Journal of Theoretical Biology 255 (2008) 53–63

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

# A stochastic model of gene evolution with chaotic mutations

Jacques M. Bahi [a,1], 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

ARTICLE INFO

*Article history:*
Received 29 November 2007
Received in revised form
22 July 2008
Accepted 22 July 2008
Available online 25 July 2008

*Keywords:*
Stochastic model
Chaotic mutations
Gene evolution
Trinucleotides
Mutation matrix
Substitution parameters
Circular code

ABSTRACT

We develop here a new class of stochastic models of gene evolution in which the mutations are chaotic, i.e. a random subset of the 64 possible trinucleotides mutates at each evolutionary time $t$ according to some 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 $t$. The chaotic 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 chaotically mutate according to three substitution parameters associated with the three trinucleotide sites. Two theorems prove that this chaotic model has a probability vector at each time $t$ and that it converges to a uniform probability vector identical to that of the standard model. Furthermore, four applications of this chaotic model (with a uniform random strategy for the 64 trinucleotides and with a particular strategy for the three stop codons) allow an evolutionary study of the three 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 values for the three substitution parameters, the chaotic models retrieve the main statistical properties of the three circular codes observed in genes. These applications also allow an evolutionary comparison between the standard and chaotic models.



## 1. Introduction

Models of gene evolution were initially developed on the basis of nucleotide information. The first model with one substitution parameter (substitutions $\alpha$: all types) was proposed by Jukes and Cantor (1969) and generalized to two parameters (transitions $\alpha: A \longleftrightarrow G$ and $C \longleftrightarrow T$ and transversions $\beta: A \longleftrightarrow C$, $A \longleftrightarrow T$, $C \longleftrightarrow G$ and $G \longleftrightarrow T$) (Kimura, 1980), three parameters (transitions $\alpha: A \longleftrightarrow G$ and $C \longleftrightarrow T$, transversions $\beta: A \longleftrightarrow T$ and $C \longleftrightarrow G$ and transversions $\gamma: A \longleftrightarrow C$ and $G \longleftrightarrow T$) (Kimura, 1981), five parameters (transitions $\alpha: A \longrightarrow G$ and $T \longrightarrow C$, transitions $\beta: G \longrightarrow A$ and $C \longrightarrow T$, transversions $\gamma: A \longleftrightarrow T$ and $C \longleftrightarrow G$, transversions $\delta: A \longrightarrow C$ and $T \longrightarrow G$ and transversions $\varepsilon: C \longrightarrow A$ and $G \longrightarrow T$) (Takahata and Kimura, 1981) and six parameters ($\alpha: A \longrightarrow C$, $A \longrightarrow G$, $T \longrightarrow C$ and $T \longrightarrow G$, $\alpha_1: A \longrightarrow T$, $\alpha_2: C \longrightarrow G$, $\beta: C \longrightarrow A$, $C \longrightarrow T$, $G \longrightarrow A$ and $G \longrightarrow T$, $\beta_1: T \longrightarrow A$ and $\beta_2: G \longrightarrow C$) (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 application of random mutations) has been proposed by Fryxell and Zuckerkandl (2000) while a discrete version with time steps $\Delta t/L$ where $\Delta t$ 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 $61 \times 61$ mutation matrix based on numerical solutions (Goldman and Yang, 1994) and its extensions to the non-synonymous/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

* Corresponding author. Tel.: +33 3 90 24 44 62.
*E-mail addresses:* jacques.bahi@univ-fcomte.fr (J.M. Bahi), michel@dpt-info.u-strasbg.fr (C.J. Michel).
[1] Tel.: +33 3 84 58 77 94.

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.07.028

"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 different types of substitutions in the trinucleotide sites. This trinucleotide set can obviously be reduced to only one trinucleotide.

In the standard stochastic evolution model, 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 non-zero elements in this matrix are always in the same positions. Furthermore, non-zero 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 $64 \times 64$ trinucleotide mutation matrix with nine substitution parameters associated with the three types of substitutions in the three trinucleotide sites (Michel, 2007). It generalizes the models based on the $4 \times 4$ nucleotide mutation matrices, particularly (Jukes and Cantor, 1969; Kimura, 1980, 1981; Takahata and Kimura, 1981), and the $64 \times 64$ trinucleotide mutation matrices with three and six 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 chaotic mutations. In this chaotic model, a random subset of the 64 possible trinucleotides mutates at each evolutionary time $t$ according to some substitution probabilities while the other trinucleotides do not mutate. Thus, the mutation matrix changes in time. This chaotic model generalizes the standard one. It differs from the previous ones such as the covarion-style model (Tuffley and Steel, 1998) in which the sites have a chaotic evolution modelled by two Markov processes, by the fact that in our approach the motifs have a chaotic evolution modelled by one Markov process. As the chaotic model is 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. For example, the degeneracy of the genetic code allows several codons to code the same amino acid. This codon usage bias (Grantham et al., 1980) is generally correlated with gene expressivity (Grantham et al., 1981; Ikemura, 1985; Sharp and Matassi, 1994), even if its strength varies among bacterial species (Sharp et al., 2005). A proposed explanation is that codon usage reflects the variations in the concentrations of tRNAs. Major codons encoded by more abundant tRNAs should increase translational efficacy (Bulmer, 1991; Akashi and Eyre-Walker, 1998). Nevertheless, tRNA abundance could also have evolved for matching codon pattern in a genome (Fedorov et al., 2002) and then would rather be a consequence of the synonymous codon bias. Several other processes may influence codon usage (see the review in Ermolaeva, 2001). The extreme case could be the stop codons which mutate rarely, perhaps even never. Therefore, codon usage bias could also be studied by evolution models with codon substitution probabilities which remain constant in time (classical approach) or which are chaotic, i.e. for a certain period of time, a set of codons mutates, then in another period of time, another set of codons mutates. This chaotic process can be based on random or particular strategies.

Two types of results are presented in this paper:

(i) A mathematical model of gene evolution with chaotic mutations is developed and two theorems prove that it leads to a probability vector at each time $t$ and that it converges to a uniform probability vector identical to that of the standard model (Section 2).

(ii) Four applications of this chaotic model (with a uniform random strategy for the 64 trinucleotides and with a particular strategy for the three stop codons) to the evolution of the three circular codes identified in both eukaryotic and prokaryotic genes allow the retrieval of their main statistical properties. Furthermore, they also allow an evolutionary comparison between the standard and chaotic models (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 chaotically mutate according to three substitution probabilities $p$, $q$ and $r$ associated with the three 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 $t + \Delta t$, the occurrence probability of the trinucleotide $i$ is $P_i(t + \Delta t)$ so that $P_i(t + \Delta t) - P_i(t)$ represents the probabilities of trinucleotides $i$ which appear and disappear during the time interval $\Delta t$

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

where $\alpha$ is the probability that a trinucleotide is subjected to one substitution during $\Delta t$ 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 $\Delta t$ is assumed to be small enough that a trinucleotide cannot mutate twice in a row during $\Delta t$) otherwise it is given as a function of the three substitution rates $p$, $q$ and $r$. For example, with the trinucleotide $AAA$ associated with $i = 1$, $P(CAA \to AAA) = P(GAA \to AAA) = P(TAA \to AAA) = p/3$, $P(ACA \to AAA) = P(AGA \to AAA) = P(ATA \to AAA) = q/3$, $P(AAC \to AAA) = P(AAG \to AAA) = P(AAT \to AAA) = r/3$ 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 an average one substitution per trinucleotide per time interval. Then,

$$P_i(t + \Delta t) - P_i(t) = \Delta t \sum_{j=1}^{64} P(j \to i) P_j(t) - \Delta t P_i(t). \tag{2.1}$$

Formula (2.1) leads to

$$\lim_{\Delta t \to 0} \frac{P_i(t + \Delta t) - P_i(t)}{\Delta t} = P_i'(t) = \sum_{j=1}^{64} P(j \to i) P_j(t) - P_i(t). \tag{2.2}$$

