Available online at www.sciencedirect.com
SCIENCE DIRECT
Computational Biology and Chemistry
Computational Biology and Chemistry 30 (2006) 1–11
www.elsevier.com/locate/compbiolchem

An analytical model of gene evolution with six mutation parameters: An application to archaeal circular codes

Gabriel Frey, Christian J. Michel*

Equipe de Bioinformatique Théorique, LSIIT (UMR CNRS-ULP 7005), Université Louis Pasteur de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France

Received 26 August 2005; received in revised form 4 September 2005; accepted 5 September 2005

Abstract

We develop here an analytical evolutionary model based on a trinucleotide mutation matrix 64×64 with six substitution parameters associated with the transitions and transversions in the three trinucleotide sites. It generalizes the previous models based on the nucleotide mutation matrices 4×4 and the trinucleotide mutation matrix 64×64 with three parameters. It determines at some time t the exact occurrence probabilities of trinucleotides mutating randomly according to six substitution parameters. An application of this model allows an evolutionary study of the common circular code *COM* and the 15 archaeal circular codes *X* which have been recently identified in several archaeal genomes. The main property of a circular code is the retrieval of the reading frames in genes, both locally, i.e. anywhere in genes and in particular without a start codon, and automatically with a window of a few nucleotides. In genes, the circular code is superimposed on the traditional genetic one. Very unexpectedly, the evolutionary model demonstrates that the archaeal circular codes can derive from the common circular code subjected to random substitutions with particular values for six substitutions parameters. It has a strong correlation with the statistical observations of three archaeal codes in actual genes. Furthermore, the properties of these substitution rates allow proposal of an evolutionary classification of the 15 archaeal codes into three main classes according to this model. In almost all the cases, they agree with the actual degeneracy of the genetic code with substitutions more frequent in the third trinucleotide site and with transitions more frequent that transversions in any trinucleotide site.

Keywords: Analytical model; Gene evolution; Mutation parameter

1. Introduction

1.1. Presentation of the approach

The trinucleotide distribution in (protein coding) genes is not random. Indeed, a few trinucleotides occur with high frequencies in the reading frame of genes (Grantham et al., 1980), the reading frame being the modulo 3 frame established by the codon ATG. This trinucleotide usage preference has been related to several biological factors, including translational selection (Shpaer, 1986; Akashi and Eyre-Walker, 1998), GC composition (Jukes and Bhushan, 1986; Konu and Li, 2002), strand-specific mutational bias (Sharp and Matassi, 1994; Berg and Silva, 1997; Campbell et al., 1999), transcriptional selection, RNA stability and tRNA content (Ikemura, 1985) (see also the review Ermolaeva, 2001).

In this line of research, we have identified several preferential subsets of 20 trinucleotides in the reading frame of genes (eukaryotes/prokaryotes and archaeal genomes) by developing three main statistical methods (frame autocorrelation function without bias, frame trinucleotide frequency and frame permuted trinucleotide frequency) (Arquès and Michel, 1996; Frey and Michel, 2003). The principle of these methods is simple and based on two steps:

(i) a computation of the occurrence frequencies of the 64 trinucleotides $\mathbb{T} = \{\mathrm{AAA}, \ldots, \mathrm{TTT}\}$ in the three frames of genes, i.e. the reading frame and the two shifted frames (the reading frame shifted in the 5′–3′ direction by one and two nucleotides), followed by
(ii) an assignment of a preferential frame for the 64 trinucleotides $\mathbb{T}$ by associating each trinucleotide with the frame in which it occurs with the highest frequency.

* Corresponding author. Tel.: +33 3 90 24 44 62.
E-mail addresses: frey@dpt-info.u-strasbg.fr (G. Frey), michel@dpt-info.u-strasbg.fr (C.J. Michel)

1476-9271/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.09.001

Totally unexpectedly, by excluding the four trinucleotides made of identical nucleotides $\tilde{\mathbb{T}} = \{AAA, CCC, GGG, TTT\}$, this approach has identified the common subset *COM* of 20 trinucleotides in the reading frame of genes belonging to two large and different populations of eukaryotes (26,757 sequences, 11,397,678 trinucleotides) and prokaryotes (13,686 sequences, 4,708,758 trinucleotides): $COM = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$. Furthermore, this set *COM* has interesting properties, in particular it is a circular code (Arquès and Michel, 1996; Lacan and Michel, 2001).

The observation of this preferential set *COM* of trinucleotides in various actual genes from the two largest domains, the eukaryotes and the prokaryotes, is the basis of our development of an evolutionary model. Indeed, if a trinucleotide preferential set occurs with a frequency higher than the random one in actual genes after (mainly) random mutations, then a realistic hypothesis consists in asserting that this set had a frequency in past higher than in actual time. In other words, the trinucleotides of *COM* are the basic words of the "primitive" genes (genes before evolution). Therefore, the evolutionary model proposed will be based on two processes: a construction process with a random mixing of the 20 trinucleotides of *COM* with equiprobability (1/20) followed by an evolutionary process with random substitutions which are modelled by six parameters *a*, *b*, *c*, *d*, *e* and *f* associated with the transitions and transversions in the three trinucleotide sites, respectively: *a* and *b* (resp. *c* and *d*, *e* and *f*) are the transitions (a substitution from one purine {A, G} to the other, or a substitution from one pyrimidine {C, T} to the other) and the transversions (a substitution from a purine to a pyrimidine, or reciprocally) in the first (resp. second, third) trinucleotide sites.

As the primitive genes will be constructed by trinucleotides, the mathematical model will be based on a trinucleotide mutation matrix 64×64 with six substitution parameters. Therefore, it generalizes the previous models, in particular the nucleotide mutation matrices 4×4 at one substitution parameter (Jukes and Cantor, 1969), two parameters (transitions and tranversions) (Kimura, 1980) and the trinucleotide mutation matrix 64×64 with three substitution parameters (Arquès et al., 1998).

The evolutionary model proposed here will show that the archaeal circular codes which have been recently identified in several archaeal genomes, can derive from the common circular code *COM* after a certain time of evolution and with particular values for the six substitution parameters. It has a strong correlation with the statistical observations of three archaeal codes in actual genes.

In next two Sections 1.2 and 1.3, the two stages of our approach are briefly detailed: the observation of circular codes in genes and the two processes of the evolutionary model.

1.2. Circular codes in genes

1.2.1. Definition and basic properties (detailed in Arquès and Michel, 1996, in particular)

$\mathbb{A}$ being a finite alphabet, $\mathbb{A}^*$ denotes the words on $\mathbb{A}$ of finite length including the empty word of length 0 and $\mathbb{A}^+$, the words on $\mathbb{A}$ of finite length greater or equal to 1. Let w_1w_2 be the concatenation of the two words w_1 and w_2.

Definition 1. A subset X of $\mathbb{A}^+$ is a circular code if $\forall n, m \geq 1$ 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 $n = m$, $r = 1$ and $x_i = y_i$, $1 \leq i \leq n$ (Berstel and Perrin, 1985; Béal, 1993).

A circular code is a set of words on 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. For example, let X be the six word set $X = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ and w, the word of nine letters $w = ATGGCCCTA$. The word w can be factorized circularly in two different ways: ATG, GCC, CTA and GGC, CCT, AAT. Therefore, the set X is not a circular code. But the set $\tilde{X}$ obtained by replacing the last word GGC of X by GTC, $\tilde{X} = \{AAT, ATG, CCT, CTA, GCC, GTC\}$, is a circular code as the factorizations of all the words, in particular w, are unique.

An important property of a circular code is the automatic retrieval of the construction frame of a word. Indeed, the construction frame of a word generated by a concatenation of the words of a circular code can be retrieved after the reading, anywhere in the word, of a certain number of nucleotides depending on the code. This series of nucleotides is called the window of the circular code. The main biological consequence of this property is the ability to retrieve the reading frames in genes, both locally, i.e. anywhere in genes and in particular without a start codon, and automatically with a window of a few nucleotides. Such an important property might be involved in the transcription and the translation apparatus of primitive genes.

1.2.2. A common circular code in eukaryotic and prokaryotic genes

Definition 2. The (left circular) permutation $\mathcal{P}$ of a trinucleotide $w = l_0l_1l_2$, $l_0l_1l_2 \in \mathbb{T}$, is the permuted trinucleotide $\mathcal{P}(w) = l_1l_2l_0$, e.g. $\mathcal{P}(AAC) = ACA$. This definition is naturally extended to the permutation of a trinucleotide set: The permutation $\mathcal{P}$ of a set of trinucleotides is the permuted trinucleotide set obtained by the permutation $\mathcal{P}$ of its trinucleotides.

