Bulletin of Mathematical Biology (2009) 71: 681–700
DOI 10.1007/s11538-008-9376-4

ORIGINAL ARTICLE

A Stochastic Model of Gene Evolution with Time Dependent Pseudochaotic Mutations

Jacques M. Bahi[a], Christian J. Michel[b,*]

[a] *LIFC—EA 4157, Université de Franche-Comté, IUT de Belfort, BP 527, 90016 Belfort Cedex, France*
[b] *Equipe de Bioinformatique Théorique, LSIIT (UMR CNRS-ULP 7005), Université Louis Pasteur de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France*

Received: 14 September 2007 / Accepted: 24 November 2008 / Published online: 6 February 2009

Abstract We develop here a new class of stochastic models of gene evolution in which a random subset of the 64 possible trinucleotides mutates at each evolutionary time t according to some time dependent substitution probabilities. Therefore, at each time t, the numbers and the types of mutable trinucleotides are unknown. Thus, the mutation matrix changes at each time t. This pseudochaotic model developed generalizes the standard model in which all the trinucleotides mutate at each time t. It determines the occurrence probabilities at time t of trinucleotides which pseudochaotically mutate according to 3 time dependent substitution parameters associated with the 3 trinucleotide sites. The main result proves that under suitable assumptions, this pseudochaotic model converges to a uniform probability vector identical to that of the standard model. Furthermore, an application of this pseudochaotic model allows an evolutionary study of the 3 circular codes identified in both eukaryotic and prokaryotic genes. A circular code is a particular set of trinucleotides whose main property is the retrieval of the frames in genes locally, i.e., anywhere in genes and particularly without start codons, and automatically with a window of a few nucleotides. After a certain evolutionary time and with particular time dependent functions for the 3 substitution parameters, precisely an exponential decrease in the 1st and 2nd trinucleotide sites and an exponential increase in the 3rd one, this pseudochaotic model retrieves the main statistical properties of the 3 circular codes observed in genes. Furthermore, it leads to a circular code asymmetry stronger than the standard model (non-pseudochaotic) and, therefore, to a better correlation with the genes.

Keywords Stochastic model · Pseudochaotic mutations · Time dependent mutations · Gene evolution · Trinucleotides · Mutation matrix · Substitution parameters · Circular code

*Corresponding author.
E-mail addresses: jacques.bahi@univ-fcomte.fr (Jacques M. Bahi), michel@dpt-info.u-strasbg.fr (Christian J. Michel).

1. Introduction

It is worthwhile noticing that the new model developed in this paper is based on a concept similar to the chaotic iteration model introduced by Chazan and Miranker (1969). This model was developed in the field of computer science in order to model asynchronous parallel computations with random faulty computers. It was also developed in the field of dynamical systems under the terminology of chaotic discrete dynamic systems (Robert, 1986). Nevertheless, chaotic systems are classically supposed to exhibit some specific properties such as sensitivity to initial conditions which is not the case in the previous cited research fields. Precisely, their authors determined the mathematical properties without which the system does not reach a stable state. In this paper, we follow the same approach and in order to avoid any confusion, we will call our model "pseudochaotic".

Models of gene evolution were initially developed on the basis of nucleotide information. The first model with 1 substitution parameter (substitutions α: all types) was proposed by Jukes and Cantor (1969) and generalized to 2 parameters (transitions α: $A \longleftrightarrow G$ and $C \longleftrightarrow T$, and transversions β: $A \longleftrightarrow C$, $A \longleftrightarrow T$, $C \longleftrightarrow G$ and $G \longleftrightarrow T$) (Kimura, 1980), 3 parameters (transitions α: $A \longleftrightarrow G$ and $C \longleftrightarrow T$, transversions β: $A \longleftrightarrow T$ and $C \longleftrightarrow G$, and transversions γ: $A \longleftrightarrow C$ and $G \longleftrightarrow T$) (Kimura, 1981), 5 parameters (transitions α: $A \longrightarrow G$ and $T \longrightarrow C$, transitions β: $G \longrightarrow A$ and $C \longrightarrow T$, transversions γ: $A \longleftrightarrow T$ and $C \longleftrightarrow G$, transversions δ: $A \longrightarrow C$ and $T \longrightarrow G$, and transversions ε: $C \longrightarrow A$ and $G \longrightarrow T$) (Takahata and Kimura, 1981) and 6 parameters (α: $A \longrightarrow C$, $A \longrightarrow G$, $T \longrightarrow C$ and $T \longrightarrow G$, α_1: $A \longrightarrow T$, α_2: $C \longrightarrow G$, β: $C \longrightarrow A$, $C \longrightarrow T$, $G \longrightarrow A$ and $G \longrightarrow T$, β_1: $T \longrightarrow A$ and β_2: $G \longrightarrow C$) (Kimura, 1981). DNA sequencing has revealed that the structure of the different genome regions are based on a variety of motifs of different sizes: dinucleotides, trinucleotides, oligonucleotides, either on a 2-letter alphabet, e.g., the purine/pyrimidine alphabet, or on the classical 4-letter alphabet. In order to study their evolutionary properties, nucleotide evolution models have been extended to motif evolution models, particularly to those of trinucleotides (Arquès and Michel, 1993) and dinucleotides (Arquès and Michel, 1995). The variety and the complexity of these motif models have then increased regularly. For example, with dinucleotide evolution models, a computer simulation approach (construction of simulated genes and then applications of random mutations) has been proposed by Fryxell and Zuckerkandl (2000) while a discrete version with time steps $\Delta t/L$ where Δt is the time increment and L, the length of the sequence, has been developed in Arndt et al. (2002). For phylogenetic inference, trinucleotide and site evolution models have been developed, e.g., a model with a 61×61 mutation matrix based on numerical solutions (Goldman and Yang, 1994) and its extensions to the nonsynonymous/synonymous substitution rate ratio (Yang et al., 2000; Yang and Swanson, 2002), a covarion-style model with a switch site process governed by a 2-state continuous-time Markov process ("on", "off") and an observable process governed be a second stationary and time-reversible Markov process based on a rate matrix (Tuffley and Steel, 1998), a gamma distribution model of site rate variation (Yang, 1994) and an extension to a "rate variation rate" with the constraint of being constant over sites and in time (Galtier, 2001).

The stochastic evolution models studied here allow the determination of the occurrence probabilities at time t of a set of trinucleotides which randomly mutates according to time dependent substitutions in the trinucleotide sites. This trinucleotide set can obviously be reduced to only one trinucleotide.

In these standard stochastic evolution models (nonpseudochaotic), the mutation matrix is constant in time, i.e., the matrix at initial time t_0 is the same at any time t. Zero elements and nonzero elements in this matrix are always in the same positions. Furthermore, nonzero elements in this matrix, called substitution parameters or substitution probabilities, can be either constant or time dependent. The most general analytical evolution model with constant parameters recently published is based on a 64×64 trinucleotide mutation matrix with 9 substitution parameters associated with the 3 types of substitutions in the 3 trinucleotide sites (Michel, 2007). It generalizes the models based on the 4×4 nucleotide mutation matrices, particularly (Jukes and Cantor, 1969; Kimura, 1980, 1981; Takahata and Kimura, 1981), and the 64×64 trinucleotide mutation matrices with 3 and 6 substitution parameters (Arquès et al., 1998; Frey and Michel, 2006). Evolution models with time dependent parameters were recently proposed to extend the constant models (Bahi and Michel, 2004).

We here develop a new class of stochastic models of gene evolution with time dependent random sets of mutable trinucleotides. In these pseudochaotic models, a random subset of the 64 possible trinucleotides mutates at each evolutionary time t according to some time dependent substitution probabilities while the other trinucleotides do not mutate. Thus, the mutation matrix changes in time. These pseudochaotic models generalize the standard ones (constant and time dependent). They differ from the previous ones such as the covarion-style model (Tuffley and Steel, 1998) in which the sites have an evolution modeled by 2 Markov processes, by the fact that in our approach the motifs have an evolution modeled by 1 Markov process. As the pseudochaotic models constitute a more general process of evolution compared to the other modes of evolution, it is mathematically interesting to study it as such. Otherwise, in genes, there are trinucleotides which do not mutate at each evolutionary time. The extreme case could be the stop codons which mutate rarely, perhaps even never. This biological observation is an obvious argument to develop evolution models of motifs and not only of sites.

