat{a} + \widehat{b}(\frac{l}{l_0})^{-\widehat{c}}$

Taxonomic group $\mathcal{G}$	$\widehat{a}$	$\widehat{b}$	$\widehat{c}$	l_0 (Mb)
Actino N	0.73	−0.28	1.00	0.93
Alpha N	0.75	−0.52	1.00	1.10
Bacteroidetes N	0.46	−0.27	1.00	0.24
Beta N	0.71	−0.31	1.26	1.56
Cyano N	0.49	−0.22	9.73	1.64
Firmicutes N	0.41	−0.16	1.34	0.73
Gamma N	0.55	−0.37	1.00	0.58
Euryarchaeota A	0.65	−0.35	1.14	1.30
Firmicutes A	0.33	0.14	4.43	2.10

is the reason why the *IDISL* model parameters are estimated for each taxonomic group.

3.3 Curve Fitting

The GC content distribution observed in the 894 bacterial genomes data clearly differs from the equiprobable distribution (Fig. 2b). For that reason, we use the *IDISL-HKY* model whose asymmetric substitution matrix allows non-equiprobable equilibrium distributions. Thus, the GC content observed in these nine bacterial taxonomic groups is analysed with the *IDISL-HKY* model using the GC modelling assumptions in Eq. (19) and $\Delta = \tau - d > 0$. It is modelled by the polynomial expression (20) derived from the *IDISL* model. The best curve fit is obtained by scanning over the power parameter c. Using the domain of definition given in Eq. (22) when $\Delta > 0$, then $c \geq 1$. Thus, parameter c was scanned from 1 to 15 with precision 10^{-3}. Then, for each value of c, the following linear regression $p_{G+C}(l) = a + bf(l) + \varepsilon(t)$, $\varepsilon(t) \sim \mathcal{N}(0, \sigma^2)$, is performed to explain the GC content by a transformation of the genome length $f(l) = (\frac{l}{l_0})^{-c}$ where l_0 is the smaller genome length of each taxonomic group and $\varepsilon(t)$ is a Gaussian noise $\mathcal{N}(0, \sigma^2)$. Note that parameters (a, b, c) are based on mutation parameter ratios only, thus avoiding identifiability issues due to Proposition 1.

Following Cook and Weisberg (1982), potentially influential points are determined by looking at the hat matrix $\mathbf{H}$ using the cut off value of $2q/n$ where $q = 2$ is the degree of freedom and n is the number of observations. Every point greater than the threshold is removed (in practice less than 5 %). For each taxonomic group, the set of parameters $(\widehat{a}, \widehat{b}, \widehat{c})$ minimizing the squared error between the observed GC content and the value $\widehat{a} + \widehat{b}(\frac{l}{l_0})^{-\widehat{c}}$ predicted by the inference procedure is retained. The estimating coefficients are given in Table 2.

For four taxonomic groups (non-anaerobic bacterial genomes: *Actino N*, *Alpha N*, *Bacteroidetes N*, *Gamma N*), the value of $\widehat{c}$ is equal to 1.00, meaning evolution without substitution or deletion, i.e. $\alpha = \beta = d = 0$ (Eq. (24)). For the five other taxonomic groups (three non-anaerobic bacterial genomes: *Beta N*, *Cyano N*, *Firmicutes N*, and two anaerobic: *Euryarchaeota A*, *Firmicutes A*), the value of $\widehat{c}$ is strictly greater than 1, implying $\alpha + \beta > 0$ and/or $d > 0$. In the latter case, in addition to insertion, gene evolution is subject to substitution and/or deletion. The value of $\widehat{b}$ is negative for eight among nine taxonomic groups, reflecting the classical relationship that GC content increases with genome length. For one taxonomic group (*Firmicutes A*), the value of $\widehat{b}$ is positive corresponding to a GC content decrease with genome length.