We cite the main properties of the circular code *COM* which are defined, proved and commented in Arquès and Michel (1996) and Lacan and Michel (2001):

(i) a maximal circular code, i.e. with 20 trinucleotides, as it cannot be contained in a larger circular code, i.e. in a set with more words,
(ii) a C^3 code, i.e. a maximal circular code such that its two permuted sets are also maximal circular codes (a circular code is not necessarily a C^3 code),
(iii) a self-complementary code, i.e. 10 trinucleotides are complementary to the 10 other trinucleotides,
(iv) a rare code, as the probability that a random set of 20 trinucleotides without permuted trinucleotides is a circular code is only 6.2×10^{-8},

(v) an evolutionary flexible code, in particular with an occurrence of the four types of nucleotides in the three trinucleotide sites,
(vi) a common code with a “universal" distribution in the eukaryotic and prokaryotic genes.

The different biological consequences of these properties, in particular on the two-letter genetic alphabets, the genetic code and the amino acid frequencies in proteins, are presented, e.g. in Arquès and Michel (1996).

1.2.3. Circular codes in archaeal genomes

Archaea have features that are either unique, typically prokaryotic or typically eukaryotic (Bernander, 2000; Woese, 2000; Forterre, 2001). They possess a prokaryotic mode of cellular organisation, e.g. no nuclear envelope, circular DNA molecules organized similarly to those of prokaryotes, etc. On the other hand, they present many eukaryotic similarities in their replication, transcription and translation processes, e.g. introns in tRNA genes, protein synthesis initiation with unformylated methionine, etc.

Very surprisingly, the method based on the frame permuted trinucleotide frequency, a quantitative, sensitive and automatic statistical method for searching circular codes in genes (detailed in Frey and Michel, 2003), has recently identified 15 new circular codes X in 16 archaeal genomes G, the two genomes archeoglobus and aeropyrum having the same code (Frey and Michel, 2003; Appendices A and B). These 15 archaeal codes X are all C^3 codes but without the important property of self-complementary existing in the common C^3 code COM.

In order to quantify the preferential occurrence of an archaeal code X compared to the common code COM in an archaeal genome G, the following probabilities are defined. Let $P_i(G)$ be the occurrence probability of a trinucleotide i, $i \in \{1, \ldots, 64\}$ representing the 64 trinucleotides $\mathbb{T}$, in the genes of a genome G. As the trinucleotides $\tilde{\mathbb{T}}$ (AAA, CCC, GGG and TTT) are not considered in a circular code (by definition) and therefore, not computed in the occurrence probability of a circular code, the occurrence probability $P(X, G)$ of a code X in a genome G is renormalizing

$$P(X, G) = \frac{\sum_{i \in X} P_i(G)}{\sum_{i \in \mathbb{T} - \tilde{\mathbb{T}}} P_i(G)}.$$

Then, in an archaeal genome G, the probability difference $\Pr(X, COM, G)$ evaluates simply the preferential occurence of the code X compared to the code COM as follows

$$\Pr(X, COM, G) = P(X, G) - P(COM, G). \quad (1.1)$$

In the 16 archaeal genomes, $\Pr(X, COM, G) > 0$, $\Pr(MSA, COM, G_{MSA}) = 1.22\%$ being the lowest value (Table 1). Therefore, the 15 archaeal codes X occur with frequencies higher than the common code COM in the archaeal genomes.

The analytical evolutionary model developed in the next section, will demonstrate that the archaeal circular codes can derive from the common circular code subjected to random substitutions.

1.3. An evolutionary model based on the common circular code

Founded on the principle described in Introduction, the model is based on a construction process which generates “primitive" genes according to a random mixing of the 20 trinucleotides of the common circular code COM with equiprobability (1/20). This code COM has been chosen not only because it is “universal", as already mentioned, but also as it has stronger properties, in particular the self-complementary one, which do not exist in the archaeal codes. This process is not sufficient for retrieving the archaeal codes and an evolutionary process is added to the construction one. This evolutionary process transforms the primitive genes into simulated actual ones. Substitutions with different rates in the three sites of the 20 trinucleotides of the code COM will generate other trinucleotides, distribute them according to a non-balanced way in the hope of retrieving preferentially the trinucleotides of the actual archaeal codes.

The aim of this mathematical model consists in determining the analytical solutions of the occurrence probabilities of the common circular code COM and the 15 archaeal circular codes X as a function of the evolutionary time t and the six substitution parameters a, b, c, d, e and f (Section 2). It should be stressed that this stochastic approach with exact solutions, relies on a gene evolutionary physical model based on random substitutions in simulated sequences. However, in order to get computer results with a good approximation, a population of large sequences must be simulated in the statistical analysis, which is time consuming.

The model will demonstrate here that the archaeal circular codes can derive from the common circular code after a certain evolutionary time of random substitutions in the common code and with particular values for the six substitution parameters.

2. Mathematical model

The mathematical model will determine at an evolutionary time t the occurrence probability $P(X, t)$ of a circular code X whose trinucleotides mutate according to six real substitution parameters a, b, c, d, e and f associated with the transitions and transversions in the three trinucleotide sites: a and b (resp. c and d, e and f) are the transitions and the transversions in the first (resp. second, third) trinucleotide sites, respectively.

Table 1
Probability difference $\Pr(X, COM, G) = P(X, G) - P(COM, G)$ (in %) between an archaeal circular code X and the common circular code COM in 16 archaeal genomes G

Genome G	*AG*	*AP*	*HB*	*MC*	*MP*	*MSA*	*MSM*	*MT*	*PB*	*PCA*	*PCF*	*PCH*	*SLS*	*SLT*	*TPA*	*TPV*
$\Pr(X, COM, G)$	4.80	6.61	3.45	9.64	5.23	1.22	2.54	2.70	5.63	7.32	7.72	5.08	7.37	11.43	5.46	5.51

By convention, the indexes $i, j \in \{1, \ldots, 64\}$ represent the 64 trinucleotides $\mathbb{T}$ in alphabetical order. The occurrence probability $P_i(t + \mathrm{d}t)$ of a trinucleotide i at a time $t + \mathrm{d}t$ is equal to the occurrence probability $P_i(t)$ of this trinucleotide i at the time t minus the substitution probability of this trinucleotide i during $[t, t + \mathrm{d}t]$ and plus the substitution probabilities of the trinucleotides j, $j \neq i$, into the trinucleotide i during $[t, t + \mathrm{d}t]$

$$P_i(t + \mathrm{d}t) = P_i(t) - \alpha \mathrm{d}t P_i(t) + \alpha \mathrm{d}t \sum_{j=1}^{64} P(j \to i) P_j(t) \quad (2.1)$$

where α is the probability that a trinucleotide is subjected to one substitution during an unit interval of time and where $P(j \to i)$ is the substitution probability 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 more than one nucleotide as $\mathrm{d}t$ is assumed to be enough small that a trinucleotide cannot mutate successively two times during $\mathrm{d}t$) otherwise it is given as a function of the six substitution rates a, b, c, d, e and f. For example with the trinucleotide AAA associated with $i = 1$, $P(\mathrm{GAA} \to \mathrm{AAA}) = a$, $P(\mathrm{CAA} \to \mathrm{AAA}) = P(\mathrm{TAA} \to \mathrm{AAA}) = b/2$, $P(\mathrm{AGA} \to \mathrm{AAA}) = c$, $P(\mathrm{ACA} \to \mathrm{AAA}) = P(\mathrm{ATA} \to \mathrm{AAA}) = d/2$, $P(\mathrm{AAG} \to \mathrm{AAA}) = e$, $P(\mathrm{AAC} \to \mathrm{AAA}) = P(\mathrm{AAT} \to \mathrm{AAA}) = f/2$ and $P(j \to \mathrm{AAA}) = 0$ with $j \notin \{\mathrm{AAC}, \mathrm{AAG}, \mathrm{AAT}, \mathrm{ACA}, \mathrm{AGA}, \mathrm{ATA}, \mathrm{CAA}, \mathrm{GAA}, \mathrm{TAA}\}$.

With an appropriate unit of time, the probability α is equal to 1, i.e. there is one substitution per trinucleotide per unit of time. Then, the formula (2.1) becomes

$$\frac{P_i(t + \mathrm{d}t) - P_i(t)}{\mathrm{d}t} \approx P_i'(t) = -P_i(t) + \sum_{j=1}^{64} P(j \to i) P_j(t). \quad (2.2)$$

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

$$P'(t) = -P(t) + A \cdot P(t) = (A - I) \cdot P(t) \quad (2.3)$$

where I represents the identity matrix and the symbol $\cdot$, the matrix product.