By considering the column vector $P(t) = [P_i(t)]_{1 \leqslant i \leqslant 64}$ made of the 64 $P_i(t)$ and the mutation matrix $A$ $(64, 64)$ of the 4096 trinucleotide substitution probabilities, i.e. $A_{ij} = P(i \to j)$, differential equation (2.2) can be represented by the following matrix equation:

$$P'(t) = AP(t) - P(t) = (A - I)P(t), \tag{2.3}$$

where $I$ represents the identity matrix, or similarly by

$$\forall i \in \{1, \ldots, 64\}, \quad \forall t \geqslant 0, \quad P_i'(t) = \sum_{j=1}^{64} (A - I)_{ij} P_j(t). \tag{2.4}$$

The mutation matrix $A$ $(64, 64)$ can be defined by a square block matrix $(4, 4)$ whose four diagonal elements are formed by four identical square submatrices $B$ $(16, 16)$ and whose

12 non-diagonal elements are formed by 12 identical square submatrices $(p/3)I$ (16, 16)

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

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 square submatrix $B(16,16)$ can again be defined by a square block matrix (4, 4) whose four diagonal elements are formed by four identical square submatrices $C(4,4)$ and whose 12 non-diagonal elements are formed by 12 identical square submatrices $(q/3)I$ (4, 4)

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

Finally, the square submatrix $C(4,4)$ is equal to

$$C = \begin{pmatrix} 0 & r/3 & r/3 & r/3 \\ r/3 & 0 & r/3 & r/3 \\ r/3 & r/3 & 0 & r/3 \\ r/3 & r/3 & r/3 & 0 \end{pmatrix}.$$

Matrix $A$ is stochastic when $p+q+r=1$. This block matrix property of the mutation matrix $A$ is important, particularly to determine the eigenvalues in the standard model or for a parallel computing of eigenelements.

### 2.1. Standard stochastic evolution model (SM)

In the standard stochastic evolution model *SM*, analytical solutions can be derived.

The differential equation (2.3) can then be written in the following form:

$$P'(t) = MP(t)$$

with

$$M = A - I.$$

As the three substitution parameters are real, matrix $A$ is real. It is also symmetric by construction. Therefore, matrix $M$ is also real and symmetric. There exist an eigenvector matrix $Q$ and a diagonal matrix $D$ of eigenvalues $\lambda_k$ of $M$ ordered in the same way as the eigenvector columns in $Q$ so that $M = QDQ^{-1}$. Then

$$P'(t) = QDQ^{-1}P(t).$$

This equation has the classical solution (see e.g. Lange, 2005)

$$P(t) = Qe^{Dt}Q^{-1}P(0), \tag{2.5}$$

where $e^{Dt}$ is the diagonal matrix of exponential eigenvalues $e^{\lambda_k t}$ and $P(0)$, the given initial probability vector.

The evolutionary analytical formulas $P(t)$ of the 64 trinucleotides $\mathbb{T}$ as a function of the three substitution parameters $p$, $q$ and $r$ associated with the three trinucleotide sites, respectively, can be deduced from the evolutionary analytical formulas $P(t)$ of the 64 trinucleotides $\mathbb{T}$ as a function of the nine parameters $a$, $b$, $c$, $d$, $e$, $f$, $g$, $h$ and $k$ associated with the three types of substitutions in the three trinucleotide sites (Michel, 2007). These nine parameters are $a$, $d$ and $g$ associated with the probabilities of transitions $A\longleftrightarrow G$ (a substitution from one purine $\{A,G\}$ to the other) and $C\longleftrightarrow T$ (a substitution from one pyrimidine $\{C,T\}$ to the other) in the three sites, respectively; $b$, $e$ and $h$ associated with the probabilities of transversions (a substitution from a purine to a pyrimidine, or reciprocally) $A\longleftrightarrow T$ and $C\longleftrightarrow G$ in the three sites, respectively; $c$, $f$ and $k$ associated with the probabilities of transversions $A\longleftrightarrow C$ and $G\longleftrightarrow T$ in the three sites, respectively. Indeed, the three parameter model is a particular case of the nine parameter model with $a=b=c=p/3$, $d=e=f=q/3$ and $g=h=k=r/3$. We refer to Michel (2007), particularly for the determination of the eigenvalues of matrices $A$ and $M$ as a function of the nine parameters and their eigenvectors which can be expressed in a form which does not dependent on these nine parameters.

Formula (2.5) with the 64 initial trinucleotide probabilities $P_j(0)$ before the substitution process ($t=0$), the diagonal matrix $e^{Dt}$ of exponential eigenvalues $e^{\lambda_k t}$ of $M$, its eigenvector matrix $Q$ and its inverse $Q^{-1}$ will determine the 64 trinucleotide probabilities $P_i(t)$ after $t$ substitutions as a function of the three substitution parameters $p$, $q$ and $r$ (Section 3.2.2).

### 2.2. Continuous time chaotic evolution model

The new chaotic stochastic evolution model will generalize the standard model to a random subset of the 64 possible trinucleotides which chaotically mutate at each evolutionary time $t$ according to some probabilities. Therefore, the numbers and the types of mutable trinucleotides at each time $t$ are unknown. In addition, the randomly conserved subset of trinucleotides is conserved not only in the sense that they do not mutate, but also in the sense that they cannot be generated through mutations. We assume that in our standard continuous model some trinucleotides which do not belong to a random subset $J(t)$, do not mutate. So their probabilities $P_i(t)$ are constant, thus leading to $P_i'(t)=0$. Therefore, the continuous time chaotic evolution model is described by the following system: for any arbitrary time $T$ and for all $t\in[0,T]$,

$$\begin{cases} P_i'(t) = 0 & \text{if } i\notin J(t), \\ P_i'(t) = \sum_{j=1}^{64}(A-I)_{ij}P_j(t) & \text{if } i\in J(t). \end{cases} \tag{2.6}$$

### 2.3. Discrete time chaotic evolution model (CM)

Let us discretize the time interval $[0,T]$ as follows: $0=t_0<t_1<\cdots<t_k=T$. For the sake of simplicity, we will suppose that $t_{i+1}-t_i=h$, 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 "non-mutable 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 order to obtain the discrete analogue of the chaotic continuous model (2.6), 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:

$$\begin{cases} P_i(t_k) = P_i(t_{k-1}) & \text{if } i\notin J(k), \\ P_i(t_k) = h\sum_{j=1}^{64}(A^{(k)}-I)_{ij}P_j(t_{k-1}) + P_i(t_{k-1}) & \text{if } i\in J(k). \end{cases} \tag{2.7}$$

This model, which we call the discrete time chaotic evolution model *CM*, is a discrete version of the continuous chaotic one (2.6).

**Remark 1.** It should be noticed that the continuous chaotic model is a generalization of the standard model. Indeed, we only have to suppose that at each time $t$ all the trinucleotides mutate. This is achieved in the discrete chaotic model by taking $J(k)=\{1,\ldots,64\}$ for all $k$.

**Remark 2.** In contrast to the standard model, the continuous chaotic model has no analytical solution. The discrete chaotic model gives a discrete solution to the continuous chaotic model by taking a sufficiently small $h$.

*2.3.1. 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 $A_{ii}\neq 0$ (non-zero 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 chaotic model *CM* leads to some new properties with the mutation matrix $A$ compared to the classical standard models. For any time step $t_k$ in the model *CM*, 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, matrix $A^{(k)}$ changes at each step $t_k$. The two laws below explain why the stochasticity and the symmetry of matrix $A^{(k)}$ are preserved in time with both mutable and non-mutable trinucleotides.