For each taxonomic group of bacterial genomes, Fig. 3 shows the dotplot of the GC content probability as a function of the genome length l and the best fit curve defined by the parameters $(\widehat{a}, \widehat{b}, \widehat{c})$ in Table 2 which are obtained with the *IDISL-HKY* model. Apart from the mentioned exception of *Firmicutes A*, the GC content increases with the genome length in eight taxonomic groups. The value of the asymptote of a fit curve is given by $\widehat{a}$ as $\lim_{l\to\infty} \widehat{p}_{G+C}(l) = \widehat{a} + \widehat{b}(\frac{l}{l_0})^{-\widehat{c}} = \widehat{a}$. The value of $\widehat{a}$ is below 0.5 for four taxonomic groups (*Bacteroidetes N, Cyano N, Firmicutes N, Firmicutes A*).

3.4 Biological Comparison and Interpretation of Results

In order to compare the non-linear fitting $\widehat{p}_{G+C} = \widehat{a} + \widehat{b}(\frac{l}{l_0})^{-\widehat{c}}$ obtained with the *IDISL-HKY* model with a linear fitting $\widetilde{p}_{G+C} = \widetilde{a} + \widetilde{b}l$ as proposed by Musto et al. (2006), we use the classical coefficient of determination $R^2 = 1 - \frac{SS_{\text{err}}}{SS_{\text{tot}}} = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$ where for each genome i in a given taxonomic group $\mathcal{G}$, y_i is the observed GC content percentage, $\widehat{y}_i$ is the predicted percentage and $\overline{y} = \frac{1}{Card(\mathcal{G})} \sum_i y_i$ is the average observed percentage. SS_{err} is the sum of squares of differences between the observed value y_i and the predicted value $\widehat{y}_i$, and SS_{tot} is the sum of squares of residuals. The coefficient R^2 of determination is a statistical measure of the quality of approximation of the regression line with the real data points. A coefficient $R^2 = 1.0$ indicates that the regression line perfectly fits the data while $R^2 = 0.0$ means that the model gives no improvement over the average fitting $\overline{y}$. Figure 4 shows the coefficients R^2 of determination obtained with the *IDISL-HKY* model and the linear model.

A very good fit ($R^2 > 0.70$) with the *IDISL-HKY* model is obtained for two non-anaerobic taxonomic groups (*Alpha N*, *Beta N*). In particular, for *Alpha N* containing 129 genomes, R^2 is equal to 0.85, i.e. approximately 85 % of the variation in the GC content can be explained by the genome length l using the *IDISL-HKY* model. The *IDISL-HKY* model confirms the increase of the GC content with the genome length in these two non-anaerobic taxonomic groups. For the other non-anaerobic groups (*Actino N*, *Bacteroidetes N*, *Cyano N*, *Firmicutes N*, *Gamma N*), R^2 with the *IDISL-HKY* model is comprised between 0.30 and 0.50. For the two anaerobic taxonomic groups, neither the *IDISL-HKY* model nor the linear model offers a good representation of the real data ($R^2 < 0.25$). The relationship between GC content and genome length for anaerobic bacteria might be subject to other constraints as already observed by Musto et al. (2006) when using a linear model. Thus, the modelling and interpretation of the GC content in anaerobic bacterial genomes remains still open.

In general, the non-linear modelling of the *IDISL-HKY* model outperforms the most recent linear modelling for all the non-anaerobic bacterial genomes, except for