The square 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 four square submatrices aI (16, 16) and eight square submatrices $(b/2)I$ (16, 16) as follows

$$A = \left(\begin{array}{c|cccc} & 1 \cdots 16 & 17 \cdots 32 & 33 \cdots 48 & 49 \cdots 64 \\ \hline 1 \cdots 16 & B & (b/2)I & aI & (b/2)I \\ 17 \cdots 32 & (b/2)I & B & (b/2)I & aI \\ 33 \cdots 48 & aI & (b/2)I & B & (b/2)I \\ 49 \cdots 64 & (b/2)I & aI & (b/2)I & B \end{array}\right).$$

The index ranges $\{1, \ldots, 16\}$, $\{17, \ldots, 32\}$, $\{33, \ldots, 48\}$ and $\{49, \ldots, 64\}$ are associated with the trinucleotides $\{\mathrm{AAA}, \ldots, \mathrm{ATT}\}$, $\{\mathrm{CAA}, \ldots, \mathrm{CTT}\}$, $\{\mathrm{GAA}, \ldots, \mathrm{GTT}\}$ and $\{\mathrm{TAA}, \ldots, \mathrm{TTT}\}$, 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 four square submatrices cI (4, 4) and eight square submatrices $(d/2)I$ (4, 4) as follows

$$B = \begin{pmatrix} C & (d/2)I & cI & (d/2)I \\ (d/2)I & C & (d/2)I & cI \\ cI & (d/2)I & C & (d/2)I \\ (d/2)I & cI & (d/2)I & C \end{pmatrix}.$$

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

$$C = \begin{pmatrix} 0 & f/2 & e & f/2 \\ f/2 & 0 & f/2 & e \\ e & f/2 & 0 & f/2 \\ f/2 & e & f/2 & 0 \end{pmatrix}.$$

The matrix A is stochastic when $a + b + c + d + e + f = 1$.

The differential Eq. (2.3) can then be written in the following form

$$P'(t) = M \cdot P(t)$$

with

$$M = A - I.$$

As the six substitution parameters are real, the matrix A is real and also symmetrical by construction. Therefore, the matrix M is also real and symmetrical. There exists an eigenvector matrix Q and a diagonal matrix D of eigenvalues λ_k of M ordered in the same way as the eigenvector columns in Q so that $M = Q \cdot D \cdot Q^{-1}$. Then,

$$P'(t) = Q \cdot D \cdot Q^{-1} \cdot P(t).$$

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

$$P(t) = Q \cdot \mathrm{e}^{Dt} \cdot Q^{-1} \cdot P(0) \quad (2.4)$$

where e^{Dt} is the diagonal matrix of exponential eigenvalues $\mathrm{e}^{\lambda_k t}$.

The eigenvalues λ_k of M are deduced from the eigenvalues μ_k of A such that $\lambda_k = \mu_k - 1$. The eigenvalues μ_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, the determinant $\det(A - \mu I)$ is equal to

$$\begin{aligned} \det(A - \mu I) = {} & \det(B - (-a + b + \mu)I) \\ & \times \det(B - (-a - b + \mu)I) \\ & \times [\det(B - (a + \mu)I)]^2. \end{aligned} \quad (2.5)$$

As the matrix B has a block structure similar to the matrix A, the form of the determinant $\det(B - \nu I)$ can be easily deduced from $\det(A - \mu I)$

$$\begin{aligned} \det(B - \nu I) = {} & \det(C - (-c + d + \nu)I) \\ & \times \det(C - (-c - d + \nu)I) \\ & \times [\det(C - (c + \nu)I)]^2. \end{aligned}$$

Therefore, by substituting in (2.5) $\nu = -a+b+\mu$, $\nu = -a-b+\mu$ or $\nu = a+\mu$, the determinant $\det(A-\mu I)$ becomes

$$\begin{aligned}\det(A-\mu I) = {} & \det(C-(-a+b-c+d+\mu)I)\\ & \times \det(C-(-a+b-c-d+\mu)I)\\ & \times \det(C-(-a-b-c+d+\mu)I)\\ & \times \det(C-(-a-b-c-d+\mu)I)\\ & \times [\det(C-(-a+b+c+\mu)I)]^2\\ & \times [\det(C-(-a-b+c+\mu)I)]^2\\ & \times [\det(C-(a-c+d+\mu)I)]^2\\ & \times [\det(C-(a-c-d+\mu)I)]^2\\ & \times [\det(C-(a+c+\mu)I)]^4. \end{aligned} \tag{2.6}$$

After linear combinations, the determinant $\det(C-\xi I)$ is equal to

$$\det(C-\xi I) = (e-f-\xi)(e+f-\xi)(-e-\xi)^2.$$

Therefore, by substituting in (2.6) $\xi = -a+b-c+d+\mu$, $\xi = -a+b-c-d+\mu$, $\xi = -a-b-c+d+\mu$, $\xi = -a-b-c-d+\mu$, $\xi = -a+b+c+\mu$, $\xi = -a-b+c+\mu$, $\xi = a-c+d+\mu$, $\xi = a-c-d+\mu$ or $\xi = a+c+\mu$, the determinant $\det(A-\mu I)$ is obtained

$$\begin{aligned}\det(A-\mu I) = {} & (a+b+c+d+e+f-\mu)(a+b+c+d+e-f-\mu)(a+b+c-d+e+f-\mu)\\ & \times (a+b+c-d+e-f-\mu)(a-b+c+d+e+f-\mu)(a-b+c+d+e-f-\mu)\\ & \times (a-b+c-d+e+f-\mu)(a-b+c-d+e-f-\mu)\\ & \times (a+b+c+d-e-\mu)^2(a+b+c-d-e-\mu)^2(a+b-c+e+f-\mu)^2\\ & \times (a+b-c+e-f-\mu)^2(a-b+c+d-e-\mu)^2(a-b+c-d-e-\mu)^2\\ & \times (a-b-c+e+f-\mu)^2(a-b-c+e-f-\mu)^2(-a+c+d+e+f-\mu)^2\\ & \times (-a+c+d+e-f-\mu)^2(-a+c-d+e+f-\mu)^2(-a+c-d+e-f-\mu)^2\\ & \times (a+b-c-e-\mu)^4(a-b-c-e-\mu)^4(-a+c+d-e-\mu)^4\\ & \times (-a+c-d-e-\mu)^4(-a-c+e+f-\mu)^4(-a-c+e-f-\mu)^4(-a-c-e-\mu)^8. \end{aligned}$$

Therefore, there are 27 eigenvalues λ_k of M. There are eight eigenvalues of algebraic multiplicity 1: $\lambda_1 = -1+a+b+c+d+e+f$, $\lambda_2 = -1+a+b+c+d+e-f$, $\lambda_3 = -1+a+b+c-d+e+f$, $\lambda_4 = -1+a+b+c-d+e-f$, $\lambda_5 = -1+a-b+c+d+e+f$, $\lambda_6 = -1+a-b+c+d+e-f$, $\lambda_7 = -1+a-b+c-d+e+f$ and $\lambda_8 = -1+a-b+c-d+e-f$. There are 12 eigenvalues of algebraic multiplicity 2: $\lambda_9 = -1+a+b+c+d-e$, $\lambda_{10} = -1+a+b+c-d-e$, $\lambda_{11} = -1+a+b-c+e+f$, $\lambda_{12} = -1+a+b-c+e-f$, $\lambda_{13} = -1+a-b+c+d-e$, $\lambda_{14} = -1+a-b+c-d-e$, $\lambda_{15} = -1+a-b-c+e+f$, $\lambda_{16} = -1+a-b-c+e-f$, $\lambda_{17} = -1-a+c+d+e+f$, $\lambda_{18} = -1-a+c+d+e-f$, $\lambda_{19} = -1-a+c-d+e+f$ and $\lambda_{20} = -1-a+c-d+e-f$. There are six eigenvalues of algebraic multiplicity 4: $\lambda_{21} = -1+a+b-c-e$, $\lambda_{22} = -1+a-b-c-e$, $\lambda_{23} = -1-a+c+d-e$, $\lambda_{24} = -1-a+c-d-e$, $\lambda_{25} = -1-a-c+e+f-\mu$, $\lambda_{26} = -1-a-c+e-f$. There is one eigenvalue of algebraic multiplicity 8: $\lambda_{27} = -1-a-c-e$.

The 64 eigenvectors of M associated with these 27 eigenvalues λ_k computed by formal calculus can be put in a form independent of a, b, c, d, e and f (data not shown).

The independent mixing of the 20 trinucleotides of *COM* with equiprobability (1/20) leads to the following initial vector $P(0)$ =[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, 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].