(i) Matrix conservation law for non-mutable trinucleotides. The probability of a non-mutable 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 non-mutable trinucleotide $i$ are equal to 0 except their $i$th diagonal element which is equal to 1. Furthermore, this matrix conservation law keeps the symmetry of matrix $A^{(k)}$ for a non-mutable trinucleotide. Finally, this property can obviously be extended to a set $\overline{J(k)}$ of non-mutable 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. 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 non-mutable elements are set to a probability equal to 0, matrix $A^{(k)}$ is also symmetric.

The symmetry of the mutation matrix $A^{(k)}$ with the mutable trinucleotides implies that the trinucleotide probabilities are all equal in the limiting distribution. This assumption is found in the standard models of Jukes and Cantor (1969), Kimura (1980, 1981), Takahata and Kimura (1981), Arquès et al. (1998), Frey and Michel (2006), Michel (2007), etc. Standard models analyzing non-uniform trinucleotide distributions do not have this assumption, e.g. Dayhoff et al. (1978), Henikoff and Henikoff (1992), Kelly and Churchill (1996), Thorne and Goldman (2003).

*2.3.2. Probability vector at each time step*

**Theorem 1.** *In the chaotic stochastic evolution model CM, $P(t_k)$ is a probability vector at each time step $t_k$.*

**Proof.** See Appendix A.

*2.3.3. Convergence analysis*

Let us define 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 chaotic model *CM* can be written as follows: at a time step $t_k\in\mathbb{R}$

$$P(t_k)=B^{(k)}P(t_{k-1}). \tag{2.8}$$

This model *CM* can be associated with a sequence of undirected connected graphs $G^{(k)}=(V^{(k)},E^{(k)})$ where $V^{(k)}$ is the set of 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 a time step $k$. By definition, each vertex is labelled from 1 to $|V|=64$. Let $|E|$ be the number of mutation edges. In the sequel, we will call $G^{(k)}$, the mutation probability graph at the time step $k$.

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 model *CM*, these elements change in time.

**Condition 1.** *Infinitely often, a trinucleotide mutates and the mutation graphs are connected (i.e. the corresponding incidence matrix is irreducible) and non-bipartite.*

**Remark 3.** The condition above is equivalent to the following one: there exists a subsequence $\{k_p\}_{k\in\mathbb{N}}$ such that, at the time steps $k_p$, the connection graphs are connected and non-bipartite.

**Theorem 2.** *Under Condition 1 and for all $p,q,r\in]0,1[$, the chaotic stochastic evolution model CM converges to the uniform probability vector $(1/64,\ldots,1/64)^{\mathrm{T}}$.*

**Proof.** see Appendix B.

**Remark 4.** Theorem 1 cannot be derived from the known result of Wolfowitz (1963) on stochastic matrices. Indeed, in our model, words (products of $B^{(k)}$) do not have to be indecomposable and aperiodic (SIA), i.e. in our framework, it is possible that $\lim_{n\to\infty}W^n$ does not exist for some words $W$.

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

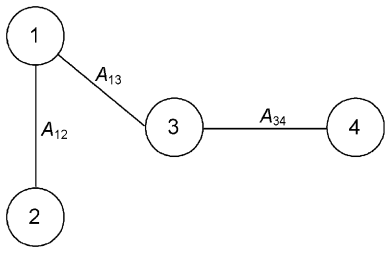


**Fig. 1.** A fixed mutation probability graph.

## 3. Applications of chaotic stochastic evolution models 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 1.** By convention, the reading frame established by a start codon belonging to $\mathbb{T}_{start} = \{ATG, GTG, TTG\}$ 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 three frames of genes showed that the trinucleotides are not uniformly distributed in these three frames. By excluding the four trinucleotides with identical nucleotides $\mathbb{T}_{id} = \{AAA, CCC, GGG, TTT\}$, the same three subsets $X_0$, $X_1$ and $X_2$ of 20 trinucleotides are found in the frames 0, 1 and 2, respectively, of two 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 three 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 $X_0 = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$.

The subsets $X_1$ and $X_2$ of 20 trinucleotides are deduced from $X_0$ thanks to the permutation property.

**Definition 2.** The (left circular) permutation $\mathscr{P}$ of a trinucleotide $w_0 = l_0 l_1 l_2 \in \mathbb{T}$ is the permuted trinucleotide $\mathscr{P}(w_0) = w_1 = l_1 l_2 l_0$, e.g. $\mathscr{P}(AAC) = ACA$, and $\mathscr{P}(\mathscr{P}(w_0)) = \mathscr{P}(w_1) = w_2 = l_2 l_0 l_1$, e.g. $\mathscr{P}(\mathscr{P}(AAC)) = CAA$. This definition is naturally extended to the trinucleotide set permutation: the permutation $\mathscr{P}$ of a set of trinucleotides is the permuted trinucleotide set obtained by the permutation $\mathscr{P}$ of all its trinucleotides.

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

These three trinucleotide subsets present several strong biomathematical properties, particularly the fact that they are circular codes.

**Notation 1.** $\mathbb{A}$ *being a finite alphabet*, $\mathbb{A}^*$ *denotes the words over* $\mathbb{A}$ *of finite length including the empty word* $\varepsilon$ *of length* 0 *and* $\mathbb{A}^+$, *the words over* $\mathbb{A}$ *of finite length greater or equal to* 1. *Let* $w_1 w_2$ *be the concatenation of the two words* $w_1$ *and* $w_2$.

**Definition 3.** A subset $X$ of $\mathbb{A}^+$ is a circular code if $\forall n, m \geqslant 1$ and $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m \in X$, and $r \in \mathbb{A}^*$, $s \in \mathbb{A}^+$, the equalities $sx_2 \ldots x_n r = y_1 y_2 \ldots y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ and $x_i = y_i$, $1 \leqslant i \leqslant n$.

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 six following words: $Y = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ and the word $w$ be a series of the nine following letters: $w = ATGGCCCTA$. The word $w$, written on a circle, can be factorized into words of $Y$ according to two different ways: $ATG, GCC, CTA$ and $AAT, GGC, CCT$, the commas showing the way of decomposition. Therefore, $Y$ is not a circular code (Fig. 2). 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 (called the window of the circular code).

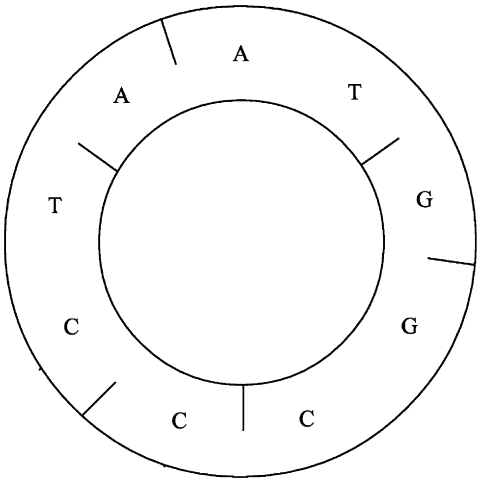


**Fig. 2.** The set $Y = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ is not a circular code as the word $w = ATGGCCCTA$, written on a circle, can be factorized into words of $Y$ according to two different ways: $ATG, GCC, CTA$ and $AAT, GGC, CCT$.

**Property 2.** *Maximal circular code*: $X_0$, $X_1$ *and* $X_2$ *are maximal circular codes* (20 *trinucleotides*) *as they are not contained in larger circular codes, i.e. in circular codes with more words* (*proof in* Michel, 2008).

These three trinucleotide subsets are also related by the complementary property.

**Definition 4.** The complementarity $\mathscr{C}$ of a trinucleotide $w_0 = l_0 l_1 l_2 \in \mathbb{T}$ is the complementary trinucleotide $\mathscr{C}(w_0) = \mathscr{C}(l_2)\mathscr{C}(l_1)\mathscr{C}(l_0)$ with $\mathscr{C}(A) = T$, $\mathscr{C}(C) = G$, $\mathscr{C}(G) = C$, $\mathscr{C}(T) = A$, e.g. $\mathscr{C}(AAC) = GTT$. This definition is naturally extended to the trinucleotide set complementarity.

