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.