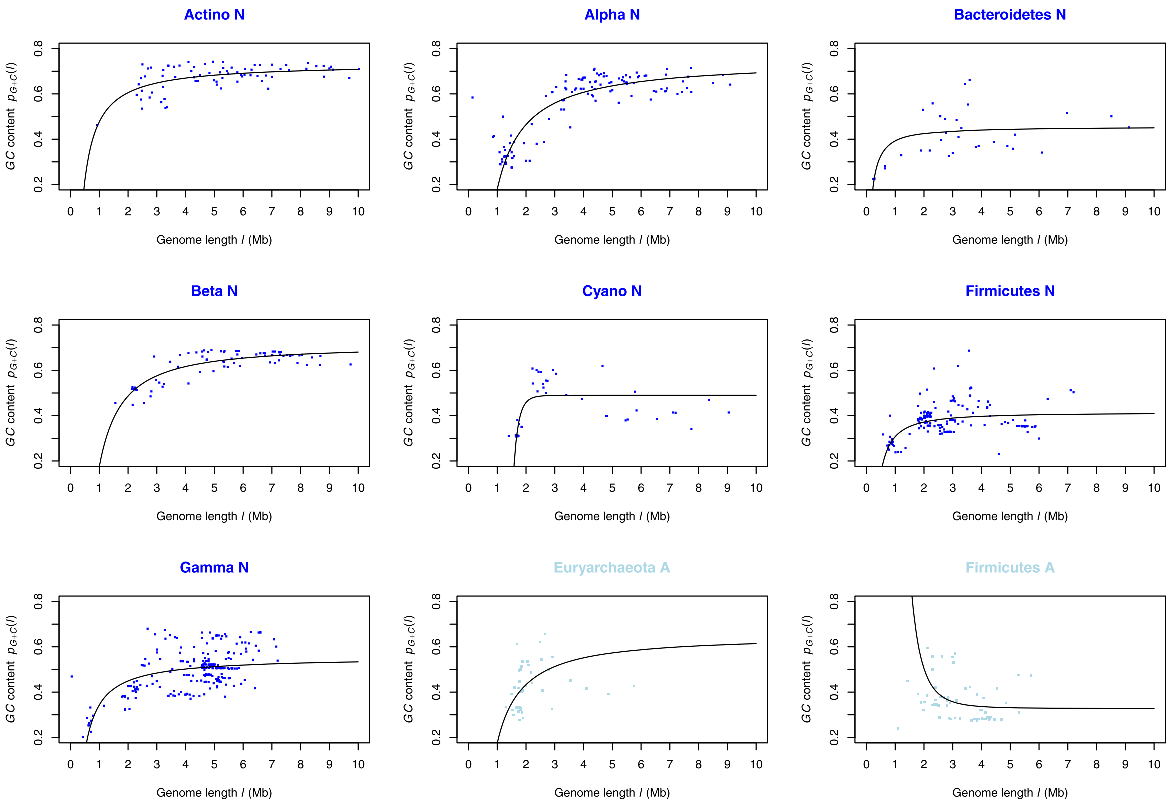

Fig. 3 GC content as a function of the genome length l and best fit curve $\widehat{p}_{G+C}(l)$ obtained with the *IDISL-HKY* model for the nine taxonomic groups of bacterial genomes. The x-axis represents the genome length l (Mb) and the y-axis, the GC content probability $p_{G+C}(l)$. The *dark data points* represent the non-anaerobic bacteria (N), the *light ones*, the anaerobic bacteria (A)