The formula (2.4) with the 64 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}, determine the 64 trinucleotide probabilities $P_i(t)$ after t substitutions as a function of the six substitution parameters a, b, c, d, e and f. As a circular code X cannot contain a trinucleotide $\tilde{\mathbb{T}}$ by definition, the occurrence probability $P(X,t)$ of a circular code X at the substitution step t, is

$$P(X,t) = \frac{\sum_{i\in X} P_i(t)}{\sum_{i\in \mathbb{T}-\tilde{\mathbb{T}}} P_i(t)}.$$

Finally, the evolutionary analytical formula $P(COM,t)$ of the common circular code *COM* as a function of the six substitution rates a, b, c, d, e and f associated with the transitions and the transversions in the three trinucleotide sites, can be expressed as a function of eigenvalues λ_k of M

$$\begin{aligned}P(COM,t) = \frac{1}{2D}(100 & + 25e^{\lambda_2 t} + e^{\lambda_4 t} + 25e^{\lambda_5 t} + 16e^{\lambda_6 t}\\ & + e^{\lambda_7 t} + 13e^{\lambda_9 t} + 5e^{\lambda_{10} t} + 36e^{\lambda_{11} t} + 2e^{\lambda_{12} t}\\ & + 5e^{\lambda_{13} t} + e^{\lambda_{14} t} + 2e^{\lambda_{15} t} + 13e^{\lambda_{17} t} + 5e^{\lambda_{18} t}\\ & + 5e^{\lambda_{19} t} + e^{\lambda_{20} t} + 2e^{\lambda_{21} t} + 22e^{\lambda_{22} t} + 6e^{\lambda_{23} t}\\ & + 2e^{\lambda_{24} t} + 2e^{\lambda_{25} t} + 22e^{\lambda_{26} t} + 8e^{\lambda_{27} t}) \end{aligned} \tag{2.7}$$

with the denominator D

$$\begin{aligned}D = {} & 150 - e^{\lambda_4 t} + 4e^{\lambda_6 t} - e^{\lambda_7 t} + e^{\lambda_{21} t} + 3e^{\lambda_{22} t}\\ & + 2e^{\lambda_{23} t} - 2e^{\lambda_{24} t} + e^{\lambda_{25} t} + 3e^{\lambda_{26} t}. \end{aligned} \tag{2.8}$$

Table 2
Initial probabilities $P(X, 0)$ of the 15 archaeal circular codes X

X	*AG/AP*	*HB*	*MC*	*MP*	*MSA*	*MSM*	*MT*	*PB*	*PCA*	*PCF*	*PCH*	*SLS*	*SLT*	*TPA*	*TPV*
$P(X, 0)$	$\frac{7}{10}$	$\frac{7}{10}$	$\frac{1}{2}$	$\frac{7}{10}$	$\frac{3}{4}$	$\frac{7}{10}$	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{7}{10}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{11}{20}$	$\frac{1}{2}$	$\frac{7}{10}$	$\frac{7}{10}$

In Appendix C, we give the evolutionary analytical formulas $P(X, t)$ of the 15 archaeal circular codes X for the reader who wants detailed results (see also Section 4).

Property 1. *The initial probability $P(COM, 0)$ (resp. $P(X, 0)$) of the code COM (resp. an archaeal code X) at the time $t = 0$ can (obviously) be obtained from the analytical solution $P(COM, t)$ (resp. $P(X, t)$) with $t = 0$ or also by a simple probability calculus.*

The probability $P(COM, 0)$ is equal to 1 as the primitive genes in this evolutionary model are generated by the code COM (20 among 20 trinucleotides).

The probability $P(X, 0)$ is also equal to the ratio of the number of common trinucleotides between COM and X to 20. These probabilities $P(X, 0)$ are given in Table 2.

Property 2. *The probability $P(COM, t)$ (resp. $P(X, t)$) of the code COM (resp. an archaeal code X) at the limit time $t \to \infty$ can (obviously) be obtained from their limit study or also by a simple probability calculus.*

Whatever $a, b, c, d, e, f \in]0, 1[$ such that $a + b + c + d + e + f = 1$, $\lim_{t\to\infty} P(COM, t) = \lim_{t\to\infty} P(X, t) = 1/3$. Indeed, the six substitutions in the 20 trinucleotides of COM, or X, generate the 44 other trinucleotides. When $t \to \infty$, the 64 trinucleotides $\mathbb{T}$ occur with the same probabilities and therefore, the probabilities of COM and X are equal to $20/60 = 1/3$ (the four trinucleotides $\tilde{\mathbb{T}}$ being not considered).

Property 3. *When one (or more) substitution has a rate equal to 0, some trinucleotides may be either not generated or generated without equiprobability and $\lim_{t\to\infty} P(COM, t) \neq 1/3$, or $\lim_{t\to\infty} P(X, t) \neq 1/3$. As an example, we explain by a simple probability calculus why $\lim_{t\to\infty} P(COM, t) = 5/12$ when $b = 0$. The code COM has 20 trinucleotides with 15 trinucleotides beginning with a purine base forming the set COM_R and five trinucleotides beginning with a pyrimidine base forming the set COM_Y, i.e. $COM = COM_R \cup COM_Y$. Each trinucleotide $w \in COM$ occurs with the same probability $P(w) = 1/20$. As there are purine and pyrimidine bases in the first trinucleotide sites of COM and as the transitions and the transversions are allowed in the second and third sites of COM ($c, d, e, f > 0$), the 64 trinucleotides $\mathbb{T}$ are generated during the evolutionary process. Among these 64 trinucleotides $\mathbb{T}$, let $\mathbb{T}_R$ be the set of the 32 trinucleotides beginning with a purine base and $\mathbb{T}_Y$, the set of the 32 trinucleotides beginning with a pyrimidine base, i.e. $\mathbb{T} = \mathbb{T}_R \cup \mathbb{T}_Y$. As in the first sites of COM the transitions are allowed ($a > 0$) but not the transversions ($b = 0$), the trinucleotide set $\mathbb{T}_R$ can only be generated from COM_R. When $t \to \infty$, the trinucleotides of COM_R and $\mathbb{T}_R$ occur with the same probability $P(w, t) = (15/20)/32 = 3/128$ with $w \in COM_R$. Similarly, when $t \to \infty$, the trinucleotides of COM_Y and $\mathbb{T}_Y$ occur with the same probability $P(w, t) = (5/20)/32 = 1/128$ with $w \in COM_Y$. The trinucleotides AAA and GGG (resp. CCC and TTT) belong to $\mathbb{T}_R$ (resp. $\mathbb{T}_Y$). Therefore, when $t \to \infty$, the trinucleotides $\tilde{\mathbb{T}}$ occur with the same probability $P(w, t) = (6 + 2)/128 = 1/16$ with $w \in \tilde{\mathbb{T}}$. Finally, $\lim_{t\to\infty} P(COM, t)$ is equal to*

$$\lim_{t\to\infty} P(COM, t) = \frac{\sum_{w\in COM_R\cup COM_Y} \lim_{t\to\infty} P(w, t)}{1 - \lim_{\substack{t\to\infty \\ w\in\tilde{\mathbb{T}}}} P(w, t)}$$

$$= \frac{(45 + 5)/128}{1 - (1/16)} = \frac{5}{12}.$$

Property 4. *The evolutionary analytical formula $Q(COM, t)$ of the common circular code COM as a function of the three substitution rates p, q and r associated with the three trinucleotide sites, respectively, is a particular case of $P(COM, t)$ with $a = p/3$, $b = 2p/3$, $c = q/3$, $d = 2q/3$, $e = r/3$ and $f = 2r/3$*

$$Q(COM, t) = \frac{1}{2\mathcal{D}}\Big(50 + 28e^{-(4/3)t} + 5e^{-(4/3)(1-p)t} + 16e^{-(4/3)(1-q)t} + 19e^{-(4/3)(1-p-q)t} + 5e^{-(4/3)(1-r)t} + 18e^{-(4/3)(1-p-r)t} + 19e^{-(4/3)(1-q-r)t}\Big)$$

with the denominator $\mathcal{D}$

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

3. Results

The 15 archaeal codes X have initial probabilities $P(X, 0)$ ranging from 0.5 to 0.75, the two codes *MC* and *SLT* having the lowest ones, and the three codes *MSA*, *MT* and *PB*, the highest ones (Table 2). All these 15 probabilities $P(X, 0)$ are significantly below than the initial probability $P(COM, 0) = 1$ of the common code *COM*. Therefore, a random mutation process seems a priori completely unable to derive an archaeal code X from the common code *COM* by decreasing the probability curve *COM* faster than an X one and then, by crossing it in order that a code X occurs with a higher probability.