**Property 3.** *Complementarity*: $\mathscr{C}(X_0) = X_0$ ($X_0$ *is self-complementary*) *and*, $\mathscr{C}(X_1) = X_2$ *and* $\mathscr{C}(X_2) = X_1$ ($X_1$ *and* $X_2$ *are complementary to each other*).

Thanks to Property 3, the circular code $X_0$ and its two permuted circular codes $X_1$ and $X_2$ can exist 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 one nucleotide in the $5'-3'$ direction) in a given DNA strand can be paired with $X_2$ ($X_0$ shifted by two 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).

The research context, the history and the other properties ($C^3$ code, rarity, largest window length, higher frequency of "misplaced" trinucleotides, flexibility) of these circular codes $X_0$, $X_1$ and $X_2$ are detailed in Michel (2008).

#### 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 three 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 three codes $X_j$ leads to the following values:

$$\begin{cases} P(X_0) = 48.8\%, \\ P(X_1) = 28.0\%, \\ P(X_2) = 23.2\%. \end{cases} \tag{3.1}$$

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

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 3) would imply the same probabilities, even with the increase in noise during evolution. This code probability difference is equal to

$$\Delta P_{Genes} = P(X_1) - P(X_2) = 4.8\%. \tag{3.2}$$

The standard and chaotic 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 chaotic stochastic evolution models of circular codes

#### 3.2.1. Principle

The observation of a preferential trinucleotide set $X_0$ in various genes (reading frames) from the two largest domains, the eukaryotes and the prokaryotes, 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 the 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 $64 \times 64$ trinucleotide mutation matrix. The standard and chaotic models *SM* and *CM* will be based on two 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 ($\frac{1}{20}$). Therefore, the occurrence probability $P(X_j, t)$ of the three circular codes $X_j$, $j \in \{0, 1, 2\}$, at initial past time $t = 0$ is obvious: $P(X_0, 0) = 1$ and $P(X_1, 0) = P(X_2, 0) = 0$ (absence of the codes $X_1$ and $X_2$ according to the construction of the model). Thus, the initial vector in *SM* and *CM* is $P(0) = [0, 1/20, 0, 1/20, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 1/20, 0, 1/20, 0, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/20, 1/20, 0, 1/20, 1/20, 1/20, 1/20, 0, 1/20, 0, 0, 0, 0, 1/20, 0, 1/20, 1/20, 1/20, 0, 1/20, 0, 1/20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/20, 0, 0]$.

These past code probabilities with the absence of $X_1$ and $X_2$ in primitive genes of this model 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 actual ones (simulated) in order to retrieve a correlation with the genes (real). Random substitutions with different rates $p$, $q$ and $r$ in the three 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 three codes in genes.

**Remark 6.** 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 chaotic 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 values for the three substitution parameters.

#### 3.2.2. Solution to the standard stochastic evolution model (SM)

As the trinucleotides $\mathbb{T}_{id}$ are not considered in the definition of a circular code, the probability $P(X_j, t)$ at time $t$ of the three circular codes $X_j$, $j \in \{0, 1, 2\}$, in the standard stochastic evolution model *SM* is

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

with $P_i(t)$ obtained from formula (2.5) according to Property 4 of the nine parameter model (Michel, 2007).

Therefore, the analytical formulas $P(X_0, t)$ (equal to $P_1(C, t)$ of Property 4 in Michel, 2007), $P(X_1, t)$ and $P(X_2, t)$ at time $t$ of the three circular codes $X_0$, $X_1$ and $X_2$ and as a function of the three substitution parameters $p$, $q$ and $r$ (associated with the three trinucleotide sites) are

$$\begin{aligned} P(X_0, t) = \frac{1}{2D}(50 &+ 28e^{-(4/3)t} + 19e^{-(4/3)pt} + 18e^{-(4/3)qt} \\ &+ 19e^{-(4/3)rt} + 5e^{-(4/3)(1-p)t} + 16e^{-(4/3)(1-q)t} \\ &+ 5e^{-(4/3)(1-r)t}), \end{aligned} \tag{3.3}$$

$$\begin{aligned} P(X_1, t) = \frac{1}{2D}(50 &- 12e^{-(4/3)t} - 9e^{-(4/3)pt} - 9e^{-(4/3)qt} \\ &- 10e^{-(4/3)rt} - 5e^{-(4/3)(1-p)t} - 5e^{-(4/3)(1-q)t}), \end{aligned} \tag{3.4}$$

$$\begin{aligned} P(X_2, t) = \frac{1}{2D}(50 &- 12e^{-(4/3)t} - 10e^{-(4/3)pt} - 9e^{-(4/3)qt} \\ &- 9e^{-(4/3)rt} - 5e^{-(4/3)(1-q)t} - 5e^{-(4/3)(1-r)t}) \end{aligned} \tag{3.5}$$

with the denominator $D$

$$D = 75 + 2e^{-(4/3)t} + 3e^{-(4/3)(1-q)t}.$$

**Remark 7.** Formulas $P(X_0, t)$, $P(X_1, t)$ and $P(X_2, t)$ are the solutions to $P(t) = Qe^{Dt}Q^{-1}P(0)$ (2.5) where $e^{Dt}$ is the diagonal matrix of exponential eigenvalues $e^{\lambda_k t}$ and $P(0)$, the given initial probability vector. The eigenvalues $\lambda_k$ of $M = A - I$ are deduced from the eigenvalues $\mu_k$ of $A$ such that $\lambda_k = \mu_k - 1$. The eigenvalues $\mu_k$ of $A$ can be obtained by determining the roots of the characteristic equation $\det(A - \mu I) = 0$ of $A$ using its block matrix properties. Therefore, after linear combinations, 64 eigenvalues $\lambda_k$ of $M$ of algebraic multiplicity 1 are obtained. The 64 eigenvectors of $M$ associated with these 64 eigenvalues $\lambda_k$ computed by formal calculus can be expressed in a form which does not dependent on the substitution parameters.

**Proposition 1.** $\forall p, q, r, t, \sum_{j=1}^{3} P(X_j, t) = 1$.

**Proposition 2.** $P(X_0, 0) = 1, P(X_1, 0) = P(X_2, 0) = 0$.

**Proof.** The initial probability $P(X_j,0)$ at time $t=0$ of a code $X_j$ can obviously be obtained from the analytical solution $P(X_j,t)$ with $t=0$ (3.3) (3.4) (3.5) and also by a simple probability calculus. Indeed, the probability $P(X_0,0)$ is equal to 1 as the primitive genes in this evolution model are generated by the code $X_0$ (20 among 20 trinucleotides) and then, $P(X_1,0)=P(X_2,0)=0$. □

**Proposition 3.** $\forall p,q,r \in ]0,1[$, $\lim_{t\to+\infty} P(X_j,t)=\frac{1}{3}$.

**Proof.** The probability $P(X_j,t)$ at limit time $t\to\infty$ of a code $X_j$ can obviously be obtained from the limit of the analytical solution $P(X_j,t)$ (3.3) (3.4) (3.5) and also by a simple probability calculus. Indeed, the three substitutions in the 20 trinucleotides of $X_0$ generate the 44 other trinucleotides. When $t\to\infty$, the 64 trinucleotides $\mathbb{T}$ occur with the same probabilities and therefore, the probabilities of $X_0$, $X_1$ and $X_2$ are equal to $20/60=\frac{1}{3}$ (the four trinucleotides $\mathbb{T}_{id}$ being ignored). □

**Proposition 4.** *For some values* $p,q,r \in \{0,1\}$, $\lim_{t\to+\infty} P(X_j,t)\neq\frac{1}{3}$ *as some trinucleotides may be either not generated or generated without equiprobability.*

**Example 1.** With $p=0$, $\lim_{t\to+\infty}P(X_0,t)=\frac{23}{50}$ (from the analytical solution (3.3) or a probability calculus not given here).