Two types of results are presented in this paper:

(i) A mathematical model of gene evolution with time dependent pseudochaotic mutations is developed. It determines the occurrence probability at time t of a random subset of trinucleotides which mutates according to 3 time dependent substitution parameters associated with the 3 trinucleotide sites. A theorem proves that it converges to a uniform probability vector identical to that of the standard model (Section 2).

(ii) An application of this pseudochaotic model to the evolution of the 3 circular codes identified in both eukaryotic and prokaryotic genes allows the retrieval of their main statistical properties after a certain evolutionary time and with particular time dependent functions for the 3 substitution parameters. Furthermore, it also allows an evolutionary comparison between the standard and pseudochaotic models. Unexpectedly, the pseudochaotic model leads to a circular code asymmetry stronger than the standard model (nonpseudochaotic) and, therefore, to a better correlation with the genes (Section 3).

2. Mathematical model

The mathematical model will determine the occurrence probabilities $P(t)$ at time t of the 64 trinucleotides $\mathbb{T} = \{AAA, \ldots, TTT\}$ which mutate in a pseudochaotic way according

to 3 substitution probabilities $p^{(t)}$, $q^{(t)}$ and $r^{(t)}$ associated with the 3 trinucleotide sites, respectively.

By convention, the indices $i, j \in \{1, \ldots, 64\}$ represent the 64 trinucleotides $\mathbb{T}$ in alphabetical order. Let $P_i^{(t)}$ be the occurrence probability at time t of a trinucleotide i. At time $t+T$, the occurrence probability of the trinucleotide i is $P_i^{(t+T)}$ so that $P_i^{(t+T)} - P_i^{(t)}$ represents the probabilities of trinucleotides i which appear and disappear during the time interval T

$$P_i^{(t+T)} - P_i^{(t)} = \alpha T \sum_{j=1}^{64} P(j \to i) P_j^{(t)} - \alpha T P_i^{(t)},$$

where α is the probability that a trinucleotide is subjected to 1 substitution during T and where $P(j \to i)$ is the probability of the substitution of a trinucleotide j into a trinucleotide i. The probability $P(j \to i)$ is equal to 0 if the substitution is impossible (j and i differ by more than one nucleotide because T is assumed to be small enough that a trinucleotide cannot mutate twice in a row during T), otherwise it is given as a function of the 3 substitution rates $p^{(t)}$, $q^{(t)}$, and $r^{(t)}$. For example, with the trinucleotide AAA associated with $i = 1$, $P(CAA \to AAA) = P(GAA \to AAA) = P(TAA \to AAA) = p^{(t)}/3$, $P(ACA \to AAA) = P(AGA \to AAA) = P(ATA \to AAA) = q^{(t)}/3$, $P(AAC \to AAA) = P(AAG \to AAA) = P(AAT \to AAA) = r^{(t)}/3$, and $P(j \to AAA) = 0$ with $j \notin \{AAC, AAG, AAT, ACA, AGA, ATA, CAA, GAA, TAA\}$.

By rescaling time, we can assume that $\alpha = 1$, i.e., there is 1 substitution per trinucleotide per time interval. Then

$$P_i^{(t+T)} - P_i^{(t)} = T \sum_{j=1}^{64} P(j \to i) P_j^{(t)} - T P_i^{(t)}. \tag{1}$$

The formula (1) leads to

$$\lim_{T \to 0} \frac{P_i^{(t+T)} - P_i^{(t)}}{T} = \frac{dP_i^{(t)}}{dt} = \sum_{j=1}^{64} P(j \to i) P_j^{(t)} - P_i^{(t)}. \tag{2}$$

By considering the column vector $P^{(t)} = [P_i^{(t)}]_{1 \le i \le 64}$ of the 64 $P_i^{(t)}$ and the 64×64 mutation matrix $A^{(t)}$ of the 4,096 trinucleotide substitution probabilities, i.e., $A_{ij}^{(t)} = P^{(t)}(i \to j)$, the differential Eq. (2) can be represented by the following matrix equation:

$$\frac{dP^{(t)}}{dt} = A^{(t)} P^{(t)} - P^{(t)} = \left(A^{(t)} - I\right) P^{(t)}, \tag{3}$$

where I represents the identity matrix and with a given $P^{(0)}$, or similarly by

$$\forall i \in \{1, \ldots, 64\}, \forall t \geqslant 0, \quad \frac{dP_i^{(t)}}{dt} = \sum_{j=1}^{64} \left(A^{(t)} - I\right)_{ij} P_j^{(t)}. \tag{4}$$

The 64×64 mutation matrix $A^{(t)}$ can be defined by a 4×4 square block matrix whose 4 diagonal elements are formed by 4 16×16 identical square submatrices $B^{(t)}$ and whose

12 nondiagonal elements are formed by 12 16×16 identical square submatrices $(p^{(t)}/3)I$

$$A^{(t)} = \left(\begin{array}{c|cccc} & 1\cdots16 & 17\cdots32 & 33\cdots48 & 49\cdots64 \\ \hline 1\cdots16 & B^{(t)} & (p^{(t)}/3)I & (p^{(t)}/3)I & (p^{(t)}/3)I \\ 17\cdots32 & (p^{(t)}/3)I & B^{(t)} & (p^{(t)}/3)I & (p^{(t)}/3)I \\ 33\cdots48 & (p^{(t)}/3)I & (p^{(t)}/3)I & B^{(t)} & (p^{(t)}/3)I \\ 49\cdots64 & (p^{(t)}/3)I & (p^{(t)}/3)I & (p^{(t)}/3)I & B^{(t)} \end{array}\right).$$

The index ranges $\{1,\ldots,16\}$, $\{17,\ldots,32\}$, $\{33,\ldots,48\}$ and $\{49,\ldots,64\}$ are associated with the trinucleotides $\{AAA,\ldots,ATT\}$, $\{CAA,\ldots,CTT\}$, $\{GAA,\ldots,GTT\}$ and $\{TAA,\ldots,TTT\}$, respectively. The 16×16 square submatrix $B^{(t)}$ can again be defined by a 4×4 square block matrix whose 4 diagonal elements are formed by 4 4×4 identical square submatrices $C^{(t)}$ and whose 12 nondiagonal elements are formed by 12 4×4 identical square submatrices $(q^{(t)}/3)I$

$$B^{(t)} = \begin{pmatrix} C^{(t)} & (q^{(t)}/3)I & (q^{(t)}/3)I & (q^{(t)}/3)I \\ (q^{(t)}/3)I & C^{(t)} & (q^{(t)}/3)I & (q^{(t)}/3)I \\ (q^{(t)}/3)I & (q^{(t)}/3)I & C^{(t)} & (q^{(t)}/3)I \\ (q^{(t)}/3)I & (q^{(t)}/3)I & (q^{(t)}/3)I & C^{(t)} \end{pmatrix}.$$

Finally, the 4×4 square submatrix $C^{(t)}$ is equal to

$$C^{(t)} = \begin{pmatrix} 0 & r^{(t)}/3 & r^{(t)}/3 & r^{(t)}/3 \\ r^{(t)}/3 & 0 & r^{(t)}/3 & r^{(t)}/3 \\ r^{(t)}/3 & r^{(t)}/3 & 0 & r^{(t)}/3 \\ r^{(t)}/3 & r^{(t)}/3 & r^{(t)}/3 & 0 \end{pmatrix}.$$

The matrix $A^{(t)}$ is stochastic when $p^{(t)} + q^{(t)} + r^{(t)} = 1$.