The stochastic model developed here, allows the investigation of such a property by searching for a probability curve cross with each archaeal code X, i.e. by searching for the existence of a positive probability difference

$$\Pr(X, COM, t) = P(X, t) - P(COM, t) > k \tag{3.1}$$

Table 3
Substitution rate barycenters (in %) of the solution spaces for the 15 archaeal codes X such that each code X occurs with a probability higher than the common code COM (Eq. (3.1))

X	a	b	c	d	e	f	$p = a + b$	$q = c + d$	$r = e + f$
AG/AP	24.9	1.9	14.5	18.8	13.4	26.5	26.8	33.3	39.9
HB	19.3	34.9	8.0	9.3	7.5	21.0	54.2	17.3	28.5
MC	10.5	2.9	20.6	20.6	17.0	28.4	13.4	41.2	45.4
MP	20.4	30.9	2.2	4.7	12.3	29.5	51.3	6.9	41.8
MSA	14.5	4.5	18.3	19.1	17.6	26.0	19.0	37.4	43.6
MSM	15.8	4.5	18.9	18.9	18.1	23.8	20.3	37.8	41.9
MT	17.8	3.5	18.0	19.9	16.0	24.8	21.3	37.9	40.8
PB	19.9	1.8	14.7	20.0	14.4	29.2	21.7	34.7	43.6
PCA	17.5	3.1	17.4	19.8	16.2	26.0	20.6	37.2	42.2
PCF	10.2	3.5	18.6	21.0	17.0	29.7	13.7	39.6	46.7
PCH	10.6	3.4	18.8	20.5	16.8	29.9	14.0	39.3	46.7
SLS	10.6	2.9	20.3	20.4	17.2	28.6	13.5	40.7	45.8
SLT	10.6	3.0	21.2	19.9	17.6	27.7	13.6	41.1	45.3
TPA	19.9	31.6	16.4	22.6	0.3	9.2	51.5	39.0	9.5
TPV	14.3	4.0	19.9	19.1	18.6	24.1	18.3	39.0	42.7

Table 4
Substitution rate barycenters (in %) of the solution spaces for the three archaeal codes $X = \{MSA, MSM, MT\}$ such that each code X has an occurrence probability difference with the common code COM higher than the one observed in its genome (Eq. (3.2) and $\Pr(X, COM, G)$ (1.1) given in % (Table 1))

X	$\Pr(X, COM, G)$	a	b	c	d	e	f	Figure
MSA	1.22	13.8	2.8	18.3	19.9	17.6	27.6	1
MSM	2.54	15.0	1.4	19.0	20.3	18.3	26.0	2
MT	2.70	17.9	0.02	17.2	21.9	16.3	26.7	3

k being chosen equal to 0.5% for a significant difference. Each substitution rate a, b, c, d, e and f varies in the range [0,1] with a step of 1% such that their probability sum is equal to 1, and t, in the range [0,15].

Very unexpectedly, all archaeal codes X can be derived from random substitutions in the common code COM. Indeed, the difference $\Pr(X, COM, t)$ can be positive for all codes X for some values of the substitution parameters. Table 3 gives the barycenters of the solution spaces (not given) of the six substitution rates a, b, c, d, e and f for the 15 archaeal codes. The barycenter rates allow proposal of a classification of the 15 archaeal codes according to this evolutionary model with six parameters. Three main classes can be observed according to the low values of the substitution rates (Table 3):

(i) the class $\mathcal{C}_r$ with low substitutions in the third site ($e < 1\%$ and $f < 10\%$, and $r < 10\%$) containing one code TPA,
(ii) the class $\mathcal{C}_q$ with low substitutions in the second site ($c < 10\%$ and $d < 10\%$, and $q \lessapprox 15\%$) containing the codes HB and MP,
(iii) the class $\mathcal{C}_b$ with low transversions in the first site ($b < 5\%$) which can be divided into five subclasses according to the values of b:
  (iiia) the class $\mathcal{C}_{b_1}$ containing the codes AG/AP and PB with $b \approx 2\%$,
  (iiib) the class $\mathcal{C}_{b_2}$ containing the codes MC, PCA, SLS and SLT with $b \approx 3\%$,
  (iiic) the class $\mathcal{C}_{b_3}$ containing the codes MT, PCF and PCH with $b \approx 3.5\%$
  (iiid) the class $\mathcal{C}_{b_4}$ containing the code TPV with $b \approx 4\%$
  (iiie) the class $\mathcal{C}_{b_5}$ containing the codes MSA and MSM with $b \approx 4.5\%$.

However, the existence of a positive difference does not simulate reality completely. Therefore, a stronger property has been studied by chosen k equal to $\Pr(X, COM, G)$ ((1.1) and Table 1), i.e. by searching for a probability difference between each archaeal code X and the code COM which is greater than the one observed in its genome

$$\Pr(X, COM, t) = P(X, t) - P(COM, t) > \Pr(X, COM, G) \tag{3.2}$$

Three applications of the model are strongly correlated with the archaeal codes MSA, MSM and MT (Table 4).

Fig. 1 (resp. 2 and 3) gives a graphical representation of the analytical solutions $P(COM, t)$ (2.7) and $P(MSA, t)$ (resp. $P(MSM, t)$ and $P(MT, t)$) (Appendix C) in its substitution rate barycenter (Table 4). The curve $P(MSA, t)$ (resp. $P(MSM, t)$ and $P(MT, t)$) crosses $P(COM, t)$ at $t_c \approx 2.51$ (resp. 2.29 and 2.69) and is correlated with the actual genes in the achaeal genome MSA (resp. MSM and MT) at $t_a \approx 3.24$ (resp. 3.44 and 5.48) as $\Pr(MSA, COM, 3.24) \approx \Pr(MSA, COM, G_{MSA}) = 1.22\%$ (resp. $\Pr(MSM, COM, 3.44) \approx \Pr(MSM, COM, G_{MSM}) = 2.54\%$ and $\Pr(MT, COM, 5.48) \approx \Pr(MT, COM, G_{MT}) = 2.70\%$).

4. Discussion

A new analytical evolutionary model has been developed here in order to generalize several previous models based on the

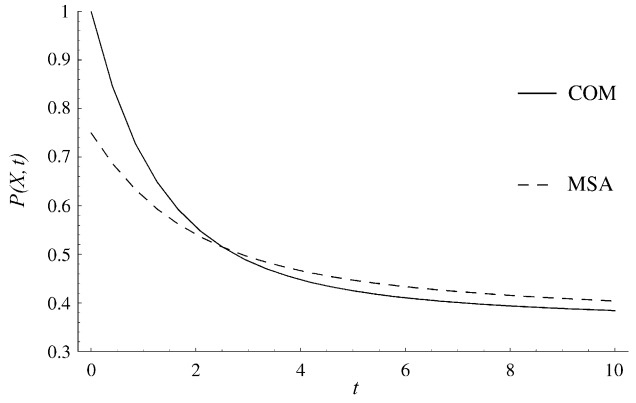

Fig. 1. Evolution of the common circular code *COM* and the archaeal circular code *MSA* in its substitution rate barycenter (in %): $a = 13.8$, $b = 2.8$, $c = 18.3$, $d = 19.9$, $e = 17.6$ and $f = 27.6$ (Table 4). The curve $P(MSA, t)$ crosses $P(COM, t)$ at $t_c \approx 2.51$ and is correlated with the actual genes of the achaeal genome *MSA* at $t_a \approx 3.24$.

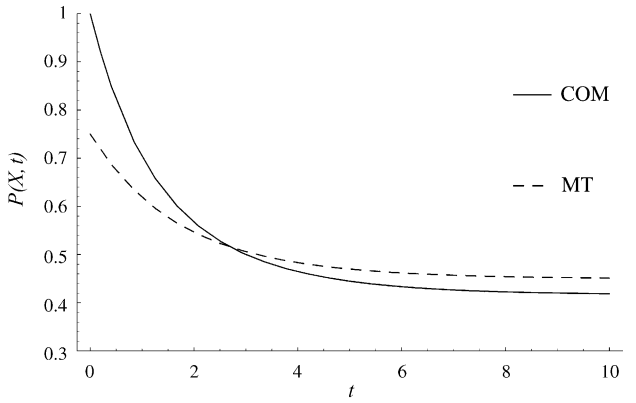

Fig. 3. Evolution of the common circular code *COM* and the archaeal circular code *MT* in its substitution rate barycenter (in %): $a = 17.9$, $b = 0.02$, $c = 17.2$, $d = 21.9$, $e = 16.3$ and $f = 26.7$ (Table 4). The curve $P(MT, t)$ crosses $P(COM, t)$ at $t_c \approx 2.69$ and is correlated with the actual genes of the archaeal genome *MT* at $t_a \approx 5.48$.

nucleotide mutation matrices 4×4 (Jukes and Cantor, 1969; Kimura, 1980) and the trinucleotide mutation matrix 64×64 at three substitution parameters (Arquès et al., 1998). It has been applied for deriving the evolutionary probabilities of the common circular code *COM* and 15 archaeal circular codes *X* as a function of the time *t* and six substitutions parameters associated with the transitions and transversions in the three trinucleotide sites.