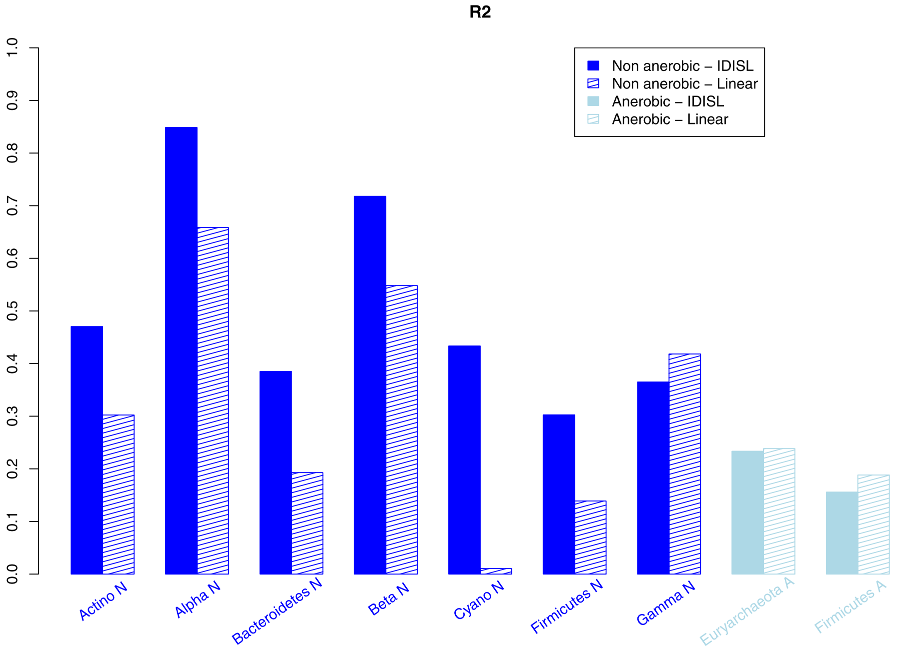

Fig. 4 Coefficients R^2 of determination obtained with the *IDISL-HKY* model $\widehat{p}_{G+C} = \widehat{a} + \widehat{b}(\frac{l}{l_0})^{-\widehat{c}}$ and the linear model $\widetilde{p}_{G+C} = \widetilde{a} + \widetilde{b}l$ for the nine taxonomic groups of bacterial genomes. The *dark boxes* represent the non-anaerobic bacteria (N), the *light ones*, the anaerobic bacteria (A)

Gamma N where the R^2 values have the same order of magnitude. Thus, the *IDISL-HKY* model which is based on the physical modelling of three mutation events (substitution, insertion, deletion) offers the best description of the GC content in bacterial genomes to date.

As already seen, four taxonomic groups have a value of $\widehat{c}$ equal to 1.00 (Table 2), thus implying $\alpha = \beta = d = 0$ (Eq. (24)). In order to analyse the weight of the insertion–deletion events for all taxonomic groups, we set the substitution parameters $\alpha = \beta = 0$ for all groups. This particular model can be studied using the *IDISL-HKY* insertion–deletion only formula (18). From Eq. (21), the deletion–insertion ratio $\frac{d}{\tau}$, the proportion $\frac{r_C}{\tau}$ of C nucleotide insertion over total nucleotide insertion rate and the initial probability $p_C(l_0)$ of nucleotide C are given as a function of the parameters (a, b, c)

$$\begin{cases} \frac{d}{\tau} = 1 - \frac{1}{c}, \\ \frac{r_C}{\tau} = \frac{r_G}{\tau} = \frac{a}{2}, \\ p_C(l_0) = p_G(l_0) = \frac{1}{2}(a + b). \end{cases}$$

Remark 6 The chosen application of the *IDISL* model is based on a large set of bacterial genomes whose lengths vary from 0.24 Mb to 10.47 Mb (Sect. 3.1). The order of magnitude difference is a factor of 40. Such a data analysis based on a wide range of genome lengths leads to a predominant insertion process: the orders of magnitude