2.1. Time dependent standard evolution model

In the time dependent standard evolution model (Bahi and Michel, 2004), it is supposed that for a sampling $t_0 < t_1 < \cdots < t_n$, $A^{(t)}$ is a constant matrix on the time interval $[t_k, t_{k+1}]$, i.e.,

$$A^{(t)} = A_k, \quad \forall t \in [t_k, t_{k+1}].$$

This equation means that although the mutation matrix $A^{(t)}$ is not constant in the entire time interval, there exists (sufficiently small) periods of time in which the mutation factors are constant.
With this realistic hypothesis, the Eq. (3) can be written as follows:

$$\frac{dP^{(t)}}{dt} = (A_k - I)P^{(t)}, \quad \forall t \in [t_k, t_{k+1}].$$

For $t \in [t_k, t_{k+1}]$, the probability $P^{(t)}$ is then computed by the formula (Bahi and Michel, 2004)

$$P^{(t)} = e^{(A_k - I)(t-t_k)} e^{(A_{k-1}-I)(t_k - t_{k-1})} \cdots e^{(A_1 - I)(t_2 - t_1)} e^{(A_0 - I)(t_1 - t_0)} P^{(0)}, \tag{5}$$

where $e^{(A_k - I)(t-t_k)}$ is the exponential of $(A_k - I)(t - t_k)$ and $P^{(0)}$, the initial vector.

The formula (5) with the 64 initial trinucleotide probabilities $P^{(0)}$ before the substitution process ($t = 0$) and the product of exponential matrices $(A_k - I)(t - t_k)$ determines the 64 trinucleotide probabilities $P^{(t)}$ after t substitutions as a function of the 3 time dependent substitution parameters $p^{(t)}$, $q^{(t)}$, and $r^{(t)}$.

2.2. *Time dependent pseudochaotic evolution model*

The new time dependent evolution model that we call pseudochaotic will generalize the standard model to a random subset of the 64 possible trinucleotides which mutates at each evolutionary time t according to some probabilities. Therefore, the numbers and the types of mutable trinucleotides at each time t are unknown. This idea of chaotic gene evolution comes from the notion of chaotic iterative algorithms introduced by Chazan and Miranker (1969).

Let us discretize the time interval $[0, t_n]$ as follows $0 = t_0 < t_1 < \cdots < t_n$. For the sake of simplicity, we will suppose that $t_{k+1} - t_k = h$ where h is a constant real value. We denote by $P_i^{(t_k)}$ the probability at the time step t_k of the trinucleotide i. We denote by $J(k) \subseteq \{1, \ldots, 64\}$ the set of trinucleotide indices mutating at t_k. If the trinucleotide $i \notin J(k)$, then i does not mutate at t_k and i is called "nonmutable trinucleotide". If the trinucleotide $i \in J(k)$, then i mutates at t_k and i is called "mutable trinucleotide". The $J(k)$ parameter randomly chooses a subset of trinucleotides at each time step t_k according to a uniform distribution, i.e., the numbers and the types of mutable trinucleotides randomly vary at each t_k. The mutation matrix changes at each time step t_k according to the trinucleotides which mutate or not. We denote by $A^{(k)}$ the mutation matrix at the time step t_k. In the sequel, we will use the explicit Euler approximation of the derivative, i.e., $dP_i^{(t_k)}/dt = (P_i^{(t_k)} - P_i^{(t_{k-1})})/h$ with a sufficiently small time step h. Then we obtain the following model from the Eq. (3):

$$\begin{cases} P_i^{(t_k)} = P_i^{(t_{k-1})} & \text{if } i \notin J(k), \\ P_i^{(t_k)} = h \displaystyle\sum_{j=1}^{64} (A^{(k)} - I)_{ij} P_j^{(t_{k-1})} + P_i^{(t_{k-1})} & \text{if } i \in J(k). \end{cases} \tag{6}$$

Remark 1. The pseudochaotic model is a generalization of the standard model by taking $J(k) = \{1, \ldots, 64\}$ for all k.

2.2.1. *Mutation matrix properties*

A mutation matrix A with $A_{ii} = 0$ (zero elements on the main diagonal) implies that the probability that the trinucleotide i does not mutate is equal to 0, i.e., the trinucleotide i cannot be conserved and always mutates according to some substitution probabilities A_{ij}, $j \neq i$. A mutation matrix A with $A_{ii} \neq 0$ (nonzero elements on the main diagonal) implies that the probability that the trinucleotide i does not mutate is not null and can be conserved according to the substitution probability A_{ii}.

The pseudochaotic model leads to some new properties with the mutation matrix A compared to the standard models. For any time step t_k in the pseudochaotic model, the mutation matrix $A^{(k)}$ is stochastic and symmetric and $A^{(k)}_{ii} = 1 - \sum_{j=1, j\neq i}^{64} A^{(k)}_{ij}$. The probability that a trinucleotide i does not mutate at a step t_k is not null. In addition, it varies according to the trinucleotides which do not mutate at the step t_k. Thus, the matrix $A^{(k)}$ changes at each step t_k. The 2 laws below explain why the stochasticity and the symmetry of the matrix $A^{(k)}$ are preserved in time with both mutable and nonmutable trinucleotides.

(i) Matrix conservation law for nonmutable trinucleotides ($i \notin J(k)$). The probability of a nonmutable trinucleotide i is identical at the steps t_{k-1} and t_k, i.e., $P_i^{(t_k)} = P_i^{(t_{k-1})}$. As this trinucleotide i cannot mutate into a trinucleotide $j \neq i$, then $A^{(k)}_{ij} = 0\ \forall j \neq i$, otherwise its probability $P_i^{(t_k)}$ would decrease, in contradiction to the conservation law. Similarly, no trinucleotide $j \neq i$ can mutate into the trinucleotide i and $A^{(k)}_{ji} = 0$ $\forall j \neq i$, otherwise its probability $P_i^{(t_k)}$ would increase, in contradiction to the conservation law. Then $A^{(k)}_{ij} = A^{(k)}_{ji} = 0\ \forall j \neq i$ and $A^{(k)}_{ii} = 1$, i.e., the line i and the column i in the matrix $A^{(k)}$ associated with a nonmutable trinucleotide i are equal to 0 except their ith diagonal element which is equal to 1. Furthermore, this matrix conservation law keeps the symmetry of the matrix $A^{(k)}$ for a nonmutable trinucleotide. Finally, this property can obviously be extended to a set $\overline{J(k)}$ of nonmutable trinucleotides at a step t_k, $\overline{J(k)}$ being the complement of the set $J(k)$ of mutable trinucleotides at t_k.

(ii) Matrix conservation law for mutable trinucleotides ($i \in J(k)$). As some trinucleotides do not mutate at a step t_k, for each mutable trinucleotide i, the probability that it does not mutate is corrected to $A^{(k)}_{ii} = 1 - \sum_{j=1, j\neq i}^{64} A^{(k)}_{ij}$. This last equality leads to a stochastic matrix $A^{(k)}$. Furthermore, as only the diagonal elements are corrected and as the nonmutable elements are set to a probability equal to 0, the matrix $A^{(k)}$ is also symmetric.

The symmetry of mutation matrix $A^{(k)}$ with the mutable trinucleotides is a classical assumption of the nucleotide and trinucleotide mutation matrices in most standard models (Jukes and Cantor, 1969; Kimura, 1980, 1981; Takahata and Kimura, 1981; Arquès et al., 1998; Frey and Michel, 2006; Michel, 2007), etc.