Very unexpectedly, the archaeal codes *X* can derive from the common code *COM* subjected to random substitutions with particular values for the six substitution parameters. The model demonstrates this existence by finding a positive probability difference $\Pr(X, COM, t)$ (3.1) for all archaeal codes (Table 3). Furthermore, it has a strong correlation with the three archaeal codes *MSA*, *MSM* and *MT*. Indeed, the probability differences $\Pr(X, COM, t)$ (3.2) obtained in this model can be greater than the probability differences $\Pr(X, COM, G)$ observed in their genomes (Table 4 and Figs. 1–3). The "crossing" times t_c are 2.51, 2.29 and 2.69 for *MSA*, *MSM* and *MT*, respectively, and their "actual" times t_a, 3.24, 3.44 and 5.48, respectively. The

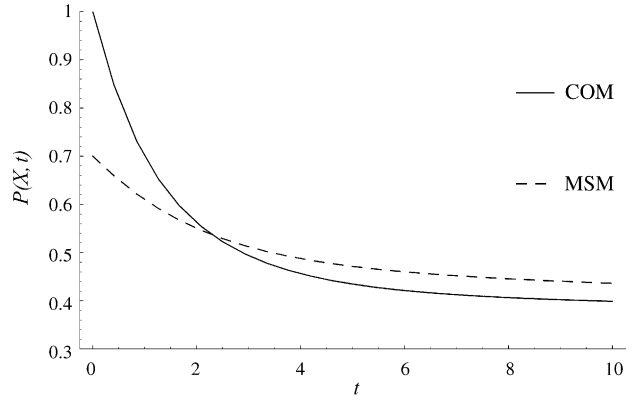

Fig. 2. Evolution of the common circular code *COM* and the archaeal circular code *MSM* in its substitution rate barycenter (in %): $a = 15.0$, $b = 1.4$, $c = 19.0$, $d = 20.3$, $e = 18.3$ and $f = 26.0$ (Table 4). The curve $P(MSM, t)$ crosses $P(COM, t)$ at $t_c \approx 2.29$ and is correlated with the actual genes of the archaeal genome *MSM* at $t_a \approx 3.44$.

values of these actual times in this model suggest a time of evolution which increases from *MSA*, *MSM* to *MT*. Note also that the shortest crossing time does not imply necessarily the shortest actual time as $t_c = 2.29$ for *MSM* and $t_a = 3.24$ for *MSA*. A strong correlation with the 12 other archaeal codes requires an improvement of this model, e.g. by adding additional parameters with a numerical model or by considering non symmetrical mutation matrices, etc.

The low values of the substitution rate barycenters allow proposal of an evolutionary classification of the 15 archaeal codes into three main classes, a class containing the code *TPA*, a class with the two codes *HB* and *MP*, and a class containing the 12 remaining codes which can be subdivided into four subclasses (Table 3 and Section 3).

The code *TPA* is the unique archaeal code with low substitutions in the third trinucleotide sites, i.e. $r < 10\%$ (Table 3), in total contradiction with the actual degeneracy of the genetic code (Ermolaeva, 2001). This result suggests that the code *TPA* is the only among the 15 identified archaeal codes which has not evolved from the common code *COM*. It can be supported by the biological fact that several genes in the archaea *TPA* have been acquired by lateral transfert from the archaea *SLS* which is only among the other archaea with living in the same thermoacidophilic environment (Ruepp et al., 2000).

The codes *HB* and *MP* have low substitutions in the second trinucleotide sites, i.e. $q < 20\%$ (Table 3). The same evolutionary class for the archaea *HB* and *MP* can be explained biologically by their high intracellular salinity which involves several specific genes not useful in the other archaea (Slesarev et al., 2002).

The 12 other archaeal codes are classified into five classes $\mathcal{C}_{b_1}$, $\mathcal{C}_{b_2}$, $\mathcal{C}_{b_3}$, $\mathcal{C}_{b_4}$ and $\mathcal{C}_{b_5}$ as a function of low transversions in the first trinucleotide sites (*b*) (Table 3). In all these five classes of codes, the rate *r* is higher than *q* which itself is higher than *p* ($r > q > p$ in $\mathcal{C}_b$, Table 3), in agreement with the actual degeneracy of the genetic code in which the substitutions are the most frequent in the third sites (Ermolaeva, 2001). Furthermore, as the transversions are associated with two mutation events com-

pared to the transitions with one mutation event (see the matrices A, B and C in Section 2), the transitions are more frequent that the transversions in each of the three sites of these five classes of codes ($a > b/2$, $c > d/2$ and $e > f/2$ in $\mathcal{C}_b$ except for one among 36 cases with *PB* in the third site, Table 3), in agreement with the chemical properties of the nucleotides (one carbon–nitrogen ring for pyrimidines and two carbon–nitrogen rings for purines) and the complementary base pairing showing a universal transition/transversion rate bias in genomes (Ochman, 2003; Rosenberg et al., 2003).

The variations of the curves $P(COM, t)$ of the common circular code *COM* and $P(X, t)$ of the 15 archaeal circular codes X, giving their trinucleotide probabilities as a function of six substitution parameters under a random evolutionary process, cannot obviously be predicted without modelling as their analytical solutions are based on a sum of several exponential terms, e.g. 23 terms for the numerator of $P(COM, t)$ (2.7). The probability differences between the trinucleotides in the primitive genes (at $t = 0$), have still some effects after a great number of random substitutions in genes, e.g. at $t = 10$ in Figs. 1–3. The primitive traces generated by these trinucleotide variations, can still be observed after a long period of random evolution, even if noise increases. Several properties with these probability curves have been observed for particular values of the six substitution rates: curves with crossings, curves with a local minimum, curves with a continuous increase during the random evolutionary process, curves with a series of fusions and separations, etc. (data not shown). These properties have not been investigated as they are not direcly in the subject of this paper. However, as already mentioned in Section 2, the evolutionary analytical formulas $P(X, t)$ of the 15 archaeal circular codes X are given in Appendix C for the reader who wants to deepen the analysis of these stochastic curves.

The biological meaning of this evolutionary model would suggest that the primitive genes (at $t = 0$), are constructed by trinucleotides of the common circular code *COM*. Only 20 among 64 trinucleotides would have been necessary. The 20 types of trinucleotides as well as the type of their concatenation are determined in this model. Indeed, the 20 trinucleotides are defined by the set *COM* which is a maximal self-complementary C^3 code (Section 1.2.2). Furthermore, the independent concatenation of these 20 trinucleotides with equiprobability is the simplest type of concatenation and therefore, compatible with a primitive stage of gene evolution. A Markov concatenation of trinucleotides (based on a stochastic matrix) would have been too complex at this primitive time. The model developed here has demonstrated that the 15 circular codes observed in archaeal genomes can derive from the common circular code subjected to random substitutions with particular values for the six substitutions parameters associated with the transitions and transversions in the three trinucleotide sites. Furthermore, it has a strong correlation with three archaeal codes.

Finally, the proposed method can be applied to other problems. In particular, the eigenvalues obtained here can be directly used to develop similar evolutionary models based on a trinucleotide mutation matrix with six substitution parameters. Such a trinucleotide mutation matrix could also improve some algorithms of phylogenetic tree reconstruction and sequence alignment.