We can attempt to explain the asymmetry between the probabilities of codes $X_1$ and $X_2$ and the decreased probability of code $X_0$ with the standard model *SM* by investigating the situation in which the probability difference

$$P(X_1,t)-P(X_2,t)>3\%, \tag{3.6}$$

3% being chosen close to the real value $\Delta P_{Genes}=4.8\%$ (3.2), and the probability inequality

$$P(X_1,t)>P(X_2)=23.2\% \tag{3.7}$$

to relate the values of the evolution model (probability $P(X_1,t)$) to the gene reality (probability $P(X_2)$ equal to 23.2% (3.1)). The occurrence probabilities $P(X_j,t)$ of the three circular codes $X_0$ (3.3), $X_1$ (3.4) and $X_2$ (3.5) are computed by varying each substitution rate $p$, $q$ and $r$ in the range $[0,1]$ with a step of 1%, such that their sum is equal to 1, and $t$ in the range $[0,30]$.

The solution space (in %) of the standard model *SM* verifying the two previous probability inequalities (3.6) and (3.7) is very restricted: $p\in\{0,\ldots,2\}$, $q\in\{1,\ldots,5\}$ and $r\in\{95,\ldots,99\}$. Its barycenter values are $\overline{p}=0.8\%$, $\overline{q}=2.6\%$ and $\overline{r}=96.6\%$ with $t\in[6.8,7.3]$. Fig. 3 gives a graphical representation of the analytical solutions $P(X_0,t)$, $P(X_1,t)$ and $P(X_2,t)$ in this substitution rate barycenter. At the construction process ($t=0$), *SM* shows the past code probabilities (Proposition 2). Totally unexpectedly, with these barycenter values $\overline{p}$, $\overline{q}$ and $\overline{r}$, the random substitution process ($t>0$) of *SM* generates the code probability inequality $P(X_1,t)>P(X_2,t)$ and increases its probability difference

$$\Delta P_{SM}(t)=P(X_1,t)-P(X_2,t) \tag{3.8}$$

during evolution which reaches a maximum value $\Delta P_{SM}(t)\approx 3.47\%$ at $t=2.8$ (Fig. 3).

### 3.2.3. Solution to the chaotic stochastic evolution model (CM)

As the trinucleotides $\mathbb{T}_{id}$ are not considered in the definition of a circular code, the probability $P(X_j,t_k)$ of the three circular codes $X_j$, $j\in\{0,1,2\}$, at the time step $t_k$ in the chaotic stochastic evolution model *CM* (with a uniform random strategy for the 64 trinucleotides) is

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

with $P_i(t_k)$ obtained from formula (2.8).

In order to compare the chaotic model *CM* with the standard model *SM*, the same barycenter values $\overline{p}$, $\overline{q}$ and $\overline{r}$ are used. As in the model *SM*, the random substitution process of the model *CM* generates the code probability inequality $P(X_1,t)>P(X_2,t)$ and increases its probability difference

$$\Delta P_{CM}(t)=P(X_1,t)-P(X_2,t) \tag{3.9}$$

during evolution, albeit more slowly, with a maximum value $\Delta P_{CM}(t)\approx 3.52\%$ at $t=16.2$. The probability inequalities (3.6) and (3.7) observed in genes are verified in the model *CM* with a time interval $t\in[39.4,43.0]$ (Fig. 4). This time interval in *CM* which is shifted and larger than in *SM* is related to some time steps with "mutation gaps".

The mathematical model of chaotic mutations can have any random strategy $J(k)$ of mutations, i.e. the numbers and the types of mutable trinucleotides are unknown at each evolutionary time. Obviously, it allows particular strategies $J(k)$ of chaotic mutations. We propose here three chaotic models with particular strategies for the stop codons.

The chaotic model $CM_{TAA}$ is based on the low mutability of the stop codon *TAA*. Precisely, the strategy $J(k)$ randomly chooses the

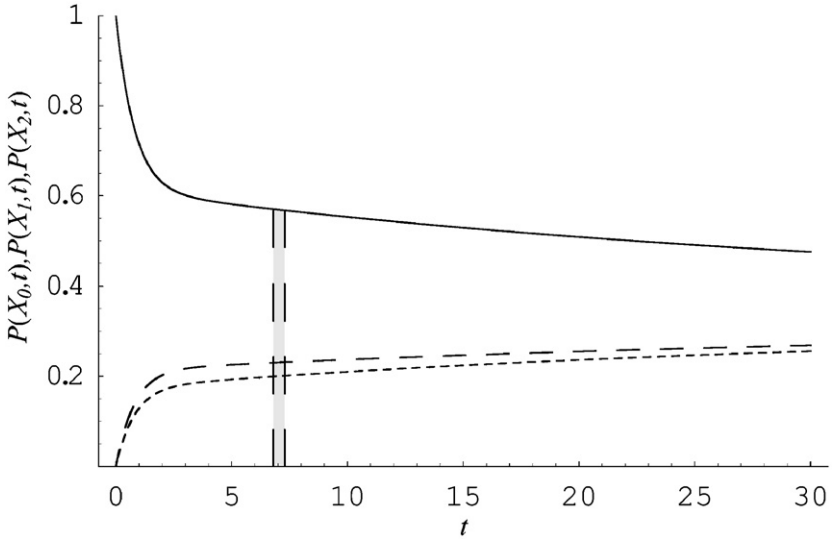


**Fig. 3.** Standard stochastic evolution model *SM* of the circular codes $X_0$, $X_1$ and $X_2$ in the substitution rate barycenter: $\overline{p}=0.8\%$, $\overline{q}=2.6\%$ and $\overline{r}=96.6\%$. The model *SM* verifies the circular code probability inequalities (3.6) and (3.7) observed in genes with $t\in[6.8,7.3]$ (in grey level).

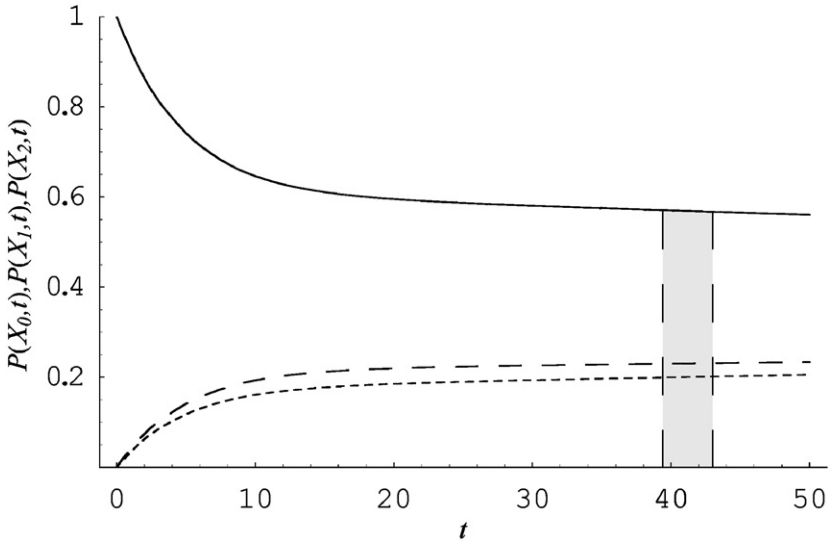


**Fig. 4.** Chaotic stochastic evolution model *CM* (with a uniform random strategy for the 64 trinucleotides) of the circular codes $X_0$, $X_1$ and $X_2$ in the substitution rate barycenter: $\overline{p} = 0.8\%$, $\overline{q} = 2.6\%$ and $\overline{r} = 96.6\%$. The model *CM* verifies the circular code probability inequalities (3.6) and (3.7) observed in genes with $t \in [39.4, 43.0]$ (in grey level).