2.2.2. Probability vector at each time step

Proposition 1. *In the time dependent pseudochaotic evolution model, $P^{(t_k)}$ is a probability vector at each time step t_k.*

Proof: Let us define the matrix $B^{(k)} = hA^{(k)} + (1-h)I$ so that $(B^{(k)}x)_i = h(\sum_{j=1}^{64}(A^{(k)} - I)_{ij}x_j) + x_i$. Then the pseudochaotic model can be written as follows: at a time step $t_k \in \mathbb{R}$

$$P^{(t_k)} = B^{(k)} P^{(t_{k-1})}. \tag{7}$$

As $B^{(k)}$ is a stochastic matrix, the proposition 1 is a direct consequence of the Eq. (7). □

2.2.3. Convergence analysis

In order to prove the convergence of the time dependent pseudochaotic evolution model, we introduce the following definitions. This pseudochaotic model can be associated with a sequence of undirected connected graphs $G^{(k)} = (V^{(k)}, E^{(k)})$ where $V^{(k)}$ is the set of

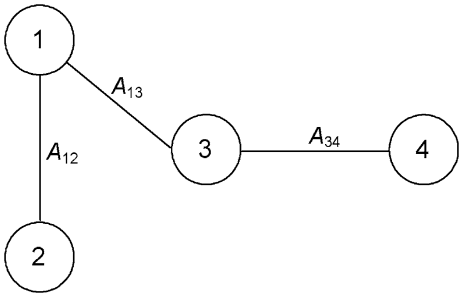

Fig. 1 An example of a fixed mutation probability graph.

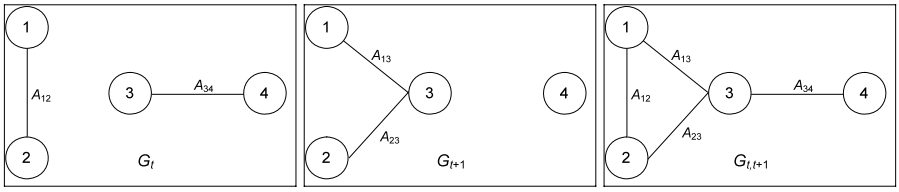

Fig. 2 An example of a superposed mutation probability graph $G_{t,t+1}$ in the pseudochaotic model.

vertexes and $E^{(k)}$ is the set of edges, $E^{(k)} \subseteq V^{(k)} \times V^{(k)}$. Each trinucleotide is a vertex of the graph. Two trinucleotides i and j are connected by an edge $\{i, j\} \in E^{(k)}$ if and only if the probability $P(j \to i) \neq 0$ at the time step k. By definition, each vertex is labeled from 1 to $|V| = 64$. Let $|E|$ be the number of mutation edges. As an example, Fig. 1 shows a fixed mutation probability graph $G = (V, E)$ with $|V| = 4$ trinucleotides, $V = \{1, 2, 3, 4\}$, and $|E| = 3$ mutation edges, $E = \{(1, 2), (1, 3), (3, 4)\}$.

It should be noticed that in this example the coefficients A_{ij} of the mutation matrix A are supposed to be constant, but in the pseudochaotic model, these elements change during time.

Definition 1. A superposed mutation probability graph $G_{t,t+m}$ between 2 times t and $t + m$ is the graph that shows all the edges corresponding to the mutable trinucleotides between the times t and $t + m$. Figure 2 shows an example of a superposed mutation probability graph $G_{t,t+1}$ in the pseudochaotic model.

Condition 1. For each time t, there is a time $t + m$ such that the superposed mutation probability graph $G_{t,t+m}$ is a connected graph.

Remark 2. It should be noted that the condition 1 is not restrictive. It allows the mutation probability graphs to be never connected, i.e., their corresponding incidence matrices may be never irreducible.

At a time t, let us define the adjacency matrix $I^{(t)}$ of a mutation matrix $M^{(t)}$ by

$$I_{ij}^{(t)} = \begin{cases} 1 & \text{if the trinucleotide } i \text{ mutates into the trinucleotide } j \\ & \text{or into itself } (i = j), \\ 0 & \text{otherwise.} \end{cases}$$

Let us define the adjacency matrix of the superposed mutation probability graph $G_{t,t+m}$ between the times t and $t+m$ by

$$I_{ij}^{(t,t+m)} = I_{ij}^{(t)} + I_{ij}^{(t+1)} + \cdots + I_{ij}^{(t+m-1)} + I_{ij}^{(t+m)}$$

with the classical boolean operations $1+1=1$; $1+0=0+1=1$ and $0+0=0$.

Lemma 1. *If the superposed mutation probability graph $G_{t,t+m}$ of $m+1$ mutation matrices $M^{(t)}, M^{(t+1)}, \ldots, M^{(t+m-1)}, M^{(t+m)}$ is connected then the matrix $M = M^{(t+m)} M^{(t+m-1)} \cdots M^{(t+1)} M^{(t)}$ is an irreducible matrix.*

Proof: The adjacency matrix of M is $I = I^{(t+m)} \times \cdots \times I^{(t)}$ and the adjacency matrix of $G_{t,t+m}$ is $I^{(t)} + \cdots + I^{(t+m)}$. By remarking that $I^{(t)} + \cdots + I^{(t+m)} \leq I^{(t+m)} \times \cdots \times I^{(t)}$, we deduce that if any entry of $G_{t,t+m}$ is 1, then the corresponding entry of I is also 1. Therefore, if $G_{t,t+m}$ is connected, then M is irreducible. □

Theorem 1. *Under condition* 1 *and for all $p, q, r \in]0,1[$, the time dependent pseudo-chaotic evolution model converges to the uniform probability vector* $(1/64, \ldots, 1/64)^T$.

Proof: (i) Sufficient condition.

By hypothesis, there always exists a time step $t \in \mathbb{N}$ such that the superposed mutation probability graph is connected. So, according to Lemma 1, for any time step $t > t_0$, there always exists irreducible matrices $T^{(p_i)}$ such that

$$M^{(t)} M^{(t-1)} \cdots M^{(2)} M^{(1)} = M^{(t)} M^{(t-1)} \cdots M^{(t-L+1)} M^{(t-L)} \times T^{(p_\alpha)} T^{(p_{\alpha-1})} \cdots T^{(p_2)} T^{(p_1)},$$

where

$$\begin{aligned} T^{(p_\alpha)} &= M^{(t-L-1)} \times \cdots \times M^{(L_\alpha)}, \\ T^{(p_{\alpha-1})} &= M^{(L_\alpha - 1)} \times \cdots \times M^{(L_{\alpha-1})}, \\ &\cdots \\ T^{(p_1)} &= M^{(L_2-1)} \times \cdots \times M^{(L_1)}, \end{aligned}$$

where L and L_i are finite integers. When t tends to infinity, then α tends also to infinity and $\lim_{\alpha \to \infty} p_\alpha = \infty$.

It can easily be seen that the matrices $T^{(p_i)}$ are doubly stochastic (not necessarily symmetric), not bipartite (-1 is not an eigenvalue) and irreducible. Therefore, if a matrix

A is irreducible and not bipartite, then there exists l such that A^l is a positive matrix. These deductions imply that

$$\forall i, \quad \lim_{k\to\infty}\left(T^{(p_i)}\right)^k = Q = \frac{1}{64}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}.$$

Let $\gamma^{(p_j)}$ be the second largest eigenvalue of $T^{(p_j)}$ then

$$0 \le \gamma^{(p_j)} < 1.$$

Recall that for a matrix M we have

$$\|M\|_2 = \sqrt{\lambda},$$

where λ is the maximum eigenvalue of $M^T M$, M^T being the matrix transpose of M. Therefore, if M is a doubly stochastic matrix, then

$$\|M\|_2 = \sqrt{\lambda} = 1. \tag{8}$$

Let $x^* = c(1,\dots,1)^T$, $c > 0$, be a uniform probability vector then

$$\begin{aligned}
\left\|x^{(t+1)} - x^*\right\|_2 &= \left\|M^{(t)} \times \cdots \times M^{(t-L)} T^{(p_\alpha)} \times \cdots \times T^{(p_1)} x^0 - x^*\right\|_2 \\
&\le \left\|M^{(t)}\right\|_2 \times \cdots \times \left\|M^{(t-L)}\right\|_2 \left\|T^{(p_\alpha)} \times \cdots \times T^{(p_1)} x^0 - x^*\right\|_2 \\
&\le \left\|T^{(p_\alpha)}\left(T^{(p_{\alpha-1})} \times \cdots \times T^{(p_1)}\right)x^0 - T^{(p_\alpha)} x^*\right\|_2 \\
&\le \gamma^{(p_\alpha)} \left\|T^{(p_{\alpha-1})}\left(T^{(p_{\alpha-2})} \times \cdots \times T^{(p_1)}\right)x^0 - T^{(p_{\alpha-1})} x^*\right\|_2 \\
&\le \gamma^{(p_\alpha)} \times \cdots \times \gamma^{(p_1)} \left\|x^0 - x^*\right\|_2.
\end{aligned}$$