Appendix A. List of the 16 archaeal genomes studied and their abbrevations

Genome G_{AG}: Archeoglobus (fulgidus) with 2407 genes containing 1989 kb (Euryarchaeota EA)
Genome G_{AP}: Aeropyrum (pernix) with 2694 genes containing 1916 kb (Crenarchaeota CA)
Genome G_{HB}: Halobacterium (sp.NCR-1) with 2058 genes containing 1761 kb (EA)
Genome G_{MC}: Methanoccocus (jannashii) with 1709 genes containing 1444 kb (EA)
Genome G_{MP}: Methanopyrus (kandleri) with 1678 genes containing 1492 kb (EA)
Genome G_{MSA}: Methanosarcina acetivorans with 4440 genes containing 4162 kb (EA)
Genome G_{MSM}: Methanosarcina mazei with 3371 genes containing 3065 kb (EA)
Genome G_{MT}: Methanothermobacter (thermautotrophicus) with 1868 genes containing 1575 kb (EA)
Genome G_{PB}: Pyrobaculum (aerophilum) with 2605 genes containing 1968 kb (CA)
Genome G_{PCA}: Pyroccocus abyssi with 1762 genes containing 1606 kb (EA)
Genome G_{PCF}: Pyroccocus furiosus with 2060 genes containing 1740 kb (EA)
Genome G_{PCH}: Pyroccocus horikoshii with 2058 genes containing 1704 kb (EA)
Genome G_{SLS}: Sulfolobus solfataricus with 2994 genes containing 2525 kb (CA)
Genome G_{SLT}: Sulfolobus tokodaii with 2826 genes containing 2276 kb (CA)
Genome G_{TPA}: Thermoplasma acidophilium with 1478 genes containing 1359 kb (EA)
Genome G_{TPV}: Thermoplasma volcanium with 1523 genes containing 1353 kb (EA)

Appendix B. List of the 15 archaeal circular codes

Code AG/AP: AAC, AAG, ATA, ACC, GAC, TAC, AGC, GAG, GTA, ATC, ATG, ATT, GCC, CTC, GCG, GTC, CTG, TTC, GTG, GTT
Code *HB*: AAC, AAG, AAT, ACC, GAC, TAC, CAG, GAG, TAG, ATC, ATG, TAT, GCC, CTC, GGC, GTC, CTG, TTC, GTG, TTG
Code *MC*: ACA, GAA, ATA, CCA, GAC, ACT, GCA, GGA, GTA, TCA, GAT, ATT, GCC, CCT, GCG, GTC, GCT, TCT, GGT, GTT
Code *MP*: AAC, AAG, ATA, ACC, GAC, TAC, CAG, GAG, GTA, ATC, ATG, ATT, GCC, CTC, GCG, GTC, CTG, TTC, GTG, TTG
Code *MSA*: AAC, GAA, ATA, ACC, GAC, TAC, GCA, GAG, GTA, ATC, GAT, ATT, GCC, CTC, GGC, GTC, GCT, CTT, GTG, GTT

Code *MSM*: AAC, GAA, ATA, ACC, GAC, ACT, GCA, GAG, GTA, ATC, GAT, ATT, GCC, CTC, GGC, GTC, GCT, CTT, GTG, GTT
Code *MT*: AAC, AAG, ATA, ACC, GAC, TAC, GCA, GAG, GTA, ATC, GAT, ATT, GCC, CTC, GGC, GTC, GCT, TTC, GTG, GTT
Code *PB*: AAC, GAA, ATA, ACC, GAC, TAC, GCA, GAG, GTA, ATC, ATG, ATT, GCC, CTC, GCG, GTC, CTG, TTC, GTG, GTT
Code *PCA*: AAC, AAG, ATA, ACC, GAC, TAC, GCA, GAG, GTA, ATC, GAT, ATT, GCC, CTC, GCG, GTC, GCT, TTC, GTG, GTT
Code *PCF*: ACA, GAA, ATA, CCA, GAC, CTA, GCA, GAG, GTA, ATC, GAT, ATT, GCC, CTC, GCG, GTC, GCT, CTT, GTG, GTT
Code *PCH*: ACA, GAA, ATA, CCA, GAC, CTA, GCA, GAG, GTA, ATC, GAT, ATT, GCC, CTC, GCG, GTC, GCT, TTC, GTG, GTT
Code *SLS*: ACA, GAA, ATA, CCA, GAC, ACT, GCA, GAG, GTA, TCA, GAT, ATT, GCC, CCT, GCG, GTC, GCT, TCT, GGT, GTT
Code *SLT*: ACA, GAA, ATA, CCA, GAC, ACT, GCA, GGA, GTA, TCA, GAT, ATT, GCC, CCT, GGC, GTC, GCT, TCT, GGT, GTT
Code *TPA*: AAC, AAG, ATA, ACC, GAC, TAC, AGC, GAG, GTA, ATC, ATG, ATT, GCC, CTC, GGC, GTC, CTG, TTC, GTG, GTT
Code *TPV*: AAC, GAA, ATA, ACC, GAC, ACT, GCA, GAG, GTA, ATC, GAT, ATT, GCC, CTC, GGC, GTC, GCT, CTT, GGT, GTT

Appendix C. Evolutionary analytical formulas of the archaeal circular codes

With the denominator D (2.8) and the eigenvalues λ_k of M, the evolutionary analytical formulas $P(X,t)$ of the 15 archaeal circular codes X obtained are

$$P(AG/AP,t)=\frac{1}{D}(50+5e^{\lambda_2 t}+15e^{\lambda_5 t}-e^{\lambda_7 t}+11e^{\lambda_9 t}+18e^{\lambda_{11} t}-2e^{\lambda_{13} t}+e^{\lambda_{15} t}+2e^{\lambda_{18} t}-e^{\lambda_{19} t}+2e^{\lambda_{21} t}+9e^{\lambda_{22} t}+e^{\lambda_{23} t}-e^{\lambda_{26} t}+3e^{\lambda_{27} t})$$

$$P(HB,t)=\frac{1}{2D}(100+10e^{\lambda_2 t}+10e^{\lambda_5 t}+8e^{\lambda_6 t}+24e^{\lambda_9 t}+6e^{\lambda_{10} t}+42e^{\lambda_{11} t}-2e^{\lambda_{12} t}-2e^{\lambda_{13} t}-e^{\lambda_{17} t}+3e^{\lambda_{18} t}+e^{\lambda_{19} t}+e^{\lambda_{20} t}+4e^{\lambda_{21} t}+8e^{\lambda_{22} t}+12e^{\lambda_{26} t})$$

$$P(MC,t)=\frac{1}{2D}(100+5e^{\lambda_2 t}+30e^{\lambda_5 t}+4e^{\lambda_6 t}-e^{\lambda_7 t}-11e^{\lambda_9 t}-6e^{\lambda_{11} t}+3e^{\lambda_{12} t}+6e^{\lambda_{13} t}-e^{\lambda_{14} t}+e^{\lambda_{15} t}+20e^{\lambda_{17} t}-e^{\lambda_{18} t}+3e^{\lambda_{19} t}+e^{\lambda_{21} t}+13e^{\lambda_{22} t}+3e^{\lambda_{23} t}-5e^{\lambda_{24} t}-3e^{\lambda_{25} t}-3e^{\lambda_{26} t}+2e^{\lambda_{27} t})$$

$$P(MP,t)=\frac{1}{2D}(100+20e^{\lambda_5 t}-2e^{\lambda_7 t}+23e^{\lambda_9 t}+3e^{\lambda_{10} t}+42e^{\lambda_{11} t}-2e^{\lambda_{12} t}-e^{\lambda_{13} t}+e^{\lambda_{14} t}+2e^{\lambda_{15} t}+4e^{\lambda_{18} t}+2e^{\lambda_{20} t}+6e^{\lambda_{21} t}+16e^{\lambda_{22} t}+2e^{\lambda_{23} t}-2e^{\lambda_{25} t}+4e^{\lambda_{26} t}+6e^{\lambda_{27} t})$$

$$P(MSA,t)=\frac{1}{2D}(100+20e^{\lambda_2 t}+e^{\lambda_4 t}+35e^{\lambda_5 t}+4e^{\lambda_6 t}-2e^{\lambda_7 t}+9e^{\lambda_9 t}-2e^{\lambda_{10} t}+30e^{\lambda_{11} t}+e^{\lambda_{12} t}+2e^{\lambda_{13} t}+e^{\lambda_{14} t}+3e^{\lambda_{15} t}+18e^{\lambda_{17} t}-e^{\lambda_{18} t}+3e^{\lambda_{19} t}+2e^{\lambda_{20} t}+e^{\lambda_{21} t}+5e^{\lambda_{22} t}+e^{\lambda_{23} t}+e^{\lambda_{24} t}+e^{\lambda_{25} t}+9e^{\lambda_{26} t}-2e^{\lambda_{27} t})$$

$$P(MSM,t)=\frac{1}{D}(50+10e^{\lambda_2 t}+e^{\lambda_4 t}+20e^{\lambda_5 t}+4e^{\lambda_6 t}-e^{\lambda_7 t}+2e^{\lambda_9 t}-e^{\lambda_{10} t}+12e^{\lambda_{11} t}+e^{\lambda_{13} t}+2e^{\lambda_{15} t}+8e^{\lambda_{17} t}-e^{\lambda_{18} t}+2e^{\lambda_{19} t}+e^{\lambda_{20} t}+e^{\lambda_{25} t}+3e^{\lambda_{26} t}-2e^{\lambda_{27} t})$$