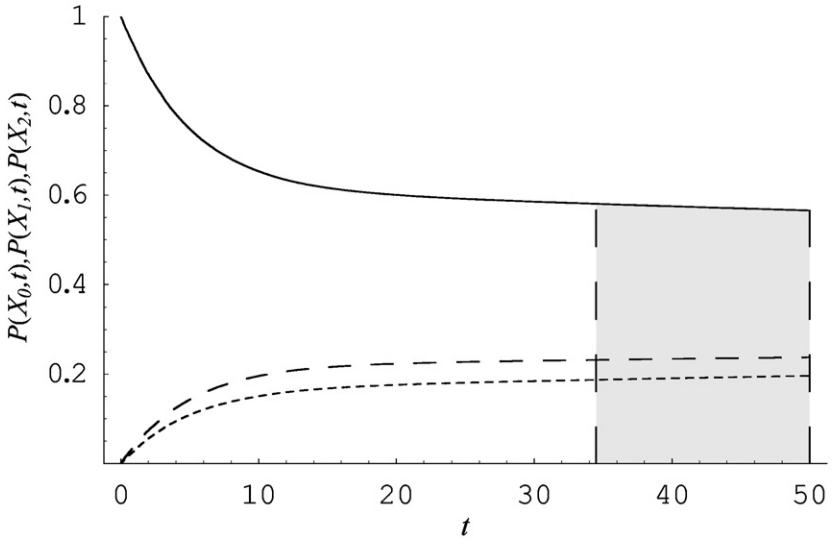


**Fig. 5.** Chaotic stochastic evolution model $CM_{TAA}$ (with low mutability of the stop codon *TAA*) of the circular codes $X_0$, $X_1$ and $X_2$ in the substitution rate barycenter: $\overline{p} = 0.8\%$, $\overline{q} = 2.6\%$ and $\overline{r} = 96.6\%$. The model $CM_{TAA}$ verifies the circular code probability inequalities (3.6) and (3.7) observed in genes with $t \geqslant 34.5$ (in grey level).

stop codon *TAA* with a probability equal to $10^{-3}$ and the other trinucleotides with equiprobability ($(1-10^{-3})/63 \approx 0.0159$). By keeping the same barycenter values $\overline{p}$, $\overline{q}$ and $\overline{r}$ (to compare the different types of evolution models), the model $CM_{TAA}$ increases the code probability difference

$$\Delta P_{CM_{TAA}}(t) = P(X_1, t) - P(X_2, t) \tag{3.10}$$

during evolution with a maximum value $\Delta P_{CM_{TAA}}(t) \approx 4.76\%$ at $t = 18.3$ which is very close to the value $\Delta P_{Genes} = 4.8\%$ (3.2) observed in genes (Fig. 5).

Similarly, the chaotic models $CM_{TAG}$ and $CM_{TGA}$ are based on low mutabilities ($10^{-3}$) of the stop codons *TAG* and *TGA*, respectively. Fig. 6 summarizes the evolution of the circular code probability differences $\Delta P(t) = P(X_1, t) - P(X_2, t)$ at time $t$ with the standard model *SM* ($\Delta P_{SM}(t)$ (3.8)) and the four chaotic models *CM* ($\Delta P_{CM}(t)$ (3.9)), $CM_{TAA}$ ($\Delta P_{CM_{TAA}}(t)$ (3.10)), $CM_{TAG}$ ($\Delta P_{CM_{TAG}}(t)$) and $CM_{TGA}$ ($\Delta P_{CM_{TGA}}(t)$). To explain the asymmetry between the codes $X_1$ and $X_2$ in genes, the model $CM_{TAG}$ is the least interesting (maximum value $\Delta P_{CM_{TAG}}(t) \approx 1.61\%$ at $t = 12.1$) and less performant than the model *SM*. The models *CM* and $CM_{TGA}$ (maximum value $\Delta P_{CM_{TGA}}(t) \approx 3.55\%$ at $t = 17.5$) have similar curve shapes and code probability differences maintained in an evolutionary time larger than the model *SM*. The chaotic model $CM_{TAA}$ with low mutability of the stop codon *TAA* matches the probability discrepancy between the circular codes $X_1$ and $X_2$ observed in real genes better than the standard model *SM* and the chaotic models *CM*, $CM_{TAG}$ and $CM_{TGA}$.

## 4. Conclusion

We have developed a new class of stochastic models of gene evolution with chaotic mutations, i.e. a random subset of the 64

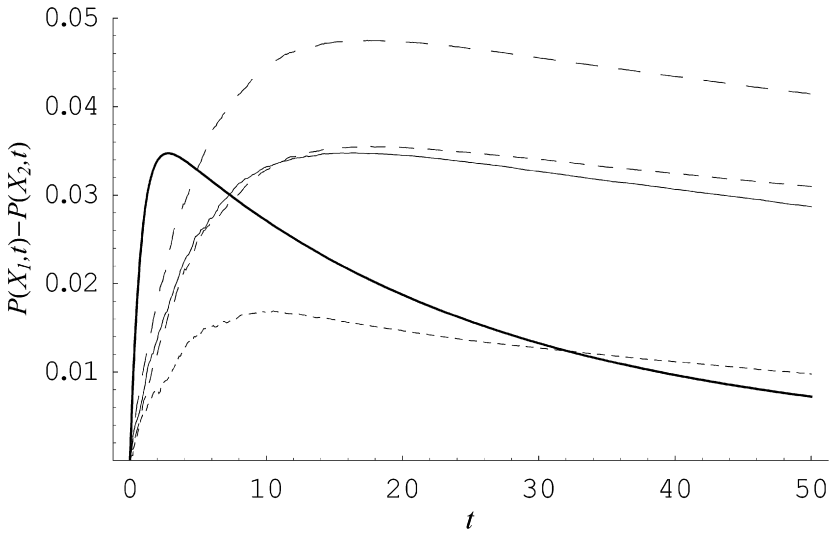


**Fig. 6.** Circular code probability differences $\Delta P = P(X_1, t) - P(X_2, t)$ between the two circular codes $X_1$ and $X_2$ with the standard model *SM* ($\Delta P_{SM}(t)$; thick full line) and the four chaotic models *CM* ($\Delta P_{CM}(t)$; thin full line), $CM_{TAA}$ ($\Delta P_{CM_{TAA}}(t)$; large dash line; top curve), $CM_{TAG}$ ($\Delta P_{CM_{TAG}}(t)$; dot line; bottom curve) and $CM_{TGA}$ ($\Delta P_{CM_{TGA}}(t)$; small dash line). The chaotic model $CM_{TAA}$ with low mutability of the stop codon *TAA* matches the probability discrepancy between the circular codes $X_1$ and $X_2$ observed in real genes better than the standard model *SM* and the chaotic models *CM*, $CM_{TAG}$ and $CM_{TGA}$.

possible trinucleotides mutates at each evolutionary time $t$ according to some substitution probabilities. Therefore, at each time $t$, the numbers and the types of mutable trinucleotides are unknown and the mutation matrix changes. Thus, the chaotic model *CM* generalizes the standard model *SM* 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 chaotically mutate according to three substitution parameters associated with the three trinucleotide sites. Two theorems prove that the chaotic model has a probability vector at each time $t$ and that it converges to a uniform probability vector identical to that of the standard model if the substitution matrix $A$ is symmetric. Four applications of this chaotic model have allowed an evolutionary study of the three 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 $X_2$.

As a chaotic model is more general than a standard model, there is no mathematical reason why a chaotic model should have similar results of a standard model. On a biological perspective, a chaotic evolution allows some trinucleotides not to mutate for a certain period of time. The evolutionary behaviour of mutable trinucleotides is modeled by the $J(k)$ parameter (see Section 2.3) 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. Any deterministic strategy for the $J(k)$ parameter converges as long as Condition 1 of Theorem 2 is satisfied.