So

$$\lim_{t\to\infty}\left\|x^{(t+1)} - x^*\right\|_2 \le \lim_{\alpha\to\infty} \gamma^{(p_\alpha)} \times \cdots \times \gamma^{(p_1)} \left\|x^{(0)} - x^*\right\|_2.$$

Since the number of edges of mutation probability graphs is finite, the number of mutation probability graphs is finite. So, there exists k such that for all $j \in \mathbb{N}$, $\gamma^{(p_j)} \le \gamma^{(p_k)} < 1$ and

$$\lim_{\alpha\to\infty} \gamma^{(p_\alpha)} \times \cdots \times \gamma^{(p_1)} \le \lim_{\alpha\to\infty} \left(\gamma^{(p_k)}\right)^\alpha = 0.$$

The last inequality implies that

$$\lim_{t\to\infty}\left\|x^{(t)} - x^*\right\|_2 = 0.$$

Therefore,

$$\forall i \in \{1,\dots,n\}, \quad x_i^{(t)} \to x_i^*.$$

In the case of a 64×64 mutation matrix, we deduce that

$$x^* = \frac{1}{64}(1, \ldots, 1)^T.$$

(ii) Necessary condition.

It is obvious. Indeed, if the condition 1 is not satisfied, then a set of trinucleotides will be forever isolated, and thus its probability will never reach the uniform probability. □

Remark 3. Theorem 1 cannot be derived from the known result of Wolfowitz on stochastic matrices (Wolfowitz, 1963). Indeed, in our model, words (products of $B^{(k)}$) do not have to be indecomposable and aperiodic (SIA). Precisely, in our model, it is possible that $\lim_{n\to\infty} W^n = \lim_{n\to\infty}(B^{(k+n)} \cdots B^{(k)})$ does not exist.

Remark 4. Theorem 1 is valid for an arbitrary number of trinucleotides. Therefore, it generalizes the result of the standard model given in Section 2.1.

Remark 5. It is possible to have nonuniform stationary probabilities if Condition 1 is not verified, i.e., a subset $J(k)$ of trinucleotides definitively stops mutating after a certain time step.

3. An application of the pseudochaotic evolution model to circular codes

3.1. Circular codes X_0, X_1, and X_2 identified in eukaryotic and prokaryotic genes

3.1.1. Identification

Definition 2. By convention, the reading frame established by a start codon belonging to $\mathbb{T}_{\text{start}} = \{ATG, GTG, TTG\}$ is the frame 0, and the frames 1 and 2 are the reading frame 0 shifted by 1 and 2 nucleotides in the 5′–3′ direction, respectively.

In 1996, a simple occurrence study of the 64 trinucleotides $\mathbb{T}$ in the 3 frames of genes showed that the trinucleotides are not uniformly distributed in these 3 frames. By excluding the 4 trinucleotides with identical nucleotides $\mathbb{T}_{\text{id}} = \{AAA, CCC, GGG, TTT\}$, the same 3 subsets X_0, X_1, and X_2 of 20 trinucleotides are found in the frames 0, 1, and 2, respectively, of 2 large and different gene populations (protein coding regions) of eukaryotes (26, 757 sequences, 11, 397, 678 trinucleotides) and prokaryotes (13, 686 sequences, 4, 708, 758 trinucleotides) (Arquès and Michel, 1996). These 3 trinucleotide subsets are obviously (law of large numbers) retrieved with the actual statistical studies (results not shown). The subset X_0 of 20 trinucleotides in frame 0 is $X_0 = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$ (Arquès and Michel, 1996).

Unexpectedly, the 2 subsets X_1 and X_2 of 20 trinucleotides found in the frames 1 and 2, respectively, of these 2 gene populations are related to X_0 thanks to the permutation property.

Notation 1. The *letters* (or *nucleotides* or *bases*) define the genetic alphabet $\mathcal{A} = \{A, C, G, T\}$. The set of *nonempty words* (resp. *words*) over $\mathcal{A}$ is denoted by $\mathcal{A}^+$ (resp. $\mathcal{A}^*$). The set of the 64 words of length 3 (or *trinucleotides* or *triletters*) is denoted by $\mathbb{T}$. Let $w_1 w_2$ be the concatenation of the 2 words w_1 and w_2.

Definition 3. The (left circular) permutation map $\mathcal{P}$: $\mathbb{T} \to \mathbb{T}$ permutes circularly each trinucleotide $w_0 = l_0 l_1 l_2$ as follows $\mathcal{P}(w_0) = w_1 = l_1 l_2 l_0$, e.g., $\mathcal{P}(AAC) = ACA$. The kth iterate of $\mathcal{P}$ is denoted by $\mathcal{P}^k$, e.g., $\mathcal{P}^2(w_0) = w_2 = l_2 l_0 l_1$. This permutation map on words is naturally extended to word sets: a permuted trinucleotide set is obtained by applying the permutation map $\mathcal{P}$ to all its trinucleotides.

Property 1. Permutation: $\mathcal{P}(X_0) = X_1$ and $\mathcal{P}^2(X_0) = X_2$ (X_0 generates X_1 by one permutation and X_2 by another permutation).

Remark 6. Two trinucleotides u and v are conjugate if there exist two words s and t such that $u = st$ and $v = ts$. Therefore, if u and v satisfy $\mathcal{P}^k(u) = v$ for some k, then u and v are conjugate.

Unexpectedly, these 3 trinucleotide subsets X_0, X_1, and X_2 are also related to each other by the complementarity property.

Definition 4. The complementarity map $\mathcal{C}$: $\mathcal{A}^+ \to \mathcal{A}^+$ is defined by $\mathcal{C}(A) = T$, $\mathcal{C}(T) = A$, $\mathcal{C}(C) = G$ and $\mathcal{C}(G) = C$ and by $\mathcal{C}(uv) = \mathcal{C}(v)\mathcal{C}(u)$ for all $u, v \in \mathcal{A}^+$, e.g., $\mathcal{C}(AAC) = GTT$. This complementarity map on words is naturally extended to word sets: a complementary trinucleotide set is obtained by applying the complementarity map $\mathcal{C}$ to all its trinucleotides.

Property 2. Complementarity: $\mathcal{C}(X_0) = X_0$ (X_0 is self-complementary) and, $\mathcal{C}(X_1) = X_2$ and $\mathcal{C}(X_2) = X_1$ (X_1 and X_2 are complementary to each other).

Remark 7. This complementarity map $\mathcal{C}$ is associated to the property of the complementary and antiparallel double helix (one DNA strand chemically oriented in a 5′–3′ direction and the other DNA strand, in the opposite 3′–5′ direction). Thanks to Property 2, the circular code X_0 and its 2 permuted circular codes X_1 and X_2 can exist simultaneously in a DNA double helix: X_0 in a given DNA strand can be paired with X_0 in the antiparallel complementary DNA (cDNA) strand, X_1 (X_0 shifted by 1 nucleotide in the 5′–3′ direction) in a given DNA strand can be paired with X_2 (X_0 shifted by 2 nucleotides in the 5′–3′ direction) in the cDNA strand and X_2 (in a given DNA strand) can similarly be paired with X_1 (in the cDNA strand).