$$P(MT,t)=\frac{1}{2D}(100+20e^{\lambda_2 t}+e^{\lambda_4 t}+35e^{\lambda_5 t}+4e^{\lambda_6 t}-2e^{\lambda_7 t}+15e^{\lambda_9 t}+2e^{\lambda_{10} t}+30e^{\lambda_{11} t}+e^{\lambda_{12} t}+e^{\lambda_{14} t}+3e^{\lambda_{15} t}+12e^{\lambda_{17} t}+e^{\lambda_{18} t}-e^{\lambda_{19} t}+2e^{\lambda_{20} t}+3e^{\lambda_{21} t}+11e^{\lambda_{22} t}+e^{\lambda_{23} t}+e^{\lambda_{24} t}-e^{\lambda_{25} t}+3e^{\lambda_{26} t}-2e^{\lambda_{27} t})$$

$$P(PB,t)=\frac{1}{2D}(100+5e^{\lambda_2 t}+30e^{\lambda_5 t}-4e^{\lambda_6 t}-3e^{\lambda_7 t}+18e^{\lambda_9 t}+e^{\lambda_{10} t}+36e^{\lambda_{11} t}+e^{\lambda_{12} t}+e^{\lambda_{13} t}+3e^{\lambda_{15} t}+10e^{\lambda_{17} t}+5e^{\lambda_{18} t}+e^{\lambda_{19} t}+3e^{\lambda_{21} t}+17e^{\lambda_{22} t}+3e^{\lambda_{23} t}+e^{\lambda_{24} t}-e^{\lambda_{25} t}+9e^{\lambda_{26} t}+4e^{\lambda_{27} t})$$

$$P(PCA,t)=\frac{1}{2D}(100+15e^{\lambda_2 t}+e^{\lambda_4 t}+35e^{\lambda_5 t}-3e^{\lambda_7 t}+13e^{\lambda_9 t}+3e^{\lambda_{10} t}+30e^{\lambda_{11} t}+2e^{\lambda_{12} t}-e^{\lambda_{13} t}+e^{\lambda_{14} t}+4e^{\lambda_{15} t}+12e^{\lambda_{17} t}+2e^{\lambda_{18} t}-4e^{\lambda_{19} t}+2e^{\lambda_{20} t}+4e^{\lambda_{21} t}+12e^{\lambda_{22} t}-2e^{\lambda_{25} t}-2e^{\lambda_{26} t})$$

$$P(PCF,t)=\frac{1}{2D}(100+5e^{\lambda_2 t}+30e^{\lambda_5 t}-4e^{\lambda_6 t}-3e^{\lambda_7 t}+e^{\lambda_9 t}+24e^{\lambda_{11} t}+5e^{\lambda_{12} t}+2e^{\lambda_{13} t}+e^{\lambda_{14} t}+3e^{\lambda_{15} t}+21e^{\lambda_{17} t}+2e^{\lambda_{19} t}+e^{\lambda_{20} t}+e^{\lambda_{21} t}+3e^{\lambda_{22} t}-3e^{\lambda_{23} t}-e^{\lambda_{24} t}-e^{\lambda_{25} t}+e^{\lambda_{26} t}+4e^{\lambda_{27} t})$$

$$P(PCH,t) = \frac{1}{2D}(100 + 30e^{\lambda_5 t} - 2e^{\lambda_7 t} + 3e^{\lambda_9 t} + 5e^{\lambda_{10} t} + 24e^{\lambda_{11} t} + 4e^{\lambda_{12} t} + e^{\lambda_{13} t} - e^{\lambda_{14} t} + 4e^{\lambda_{15} t} + 21e^{\lambda_{17} t} + 3e^{\lambda_{18} t} + e^{\lambda_{19} t} - e^{\lambda_{20} t} + 8e^{\lambda_{22} t} - 6e^{\lambda_{23} t} - 2e^{\lambda_{24} t} - 2e^{\lambda_{25} t} + 2e^{\lambda_{26} t})$$

$$P(SLS,t) = \frac{1}{2D}(100 + 5e^{\lambda_2 t} + 30e^{\lambda_5 t} + 4e^{\lambda_6 t} - e^{\lambda_7 t} - 10e^{\lambda_9 t} + e^{\lambda_{10} t} + 3e^{\lambda_{12} t} + 3e^{\lambda_{13} t} + 3e^{\lambda_{15} t} + 20e^{\lambda_{17} t} - e^{\lambda_{18} t} + 3e^{\lambda_{19} t} + e^{\lambda_{21} t} + 13e^{\lambda_{22} t} + e^{\lambda_{23} t} - 3e^{\lambda_{24} t} - 3e^{\lambda_{25} t} + 5e^{\lambda_{26} t} + 2e^{\lambda_{27} t})$$

$$P(SLT,t) = \frac{1}{D}(50 + 5e^{\lambda_2 t} + 15e^{\lambda_5 t} + 4e^{\lambda_6 t} - 7e^{\lambda_9 t} - 2e^{\lambda_{10} t} - 6e^{\lambda_{11} t} + 3e^{\lambda_{13} t} + 10e^{\lambda_{17} t} - e^{\lambda_{18} t} + 3e^{\lambda_{19} t} + 6e^{\lambda_{22} t} + e^{\lambda_{23} t} - e^{\lambda_{24} t})$$

$$P(TPA,t) = \frac{1}{2D}(100 + 15e^{\lambda_2 t} - e^{\lambda_4 t} + 25e^{\lambda_5 t} - e^{\lambda_7 t} + 24e^{\lambda_9 t} + 2e^{\lambda_{10} t} + 36e^{\lambda_{11} t} - 2e^{\lambda_{12} t} - 4e^{\lambda_{13} t} + 2e^{\lambda_{17} t} + 4e^{\lambda_{18} t} + 2e^{\lambda_{21} t} + 12e^{\lambda_{22} t} + 2e^{\lambda_{23} t} + 2e^{\lambda_{25} t} + 4e^{\lambda_{26} t} + 2e^{\lambda_{27} t})$$

$$P(TPV,t) = \frac{1}{2D}(100 + 20e^{\lambda_2 t} + e^{\lambda_4 t} + 35e^{\lambda_5 t} + 12e^{\lambda_6 t} - 2e^{\lambda_9 t} - e^{\lambda_{10} t} + 18e^{\lambda_{11} t} + 3e^{\lambda_{12} t} + 3e^{\lambda_{13} t} + 3e^{\lambda_{15} t} + 19e^{\lambda_{17} t} - 2e^{\lambda_{18} t} + 6e^{\lambda_{19} t} + e^{\lambda_{20} t} + e^{\lambda_{21} t} - e^{\lambda_{22} t} - e^{\lambda_{23} t} + e^{\lambda_{24} t} - e^{\lambda_{25} t} + 9e^{\lambda_{26} t})$$

References

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

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.

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

Béal, M.-P., 1993. Codage Symbolique. Masson, Paris.

Berg, O.G., Silva, P.J.N., 1997. Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection. Nucleic Acids Res. 25, 1397–1404.

Bernander, R., 2000. Chromosome replication, nucleotid segregation and cell division in archaea. Trends in Microbiol. 8, 278–283.

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

Campbell, A., Mrázek, J., Karlin, S., 1999. Genomic signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc. Natl. Acad. Sci. USA 96, 9184–9189.

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

Frey, G., Michel, C.J., 2003. Circular codes in archaeal genomes. J. Theor. Biol. 223, 413–431.

Forterre, P., 2001. Genomics and early cellular evolution. The origin of the DNA world. C. R. Acad. Sci. III 324, 1067–1076.

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

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

Jukes, T.H., Bhushan, V., 1986. Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. J. Mol. Evol. 24, 39–44.

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

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.

Konu, O., Li, M.D., 2002. Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents. J. Mol. Evol. 54, 35–41.

Lacan, J., Michel, C.J., 2001. Analysis of a circular code model. J. Theor. Biol. 213, 159–170.

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

Ochman, H., 2003. Neutral mutations and neutral substitutions in bacterial genomes. Mol. Biol. Evol. 20, 2091–2096.

Rosenberg, M.S., Subramanian, S., Kumar, S., 2003. Patterns of transitional mutation biases within and among mammalian genomes. Mol. Biol. Evol. 20, 988–993.

Ruepp, A., et al. 2000., The genome sequence of the thermoacidophilic scavenger thermoplasma acidophilum. Nature 407, 508–513.

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

Shpaer, E.G., 1986. Constraints on codon context in *Escherichia coli* genes. Their possible role in modulating the efficiency of translation. J. Mol. Biol. 188, 555–564.

Slesarev, A., et al. 2002., The complete genome of hypermetrohile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. Proc. Natl. Acad. Sci. USA 99, 4644–4649.

Woese, C.R., 2000. Interpreting the universal phylogenetic tree. Proc. Natl. Acad. Sci. USA 97, 8392–8396.