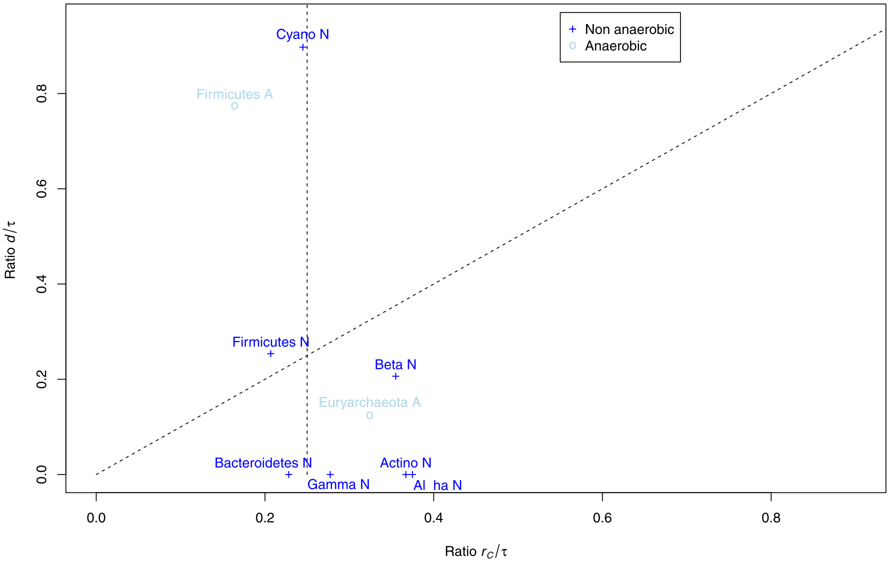

Fig. 5 *GC* deletion versus insertion rates for the nine taxonomic groups. The x-axis represents the $\frac{r_C}{\tau}$ ratio and the y-axis, the $\frac{d}{\tau}$ ratio where τ is the total nucleotide insertion rate, r_C is the insertion rate of nucleotide C and d is the total deletion rate. The *vertical dashed line* is associated with $\frac{r_C}{\tau} = 0.25$ and the bisector, with $d = r_C$. The *dark names* ('+') represent the non-anaerobic bacteria (N), the *light ones* ('o'), the anaerobic bacteria (A)

of the deletion d and the substitutions α and β are much smaller than the insertion. When the *IDISL* model is applied on genomes with similar lengths (with an order of magnitude difference of about 4), these three evolutionary processes have the same order of magnitude and their parameters differ from 0. As the *IDISL* model differs from the classical models by modelling insertion and deletion (in addition to substitution and independently), we have chosen an application in consequence.

Figure 5 plots the values of the ratios $\frac{d}{\tau}$ versus $\frac{r_C}{\tau}$ estimated from the values $(\widehat{a}, \widehat{b}, \widehat{c})$ (Table 2) for the nine taxonomic groups of bacterial genomes. The anaerobic groups (*Euryarchaeota A*, *Firmicutes A*) are plotted for indication but as already mentioned the interpretation must be cautious as their coefficients $R^2 \approx 0.20$ (Fig. 4).

With the *IDISL-HKY* model, the non-anaerobic bacterial genomes are divided into three classes according to the ratios $\frac{d}{\tau}$:

$$\begin{cases} \text{`Insertion' class: } d = 0 \text{ and} \\ \quad r_C > 0\text{: } \textit{Actino N}, \textit{Alpha N}, \textit{Bacteroidetes N}, \textit{Gamma N}, \\ \text{`Equilibrium' class: } d \approx r_C\text{: } \textit{Beta N}, \textit{Firmicutes N}, \\ \text{`Deletion' class: } d >> r_C\text{: } \textit{Cyano N}. \end{cases}$$

In the 'Insertion' class (*Actino N*, *Alpha N*, *Bacteroidetes N*, *Gamma N*), no deletion rate is detected as previously noticed. The 'Equilibrium' class (*Beta N*, *Firmicutes N*) corresponds to a deletion rate d of magnitude order of the insertion rate r_C of nucleotide C. The ‘Deletion’ class with one taxonomic group (*Cyano N*) corresponds to an intensive increase of the GC content starting from length l_0 (see the curve *Cyano N* in Fig. 3). The taxonomic groups located to the left of the vertical dashed line $\frac{r_C}{\tau}=0.25$ (*Bacteroidetes N*, *Cyano N*, *Firmicutes N*) have a GC content $\widehat{p}_{G+C}(l)$ below 0.5 for large genome lengths (as already noticed in Fig. 3). Indeed, the asymptote of the curve $\widehat{p}_{G+C}(l)$ is equal to $\widehat{a}=2\frac{r_C}{\tau}$. Their GC content increases but remains below 50 % even for large genome lengths.

4 Conclusion

We introduced here the *IDISL* model which provides an analytical expression of the residue occurrence probability according to the sequence length taking into account three independent evolutionary processes: substitution, insertion, and deletion of residues. This general property allows a great variety of evolution types to be considered, either by using the three processes simultaneously or by limiting or removing some of them. Thus, the *IDISL* model constitutes a new research approach for sequence analysis by studying the residue proportion according to sequence length under various evolutionary processes. It could also be extended for sequence alignment and phylogeny.