These 3 trinucleotide subsets X_0, X_1, and X_2 present several other strong biomathematical properties, particularly the fact that they are circular codes. We recall their main properties which will be involved in the standard and pseudochaotic evolution models. A recent review of circular codes in genes details the research context, the history, and their different properties (Michel, 2008).

Definition 5. Code: A set X of words is a code if, for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $n, m \geq 1$, the condition $x_1 \cdots x_n = x'_1 \cdots x'_m$ implies $n = m$ and $x_i = x'_i$ for $i = 1, \ldots, n$.

The set $\mathbb{T}$ itself is a code. More precisely, it is a uniform code (Berstel and Perrin, 1985). Consequently, any nonempty subset of $\mathbb{T}$ is a code.

Definition 6. Circular code: A code X is circular if, for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $n, m \geq 1$, $r \in \mathcal{A}^*$ and $s \in \mathcal{A}^+$, the conditions $sx_2 \cdots x_n r = x'_1 \cdots x'_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ (empty word) and $x_i = x'_i$ for $i = 1, \ldots, n$.

Remark 8. $\mathbb{T}$ is obviously not a circular code.

Definition 7. Maximal circular code. A circular code X is maximal if, for each $y \in \mathbb{T}$, $X \cup \{y\}$ is not a circular code.

Remark 9. Any circular code with 20 trinucleotides is maximal as it cannot be contained in a circular code with more words.

Definition 8. Self-complementary code: A code X is self-complementary if, for each $y \in X$, $\mathcal{C}(y) \in X$.

Definition 9. C^3 self-complementary code: A code X is C^3 self-complementary if X, $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ are codes satisfying the following properties: $X = \mathcal{C}(X)$ (self-complementary), $\mathcal{C}(X_1) = X_2$ and $\mathcal{C}(X_2) = X_1$.

Property 3. The set X_0 is a maximal C^3 self-complementary circular code (Arquès and Michel, 1996).

A circular code allows the reading frames in genes to be retrieved. It is a set of words over an alphabet such that any word written on a circle (the next letter after the last letter of the word being the first letter) has a unique decomposition into words of the circular code. As an example, let the set Y be composed of the 6 following words: $Y = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ and the word w be a series of the 9 following letters: $w = ATGGCCCTA$. The word w written on a circle can be factorized into words of Y according to 2 different ways: ATG, GCC, CTA and AAT, GGC, CCT, the commas showing the way of decomposition (Fig. 3). Therefore, Y is not a circular code. In contrast, if the set Z obtained by replacing the word GGC of Y by GTC is considered, i.e., $Z = \{AAT, ATG, CCT, CTA, GCC, GTC\}$, then there never exists an ambiguous word with Z, particularly w is not ambiguous, and Z is a circular code. The construction frame of a word generated by any concatenation of words of a circular code can be retrieved after the reading, anywhere in the generated word, of a certain number of nucleotides depending on the code. This series of nucleotides is called the window of the circular code. Then the minimal window length is the size of the longest ambiguous word which can be read in at least 2 frames more one letter. The flower automaton is the classical method to determine if a set of words is a circular code or not (Lassez, 1976; Berstel and Perrin, 1985).

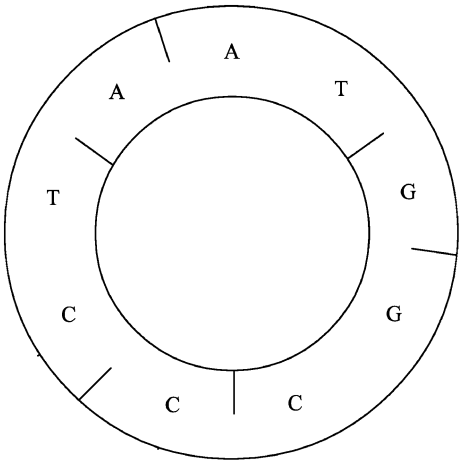

Fig. 3 The set $Y = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ is not a circular code as the word $w = ATGGCCCTA$ written on a circle can be factorized into words of Y according to 2 different ways: ATG, GCC, CTA and AAT, GGC, CCT.

3.1.2. Probabilities of the circular codes X_0, X_1, and X_2 in genes

In order to determine the probabilities of the 3 circular codes X_j, $j \in \{0, 1, 2\}$, in genes (reading frames 0), the 64 occurrence probabilities P_i of the trinucleotides i, $i \in \{1, \ldots, 64\}$ representing the 64 trinucleotides $\mathbb{T}$ in alphabetical order, are computed in genes (protein coding regions) of 175 complete genomes of prokaryotes (487, 758 genes, 454 megabases). This very large set of data allows stable and significant probabilities according to the law of large numbers. As the trinucleotides $\mathbb{T}_{id}$ are not considered in the definition of a circular code, the occurrence probability $P(X_j)$ of a code X_j is renormalizing

$$P(X_j) = \frac{\sum_{i \in X_j} P_i}{\sum_{i \in \mathbb{T} - \mathbb{T}_{id}} P_i}.$$

The frequency computation of the 3 codes X_j leads to the following values:

$$\begin{cases} P(X_0) = 48.8\%, \\ P(X_1) = 28.0\%, \\ P(X_2) = 23.2\%. \end{cases} \tag{9}$$

These values are retrieved with other gene populations (results not shown) and are very similar to those published in 1996 (Arquès and Michel, 1996) and also obviously to those in 2004 (Bahi and Michel, 2004).

As expected, the code X_0 occurs with the highest probability (48.8%) in genes as the codes X_1 and X_2 occur mainly in the frames 1 and 2, respectively. Also as expected, X_0 is not a "pure" code. Indeed, the fact that its probability is less than 1 means that it is mixed with the codes X_1 and X_2 in genes. Random mutations have introduced noise

during evolution, leading to a decreased probability of X_0. Furthermore, the probability inequality $P(X_1) > P(X_2)$, i.e., an asymmetry between the codes X_1 and X_2 in genes, is totally unexpected as the complementarity property between X_1 and X_2 (Property 2) would imply the same probabilities, even with the increase of noise during evolution. This code probability difference is equal to

$$\Delta P_{\text{Genes}} = P(X_1) - P(X_2) = 4.8\%. \tag{10}$$

The standard and pseudochaotic models will explain both the decreased probability of the code X_0 and the asymmetry between the codes X_1 and X_2 in genes.

3.2. Standard and pseudochaotic evolution models of circular codes

3.2.1. Principle

The observation in various genes (reading frames) from the 2 largest domains, the eukaryotes and the prokaryotes, of a preferential trinucleotide set X_0 which has strong biomathematical and evolutionary properties (a maximal C^3 self-complementary circular code; Property 3) is the basis of our evolution model. Indeed, if such a "universal" trinucleotide set occurs with a frequency higher than the random one in current genes after (mainly) random mutations, then a realistic hypothesis of an evolution model consists in asserting that this set had a higher frequency in past than now. In other words, the trinucleotides of X_0 are assumed to be the basic words of "primitive" genes (genes before random substitutions). As these primitive genes are constructed by trinucleotides of X_0, the mathematical model will be based on a 64×64 trinucleotide mutation matrix. The standard and pseudochaotic evolution models will be based on 2 processes.

A construction process ($t = 0$) generates primitive genes according to a random mixing of the 20 trinucleotides of the circular code X_0 with equiprobability (1/20). Therefore, the occurrence probability $P^{(t)}(X_j)$ of the 3 circular codes X_j, $j \in \{0, 1, 2\}$, at initial past time $t = 0$ is obvious

$$\begin{cases} P^{(0)}(X_0) = 1, \\ P^{(0)}(X_1) = P^{(0)}(X_2) = 0, \end{cases} \tag{11}$$

the codes X_1 and X_2 being absent according to the construction of the model. Thus, the initial vector in the standard and pseudochaotic evolution models is $P(0) =$ [0, 1/20, 0, 1/20, 0, 1/20, 0, 0, 0, 0, 0, 0, 1/20, 0, 1/20, 0, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/20, 1/20, 0, 1/20, 1/20, 1/20, 1/20, 0, 1/20, 0, 0, 0, 1/20, 0, 1/20, 1/20, 1/20, 0, 1/20, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/20, 0, 0].