Four chaotic models have allowed an evolutionary study of the three circular codes $X_0$, $X_1$ and $X_2$: the model *CM* with a uniform random strategy for the 64 trinucleotides and the models $CM_{TAA}$, $CM_{TAG}$ and $CM_{TGA}$ with low mutability (substitution probability close to 0) of the stop codons *TAA*, *TAG* and *TGA*, respectively. Among these five evolution models *SM*, *CM*, $CM_{TAA}$, $CM_{TAG}$ and $CM_{TGA}$, the chaotic model $CM_{TAA}$ leads to the best fit with the probability asymmetry between the codes $X_1$ and $X_2$ observed in genes. This result may be explained by the fact that *TAA* is strongly avoided in protein coding genes. Indeed, between the three stop codons, *TAA* is the most "universal" stop codon (overrepresented in prokaryotes and lower eukaryotes) while *TGA* is overrepresented only in higher eukaryotes and *TAG* is the least used (Sharp and Bulmer, 1988; Brown et al., 1990; Sun et al., 2005).

These standard and chaotic models construct "primitive" genes, i.e. genes before random substitutions ($t = 0$), with trinucleotides of the circular code $X_0$ according to an independent concatenation with equiprobability ($\frac{1}{20}$). The substitution process ($t > 0$) allows the generation of the circular code probability inequality $P(X_1, t) > P(X_2, t)$ and the increase of its probability difference during evolution. Furthermore, it retrieves the frequency orders of the three codes $X_0$, $X_1$ and $X_2$ in genes. The barycenter values in these stochastic models have a substitution rate $\overline{r}$ (96.6%) significantly greater than $\overline{q}$ (2.6%) and $\overline{p}$ (0.8%), in agreement with the actual degeneracy of the genetic code with a significantly highest mutation rate in the third site (see e.g. Ermolaeva, 2001).

The 20 trinucleotides of $X_0$ codes for 12 amino acids: *Ala*, *Asn*, *Asp*, *Gln*, *Glu*, *Gly*, *Ile*, *Leu*, *Phe*, *Thr*, *Tyr* and *Val* according to the standard and current genetic code. Five amino acids (*AA*): *Ala*, *Gln*, *Phe*, *Thr* and *Tyr* are coded by one trinucleotide of $X_0$, six *AA*: *Asn*, *Asp*, *Glu*, *Gly*, *Ile* and *Leu*, by two trinucleotides of $X_0$, and one *AA*: *Val*, by three trinucleotides of $X_0$. In current genes, these 12 *AA* are coded by 37 trinucleotides. The determination of prebiotic amino acids in comets, submarine hydrothermal systems and soup on earth is still controversial. This open biological problem in the context of the stochastic models developed here unfortunately can give no conclusion so far if the genetic code had been involved in $X_0$ or if it is appeared after mutations of $X_0$.

The probability variations of trinucleotides after random substitutions in the standard and chaotic models are totally unexpected and cannot be predicted without modelling. Particularly, the analytical solutions in the standard model are based on a sum of several exponential terms which are function of the three parameters $p$, $q$ and $r$, and the chaotic models have a very complex

random behaviour with different mutable and non-mutable trinucleotides during evolution. On the other hand, the traces of the probability differences between primitive trinucleotides (initial probabilities) are conserved even after a great number of substitutions, e.g. at $t = 30$ in the standard model *SM* (Fig. 3) and $t = 50$ in the chaotic model *CM* (Fig. 4).

Other applications of this chaotic model can also be developed in future, e.g. with particular strategies, with mutation matrices with a greater number of substitution parameters, etc. We are also investigating this new chaotic evolutionary approach in the field of phylogenetic inference.

## Acknowledgement

We thank the reviewers for their comments.

## Appendix A. Proof of Theorem 1

Let us introduce the following notations $x_i^{(k)} = P_i(t_k)$ and $x^{(k)} = (x_1^{(k)}, \ldots, x_{64}^{(k)})$. Let $x^0 = P(t_0)$ be the initial probability vector. By supposing that $\sum_{i=1}^{64} x_i^{(m)} = 1$, we will prove that for a fixed $m \in \mathbb{N}$, $\sum_{i=1}^{64} x_i^{(m+1)} = \sum_{i=1}^{64} x_i^{(m)}$. We have successively

$$\begin{aligned}\sum_{i=1}^{64} x_i^{(m+1)} &= \sum_{i\in J(m+1)} x_i^{(m+1)} + \sum_{i\notin J(m+1)} x_i^{(m+1)} \\ &= \sum_{i\in J(m+1)} x_i^{(m)} + \sum_{i\notin J(m+1)} x_i^{(m)} \\ &\quad + h \sum_{i\in J(m+1)} \sum_{j=1}^{64} (A^{(m)} - I)_{ij} x_j^{(m)}.\end{aligned}$$

Then,

$$\begin{aligned}\sum_{i=1}^{64} x_i^{(m+1)} &= \sum_{i\in J(m+1)} x_i^{(m)} + \sum_{i\notin J(m+1)} x_i^{(m)} \\ &\quad + h \sum_{i\in J(m+1)} \left( \sum_{j=1}^{64} A_{ij}^{(m)} x_j^{(m)} - x_i^{(m)} \right). \qquad (A.1)\end{aligned}$$

On the other hand, $A^{(m)}$ is by construction a probability matrix, so $\sum_{i=1}^{64}\sum_{j=1}^{64} A_{ij}^{(m)} x_j^{(m)} = 1$ and $\sum_{i\in J(m+1)}\sum_{j=1}^{64} A_{ij}^{(m)} x_j^{(m)} + \sum_{i\notin J(m+1)} \sum_{j=1}^{64} A_{ij}^{(m)} x_j^{(m)} = 1$.

According to the matrix conservation law of substitution probabilities, if $i \notin J(m+1)$ then $A_{ii}^{(m)} = 1$ and $A_{ij}^{(m)} = 0$ for $i \neq j$, hence $\sum_{i\in J(m+1)}\sum_{j=1}^{64} A_{ij}^{(m)} x_j^{(m)} + \sum_{i\notin J(m+1)} x_i^{(m)} = 1$. Thanks to the above equation, we have

$$\begin{aligned}&h \sum_{i\in J(m+1)} \left( \sum_{j=1}^{64} A_{ij}^{(m)} x_j^{(m)} - x_i^{(m)} \right) \\ &= h \sum_{i\in J(m+1)} \sum_{j=1}^{64} A_{ij}^{(m)} x_j^{(m)} - h \sum_{i\in J(m+1)} x_i^{(m)} \\ &= h \left( 1 - \sum_{i\notin J(m+1)} x_i^{(m)} \right) - h \sum_{i\in J(m+1)} x_i^{(m)}.\end{aligned}$$

By induction hypothesis, $\sum_{i=1}^{64} x_i^{(m)} = 1$, so $\sum_{i\in J(m+1)} x_i^{(m)} = 1 - \sum_{i\notin J(m+1)} x_i^{(m)}$. Then, (A.1) leads to

$$\sum_{i=1}^{64} x_i^{(m+1)} = \sum_{i\in J(m+1)} x_i^{(m)} + \sum_{i\notin J(m+1)} x_i^{(m)} = \sum_{i=1}^{64} x_i^{(m)}. \qquad (A.2)$$

The recurrence hypothesis implies that (A.2) is true for all $m \in \mathbb{N}$. □

## Appendix B. Proof of Theorem 2

Let us recall that $B^{(k)}$ denotes the matrix $hA^{(k)} + (1-h)I$. For all $i \in \{1, \ldots, 64\}$,

$$\begin{aligned}|(B^{(k)}x)_i| &= \left| (1-h)x_i + h \sum_{j=1}^{64} A_{ij}^{(k)} x_j \right| \\ &\leqslant \left| (1-h) \max_i x_i + h \max_i x_i \sum_{j=1}^{64} A_{ij}^{(k)} \right| \\ &= \max_i |x_i|.\end{aligned}$$

So, $\forall i \in \{1, \ldots, 64\}$, $\max_i |(B^{(k)}x)_i| \leqslant \max_i |x_i|$.