The *IDISL* model led to interesting and surprising results for the analysis of the GC content evolution in 894 complete bacterial genomes when considering that their nucleotide distribution and their length are not stationary (Assumption 1, Sect. 3.2) and that the species in a taxonomic group evolve with various speeds (Assumption 2, Sect. 3.2). From a theoretical point of view, the non-linear *IDISL* model for GC content is more general than the most recent model which is linear. Indeed, the *IDISL* model is defined by three parameters (a,b,c) (Eq. (20)), and thus allows a more refined analysis with one more degree of freedom. The linear model is a particular case of the *IDISL* model with parameter $c=-1$. In practice, the *IDISL* model for GC content outperforms the linear model on these bacterial genomes (Fig. 4). Moreover, contrary to the linear model, the *IDISL* model is based on a mathematical analysis of three evolution processes. Thus, it allows an interpretation of the model parameters in terms of substitution, insertion, and deletion. Its application to a large set of complete bacterial genomes led to the identification of three evolutionary classes: ‘Insertion’, ‘Equilibrium’, and ‘Deletion’ classes. In particular, in three taxonomic groups of non-anaerobic bacterial genomes (*Cyano N*, *Firmicutes N* and *Beta N* by decreasing values of deletion rate), deletion is an important evolution process for explaining the GC content evolution. Thus, the application of the *IDISL* model also confirms that the deletion process should be considered as a process independent of substitution and insertion in genome evolution.

The *IDISL* model is a simplified modelling of reality and, obviously, cannot completely reproduce the evolutionary process observed for each genome. A limit of the *IDISL* model is related to the fact that all the parameters (11 in total) are constant in time (six substitution rates, four insertion rates, and one deletion rate). For example, bacterial mutation rates change during the experimental colonization of the mouse gut (Giraud et al. 2001). Indeed, a high mutation rate may be initially beneficial because it allows faster adaptation, but this benefit disappears once adaptation is achieved. However, constant parameters are a useful working assumption and the *IDISL* model allows us to obtain a good representation of the GC content evolution in bacterial genomes. It should be stressed that the *IDISL* model with non-constant parameters has no analytical solution to our knowledge and such an extension could be investigated in future.

Acknowledgement We thank the three reviewers for their advice.

References

Aldous, D., & Fill, J. A. (2002). *Reversible Markov chains and random walks on graphs*. Berkeley: University of California.

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.

Bastolla, U., Moya, A., Viguera, E., & van Ham, R. C. (2004). Genomic determinants of protein folding thermodynamics in prokaryotic organisms. *J. Mol. Biol.*, *343*, 1451–1466.

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

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. London: Chapman & Hall.

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

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

Foerstner, K. U., von Mering, C., Hooper, S. D., & Bork, P. (2005). Environments shape the nucleotide composition of genomes. *EMBO Rep.*, *6*, 1208–1213.

Freese, E. (1962). On the evolution of base composition of DNA. *J. Theor. Biol.*, *3*, 82–101.

Giraud, A., Matic, I., Tenaillon, O., Clara, A., Radman, M., Fons, M., & Taddei, F. (2001). Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science*, *291*, 2606–2608.

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

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

Kelly, F. P. (1979). *Reversibility and stochastic networks*. Chichester: Wiley.

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

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA*, *78*, 454–458.

Koonin, E. V., & Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*, *36*, 6688–6719.

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

Lee, K. Y., Wahl, R., & Barbu, E. (1956). Contenu en bases puriques et pyrimidiques des acides désoxyribonucléiques des bactéries. *Ann. Inst. Pasteur*, *91*, 212–224.

Malthus, T. R. (2000). An essay on the principle of population. Library of Economics, Liberty, Fund, Inc.

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

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

Michel, C. J. (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. Evol.*, *21*, 529–540.

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

Moran, N. A. (1962). Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, *108*, 583–586.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valín, F., & Bernardi, G. (2006). Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochem. Biophys. Res. Commun.*, *347*, 1–3.

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

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

Rocha, E. P., & Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet.*, *18*, 291–294.

Satapathy, S. S., Dutta, M., & Ray, S. K. (2010). 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*, 232–242.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. USA*, *48*, 582–592.

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

Takahata, N., & Kimura, M. (1981). A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics*, *98*, 641–657.

Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, *10*, 512–526.

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

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

Wang, H. C., Susko, E., & Roger, A. J. (2006). On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. *Biochem. Biophys. Res. Commun.*, *342*, 681–684.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, *39*, 105–111.