These past code probabilities with the absence of X_1 and X_2 in primitive genes of these models are totally different from the code probabilities observed in genes with the decreased probability of X_0 and the asymmetry between X_1 and X_2. Therefore, an evolution process ($t > 0$) is added to the construction process to transform these primitive genes into current genes (simulated) in order to retrieve a correlation with the genes (real). Random time dependent substitutions with different rates $p^{(t)}$, $q^{(t)}$, and $r^{(t)}$ in the 3 sites, respectively, of the 20 trinucleotides of X_0 will generate the trinucleotides of X_1 and X_2 according to an unbalanced way in the hope of retrieving the statistical properties of the 3 codes in genes.

Remark 10. The circular codes X_0, X_1, and X_2 have the same frequencies of the nucleotides A, G, C, and T. So, the unequal frequencies of X_0, X_1, and X_2 cannot be explained by unequal nucleotide frequencies.

The standard and pseudochaotic evolution models will demonstrate here that the probabilities of the codes X_0, X_1, and X_2 in genes can be retrieved after a certain evolutionary time t of random substitutions in the circular code X_0 and with particular functions for the 3 time dependent substitution parameters.

As the trinucleotides $\mathbb{T}_{id}$ are not considered in the definition of a circular code, the probability $P^{(t)}(X_j)$ at time t of the 3 circular codes X_j, $j \in \{0, 1, 2\}$, in these evolution models is

$$P^{(t)}(X_j) = \frac{\sum_{i \in X_j} P_i^{(t)}}{\sum_{i \in \mathbb{T} - \mathbb{T}_{\mathrm{id}}} P_i^{(t)}}$$

with $P_i^{(t)}$ obtained by the formula (5) for the standard model and by the formula (6) for the pseudochaotic model.

3.2.2. Results with the time dependent standard evolution model (SEM)
Twelve classes of standard evolution models based on different time dependent substitution probabilities $p^{(t)}$, $q^{(t)}$, and $r^{(t)}$ have been studied in Bahi and Michel (2004). Only one standard evolution model SEM (model 7 in Bahi and Michel, 2004) based on the following substitution probabilities

$$\begin{aligned} p^{(t)} &= q^{(t)} = \tfrac{e^{-t}}{3}, \\ r^{(t)} &= 1 - p^{(t)} - q^{(t)} = 1 - \tfrac{2e^{-t}}{3} \end{aligned} \tag{12}$$

leads to a correlation with genes (Fig. 4). At initial past time $t = 0$, $p^{(0)} = q^{(0)} = r^{(0)} = 1/3$, leading to equiprobable substitution probabilities in the 3 trinucleotide sites. During evolution, the substitution probabilities $p^{(t)}$ and $q^{(t)}$ in the 2 first sites exponentially decrease from $1/3$ to 0 while the substitution probability $r^{(t)}$ in the 3rd site exponentially increases from $1/3$ to 1, in agreement with the actual degeneracy of the genetic code with the highest mutation rate in the 3rd site (see, e.g., Ermolaeva, 2001).

This model SEM simulates the decreased probability of the code X_0 and the asymmetry between the codes X_1 and X_2 observed in genes (Fig. 5). At the construction process ($t = 0$), SEM shows the past code probabilities (11). Totally unexpectedly, with these particular functions for $p^{(t)}$, $q^{(t)}$, and $r^{(t)}$ (12), the random substitution process ($t > 0$) of SEM generates the code probability inequality $P^{(t)}(X_1) > P^{(t)}(X_2)$ and increases its probability difference

$$\Delta P_{\mathrm{SEM}}(t) = P^{(t)}(X_1) - P^{(t)}(X_2) \tag{13}$$

during evolution which reaches a maximum value $\Delta P_{\mathrm{SEM}}(t) \approx 1.8\%$ at $t \geq 4.3$ (Fig. 5).

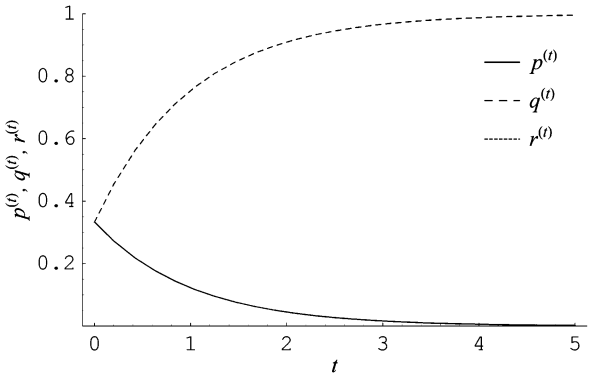

Fig. 4 Time dependent substitution probabilities in the standard and pseudochaotic evolution models SEM and CEM. The substitution probabilities $p^{(t)} = q^{(t)} = e^{-t}/3$ are associated with the 2 first trinucleotide sites (full line) and the substitution probability $r^{(t)} = 1 - 2e^{-t}/3$ is associated with the 3rd trinucleotide site (dash line) with $t \in [0, 5]$.

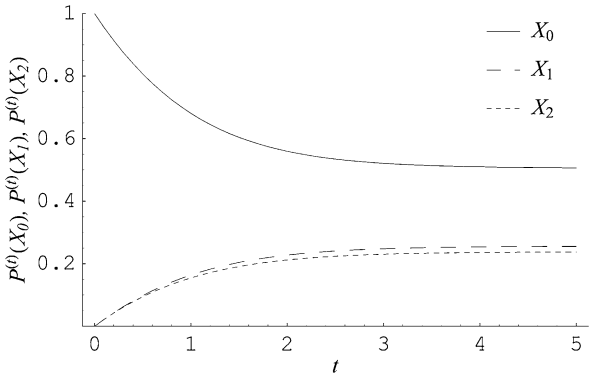

Fig. 5 Time dependent standard evolution model SEM of the circular codes X_0 (full line), X_1 (dash line), and X_2 (dot line) with $t \in [0, 5]$ as a function of the 3 substitution probabilities $p^{(t)} = q^{(t)} = e^{-t}/3$ associated with the 2 first trinucleotide sites and $r^{(t)} = 1 - 2e^{-t}/3$ associated with the 3rd trinucleotide site (Fig. 4). It simulates the asymmetry between the codes X_1 and X_2 observed in genes.

3.2.3. Results with the time dependent pseudochaotic evolution model (CEM)

In order to compare the pseudochaotic model CEM to the standard model SEM, the same substitution probabilities $p^{(t)}$, $q^{(t)}$, and $r^{(t)}$ are used. As in the model SEM, the random substitution process of the model CEM generates the decreased probability of the code X_0 and the asymmetry between the codes X_1 and X_2 observed in genes (Fig. 6). It increases the probability difference

$$\Delta P_{\mathrm{CEM}}(t) = P^{(t)}(X_1) - P^{(t)}(X_2) \tag{14}$$

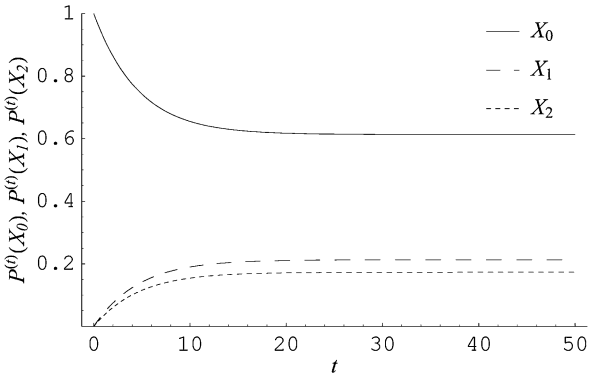

Fig. 6 Time dependent pseudochaotic evolution model CEM of the circular codes X_0 (full line), X_1 (dash line), and X_2 (dot line) with $t \in [0, 50]$ as a function of the 3 substitution probabilities $p^{(t)} = q^{(t)} = e^{-t}/3$ associated with the 2 first trinucleotide sites and $r^{(t)} = 1 - 2e^{-t}/3$ associated with the 3rd trinucleotide site (Fig. 4). It simulates significantly the asymmetry between the codes X_1 and X_2 observed in genes (Fig. 7).