Let us denote the maximum norm by $|x|_\infty = \max_i |x_i|$. Then,

$$|B^{(k)}x|_\infty \leqslant |x|_\infty. \qquad (B.1)$$

The sequence $\{x_i^{(k)}\}_k$ is bounded. Indeed, suppose that $\forall 0 \leqslant k \leqslant p$, $|x^{(k)}|_\infty \leqslant |x^{(0)}|_\infty$. Then,

$$|x_i^{(p+1)}|_\infty = |(B^{(p)}x^{(p)})_i| \leqslant \max_i |x_i^{(p)}| \leqslant |x^{(0)}|_\infty.$$

So, $\forall i \in \{1, \ldots, 64\}$

$$|x^{(p+1)}|_\infty = \max_i |x_i^{(p+1)}|_\infty \leqslant |x^{(0)}|_\infty.$$

A classical recurrence reasoning leads to

$$\forall k \in \mathbb{N}, \quad |x^{(k)}|_\infty \leqslant |x^{(0)}|_\infty.$$

Hence, the sequence $\{|x^{(k)}|_\infty\}_k$ and a fortiori $\{|x^{(k_p)}|_\infty\}_p$ (see Condition 1) is bounded. Weierstrass theorem implies that it contains a convergent subsequence. Without loss of generality, suppose that it is itself, $\lim_{p\to+\infty} x^{(k_p)} = x^*$.

On the other hand and thanks to (B.1), we can easily prove that $\{|x^{(k)}|_\infty\}_k$ is monotone decreasing, so it is convergent, $\lim_{k\to+\infty} |x^{(k)}|_\infty = \lim_{p\to+\infty} |x^{(k_p)}|_\infty = |x^*|_\infty$. Let us denote the following matrix $Q$ by

$$Q = \begin{pmatrix} 1/64 & \cdots & 1/64 \\ \vdots & \ddots & \vdots \\ 1/64 & \cdots & 1/64 \end{pmatrix}.$$

Condition 1 implies that the matrices $A^{(k_p)}$ are convergent. It implies that 1 is a simple eigenvalue of $A^{(k_p)}$ and the non-bipartite condition implies that $-1$ is not an eigenvalue of $A^{(k_p)}$. As $A^{(k_p)}$ are symmetric stochastic matrices, $\lim_{p\to+\infty} (A^{(k_p)})^p = Q$.

As the graph associated with $A^{(k_p)}$ is connected, the graph associated with the matrix $B^{(k_p)} = hA^{(k_p)} + (1-h)I$ is connected and non-bipartite. Then, $B^{(k_p)}$ are symmetric stochastic matrices and converge to $Q$

$$\lim_{p\to+\infty} (B^{(k_p)})^p = Q.$$

Then,

$$|Qx^* - (B^{(k_p)})^p x^{(k_p)}|_\infty \leqslant |(Q - (B^{(k_p)})^p)x^*|_\infty + |(B^{(k_p)})^p (x^* - x^{(k_p)})|_\infty.$$

As $\lim_{p\to\infty} x^{(k_p)} = x^*$ and $\lim_{p\to\infty} (B^{(k_p)})^p = Q$, we deduce

$$\lim_{p\to+\infty} |Qx^* - (B^{(k_p)})^p x^{(k_p)}|_\infty = 0.$$

Remark first that

$$B^{(k_p)} x^{(k_p)} = x^{(k_p+1)}.$$

Now, a classical result (Berman and Plemmons, 1994) ensures that

$$B^{(k_p)} Q = Q B^{(k_p)} = Q.$$

Thus,

$$|Qx^* - (B^{(k_p)})^p x^{(k_p)}|_\infty = |(B^{(k_p)})^{p-1}(Qx^* - B^{(k_p)}x^{(k_p)})|_\infty$$
$$= |(B^{(k_p)})^{p-1}(Qx^* - x^{(k_p+1)})|_\infty.$$

Now, as $\lim_{p\to+\infty}(B^{(k_p)})^{p-1} = Q$, by the definition of the maximum norm, we deduce

$$\lim_{p\to+\infty} |Qx^* - x^{(k_p+1)}|_\infty = 0.$$

Note that $Q$ (a doubly stochastic and positive matrix) is a paracontracting matrix, e.g. Bahi (2000). The above equation implies that

$$|Qx^*|_\infty = \lim_{p\to+\infty} |x^{(k_p+1)}|_\infty = |x^*|_\infty.$$

So $Qx^* = x^*$ and thus $x^* = c(1,\ldots,1)^{\mathrm{T}}$. Now, we have

$$|x_i^{(k+1)} - x_i^*| = |(B^{(k)}(x^{(k)} - x^*))_i| \leqslant \max_i |x_i^{(k)} - x_i^*|.$$

Thus,

$$0 \leqslant |x^{(k+1)} - x^*|_\infty \leqslant |x^{(k)} - x^*|_\infty.$$

The above development implies that the sequence $|x^{(k)} - x^*|_\infty$ is monotone decreasing, so it is convergent. Thus,

$$\lim_{k\to\infty} |x^{(k)} - x^*|_\infty = \lim_{p\to\infty} |x^{(k_p)} - x^*|_\infty = 0.$$

Therefore,

$$\lim_{k\to\infty} x^{(k)} = x^* = c(1,\ldots,1)^{\mathrm{T}}.$$

By using the above proposition, $x^*$ is a probability vector and thus,

$$x^* = \tfrac{1}{64}(1,\ldots,1)^{\mathrm{T}}. \quad \square$$

## References


Akashi, H., Eyre-Walker, A., 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. 8, 688–693.

Arndt, P.F., Burge, C.B., Hwa, T., 2002. DNA sequence evolution with neighbor-dependent mutation. RECOMB'02. In: 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., 2000. Asynchronous iterative algorithms for nonexpansive linear systems. J. Parallel Distributed Comput. 60, 92–112.

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

Berman, A., Plemmons, R.J., 1994. Nonnegative Matrices in the Mathematical Sciences. Academic Press, New York.

Brown, C.M., Stockwell, P.A., Trotman, C.N.A., Tate, W.P., 1990. The signal for the termination of protein synthesis in procaryotes. Nucleic Acids Res. 18, 2079–2086.

Bulmer, M., 1991. The selection–mutation–drift theory of synonymous codon usage. Genetics 129, 897–907.

Dayhoff, M., Schwartz, R., Orcutt, B., 1978. A model of evolutionary change in protein. Atlas Protein Sequences Struct. 5, 345–352.

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

Fedorov, A., Saxonov, S., Gilbert, W., 2002. Regularities of context-dependent codon bias in eukaryotic genes. Nucleic Acids Res. 30, 1192–1197.

Frey, G., Michel, C.J., 2006. An analytical model of gene evolution with six 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.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pave, A., 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, r49–r62.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., Mercier, R., 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9, r43–r74.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89, 10915–10919.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 12–34.

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

Kelly, C., Churchill, G., 1996. Biases in amino acid replacement matrices and alignment scores due to rate heterogeneity. J. Comput. Biol. 3, 307–318.

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.

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

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.

Sharp, P.M., Bulmer, M., 1988. Selective differences among translation termination codons. Gene 63, 141–145.

Sharp, P.M., Matassi, G., 1994. Codon usage and genome evolution. Curr. Opin. Genet. Dev. 4, 851–860.

Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F., Sockett, R.E., 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 33, 1141–1153.

Sun, J., Chen, M., Xu, J., Luo, J., 2005. Relationships among stop codon usage bias, its context, isochores and gene expression level in various eukaryotes. J. Mol. Evol. 61, 437–444.

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.

Thorne, J.L., Goldman, N., 2003. Probabilistic models for the study of protein evolution. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics, second ed. Wiley, Chichester, pp. 209–226.

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.