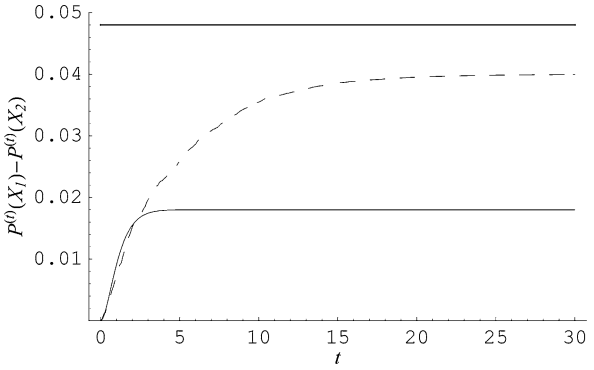

Fig. 7 Circular code probability differences $P^{(t)}(X_1) - P^{(t)}(X_2)$ between the 2 circular codes X_1 and X_2 with the standard evolution model SEM ($\Delta P_{\mathrm{SEM}}(t)$; thin full line) and the pseudochaotic model CEM ($\Delta P_{\mathrm{CEM}}(t)$; dash line). The pseudochaotic model CEM has a correlation with the circular code probability difference in genes (thick full line) better than the standard model SEM.

during evolution with a maximum value $\Delta P_{\mathrm{CEM}}(t) \approx 4.0\%$ at $t = 30$. The time interval in CEM which is shifted and larger than in SEM is related to some time steps with "mutation gaps".

Figure 7 summarizes the evolution of the circular code probability differences $P^{(t)}(X_1) - P^{(t)}(X_2)$ at time t with the standard model SEM ($\Delta P_{\mathrm{SEM}}(t)$ (13)) and the pseudochaotic model CEM ($\Delta P_{\mathrm{CEM}}(t)$ (14)). The maximum value $\Delta P_{\mathrm{CEM}}(t) \approx 4.0\%$ of CEM is close to the value 4.8% observed in genes (10) and significantly higher than the maximum value $\Delta P_{\mathrm{SEM}}(t) \approx 1.8\%$ of SEM. It could be related to some particular mutation properties of the circular code X_0. Thus, the pseudochaotic model CEM has the best correlation with the circular code probability difference in genes and so it is more realistic than the standard model SM.

4. Conclusion

We have developed a new class of stochastic models of gene evolution in which a random subset of 64 possible trinucleotides mutates at each evolutionary time t according to some time dependent substitution probabilities. Therefore, at each time t, the numbers and the types of mutable trinucleotides are unknown and the mutation matrix changes. Thus, the pseudochaotic evolution model generalizes the standard evolution model in which all the trinucleotides mutate at each time t according to constant or time dependent substitution parameters (Jukes and Cantor, 1969; Kimura, 1980, 1981; Takahata and Kimura, 1981; Arquès et al., 1998; Frey and Michel, 2006; Michel, 2007). It determines the occurrence probabilities at time t of trinucleotides which pseudochaotically mutate according to 3 time dependent substitution parameters associated with the 3 trinucleotide sites. A theorem proves that this pseudochaotic model converges to a uniform probability vector identical to that of the standard model. An application of this pseudochaotic model has allowed an evolutionary study of the 3 circular codes X_0, X_1, and X_2 identified in both eukaryotic and prokaryotic genes, particularly the decreased probability of X_0 and the asymmetry between X_1 and X_2. The pseudochaotic model has the best correlation with the circular code probability difference observed in genes, and thus it is more realistic than the standard model.

The biological meaning of these standard and pseudochaotic evolution models would suggest that the "primitive" genes, i.e., the genes before random substitutions ($t = 0$), are constructed by trinucleotides of the circular code X_0 according to an independent concatenation with equiprobability (1/20). The substitution process ($t > 0$) allows the generation of the circular code probability inequality $P^{(t)}(X_1) > P^{(t)}(X_2)$ and the increase of its probability difference during evolution. Furthermore, it retrieves the frequency orders of the 3 codes X_0, X_1, and X_2 in genes. The time dependent substitution probabilities with an exponential decrease in the 1st and 2nd trinucleotide sites and an exponential increase in the 3rd trinucleotide site agree with the actual degeneracy of the genetic code with a significantly highest mutation rate in the 3rd site (see, e.g., Ermolaeva, 2001). With random mutable and nonmutable trinucleotides during evolution, this pseudochaotic model has a very complex random behavior and its trinucleotide probability variations are totally unexpected. For example, the traces of the probability differences between primitive trinucleotides (initial probabilities) are conserved even after a great number of pseudochaotic substitutions, e.g., at $t = 50$ (Fig. 6).

As the pseudochaotic model allows some trinucleotides not to mutate for a certain period of time, it is a realistic evolutionary process from a biological point of view. The mathematical model formulated here allows the development of other biological applications (outside the circular code proposed here) in order to study evolution of genetic motifs with pseudochaotic mutations. The evolutionary behavior of mutable trinucleotides is modeled by the $J(k)$ parameter (see Section 2.2) which selects a subset of trinucleotides at a time step t_k according to any random distribution, e.g., a uniform one. However, as the mathematical model converges with any random distribution for the $J(k)$ parameter, the $J(k)$ parameter can have particular strategies. For example, one can imagine that during a certain evolutionary time interval, only the trinucleotides beginning with the nucleotide A are mutable, then during another time interval, the trinucleotides ending with the nucleotide C are mutable, etc. Evolution of trinucleotides with low mutability can also

be studied, e.g., the stop codons. Any deterministic strategy for the $J(k)$ parameter converges as long as Condition 1 of Theorem 1 is satisfied. Otherwise, different theoretical extensions of this pseudochaotic model can also be developed in the future, for example, trinucleotide mutation matrices using a greater number of substitution parameters, nucleotide and dinucleotide mutation matrices, etc.

References

Arndt, P.F., Burge, C.B., Hwa, T., 2002. DNA sequence evolution with neighbor-dependent mutation. In: RECOMB'02, Proceedings of the 6th Annual International Conference on Computational Biology, pp. 32–38.

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.

Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. J. Theor. Biol. 182, 45–58.

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

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

Berstel, J., Perrin, D., 1985. Theory of Codes, Academic, New York.

Chazan, D., Miranker, W., 1969. Chaotic relaxation. Linear Algebra Appl. 2, 199–222.

Ermolaeva, M.D., 2001. Synonymous codon usage in bacteria. Curr. Issues Mol. Biol. 3, 91–97.

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

Fryxell, K.J., Zuckerkandl, E., 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol. Biol. Evol. 17, 1371–1383.

Galtier, N., 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18, 866–873.

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11, 725–736.

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

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.

Lassez, J.-L., 1976. Circular codes and synchronization. Int. J. Comput. Syst. Sci. 5, 201–208.

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.

Michel, C.J., 2008. A 2006 review of circular codes in genes. Comput. Math. Appl. 55, 984–988.

Robert, F., 1986. Discrete Iterations: A Metric Study. Series in Computational Mathematics, vol. 6. Springer, Berlin.

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.

Tuffley, C., Steel, M., 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147, 63–91.

Wolfowitz, J., 1963. Products of indecomposable, aperiodic, stochastic matrices. Proc. Am. Math. Soc. 14, 733–737.

Yang, Z., 1994. Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314.

Yang, Z., Swanson, W.J., 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol. Biol. Evol. 19, 49–57.

Yang, Z., Nielsen, R., Goldman, N., Krabbe Pedersen, A.-M., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.