Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 223 (2003) 413–431
Journal of Theoretical Biology
www.elsevier.com/locate/jtbi

# Circular codes in archaeal genomes

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 7 November 2002; received in revised form 27 February 2003; accepted 5 March 2003

**Abstract**

A new statistical method associating each trinucleotide with a frame is developed for identifying circular codes. Its sensibility allows the detection of several circular codes in the (protein coding) genes of archaeal genomes. Several properties of these circular codes are described, in particular the lengths of the minimal windows to retrieve the construction frames, a new definition of a parameter for measuring some probabilities of words generated by the circular codes, and the types of nucleotides in the trinucleotide sites. Some biological consequences are presented in Discussion.


*Keywords:* Circular codes; Archaeal genomes; Genes; Statistical method; Trinucleotide; Frame

## 1. Introduction

The trinucleotide distribution in the (protein coding) genes is not random. Indeed, the synonymous codons (codons coding for the same amino acid) do not occur with the same frequencies in the genes of different species (Grantham et al., 1980; Sharp and Matassi, 1994; Antezana and Kreitman, 1999). This codon usage bias depends on several biological factors, mainly the gene function, the DNA spatial structure (e.g. stacking energies, curvature, superhelicity) and the evolutionary process. The gene function appears to be the most important factor (Ma et al., 2002) compared to the spatial (e.g. Tazi and Bird, 1990) and evolutionary (e.g. Smith et al., 1985) ones. Several processes involved in the gene function may determine the codon usage bias in order to increase translational efficiency and accuracy (Table 1).

The statistical methods leading to these previous results are usually simple and mainly based on the occurrence frequencies of basic motifs in genes, such as the mononucleotides, the dinucleotides, the codons, etc. In this line of research, we have added a new method called frame trinucleotide frequency (FTF) which analyses the trinucleotide occurrence frequencies in the 3 frames of genes (Arquès and Michel, 1996). By convention, the reading frame established by the codon ATG is frame 0, and frames 1 and 2 are the reading frames shifted by 1 and 2 nucleotides in the 5′–3′ direction, respectively. Then, a preferential frame for the 64 trinucleotides can be deduced by assigning to each trinucleotide the frame associated with its highest occurrence frequency. Totally unexpected, by excluding the identical trinucleotides (AAA, CCC, GGG and TTT) and with a few exceptions, this approach identifies the same 3 subsets of 20 trinucleotides per frame in the 2 gene populations $P$ of eukaryotes EUK and prokaryotes PRO. These 3 sets, $\mathscr{X}_0(\mathrm{EUK}) = \mathscr{X}_0(\mathrm{PRO})$ (resp. $\mathscr{X}_1(\mathrm{EUK}) = \mathscr{X}_1(\mathrm{PRO})$ and $\mathscr{X}_2(\mathrm{EUK}) = \mathscr{X}_2(\mathrm{PRO})$) associated with the frame 0 (resp. 1 and 2) have several interesting properties, in particular the property of circular code which is briefly recalled here without mathematical notations. The other properties, such as permutation and complementarity, are detailed in Arquès and Michel (1996, 1997).

### 1.1. The 3 sets $\mathscr{X}_0(P)$, $\mathscr{X}_1(P)$ and $\mathscr{X}_2(P)$, $P = \{\mathrm{EUK}, \mathrm{PRO}\}$, are (maximal) circular codes

A circular code is a set of words on an alphabet such as any word written on a circle (the next letter after the last letter of the word being the first letter) has at most

*Corresponding author. Tel.: +33-3-90-24-44-62; fax: +33-3-90-24-44-55.
*E-mail addresses:* frey@dpt-info.u-strasbg.fr (G. Frey), michel@dpt-info.u-strasbg.fr (C.J. Michel).

0022-5193/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0022-5193(03)00119-X

Table 1
Gene function process involved in the codon usage bias

| Type of gene function process | Some references |
|---|---|
| tRNA pool and codon–anticodon binding strength | Ikemura (1981); Ikemura (1985) |
| Codon context | Yarus and Folley (1984); Shpaer (1986); Gutman and Hatfield (1989); Berg and Silva (1997) |
| Genome signature | Campbell et al. (1999) |
| Global/local G + C content and mutational biases | Jukes and Bhushan (1986); Sharp and Matassi (1994); Sueoka (1992) |
| Gene expression | Grantham et al. (1981) |
| mRNA stability and transcription rate | Andersson and Kurland (1990) |
| Effects of DNA polymerase and repair systems, methylation, CpG islands | Hanai and Wada (1988) |
| Tertiary structure of the protein coded by the gene | Gu et al. (2002) |

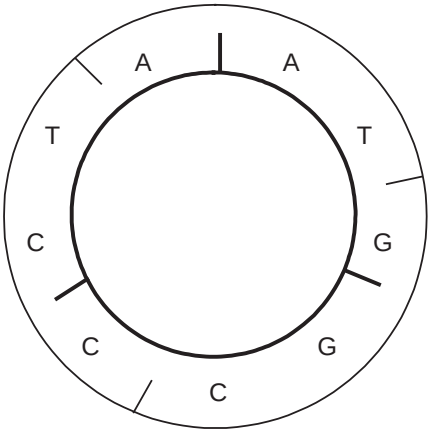


Fig. 1. The set $\mathscr{X} = \{\mathrm{AAT}, \mathrm{ATG}, \mathrm{CCT}, \mathrm{CTA}, \mathrm{GCC}, \mathrm{GGC}\}$ is not a circular code as the word $w = \mathrm{ATGGCCCTA}$, written on a circle, can be factorized into words of $\mathscr{X}$ according to 2 different ways: ATG, GCC, CTA (thick line) and AAT, GGC, CCT) (thin line).

one decomposition into words of the circular code. As an example, let the set $\mathscr{X}$ be composed of the 6 following words: $\mathscr{X} = \{\mathrm{AAT}, \mathrm{ATG}, \mathrm{CCT}, \mathrm{CTA}, \mathrm{GCC}, \mathrm{GGC}\}$ and the word $w$, be a series of the 9 following letters: $w = \mathrm{ATGGCCCTA}$. The word $w$, written on a circle, can be factorized into words of $\mathscr{X}$ according to 2 different ways: ATG, GCC, CTA and AAT, GGC, CCT (Fig. 1). Therefore, $\mathscr{X}$ is not a circular code. In contrast, if the set $\mathscr{X}'$ obtained by replacing the last word GGC of $\mathscr{X}$ by GTC is considered, i.e. $\mathscr{X}' = \{\mathrm{AAT}, \mathrm{ATG}, \mathrm{CCT}, \mathrm{CTA}, \mathrm{GCC}, \mathrm{GTC}\}$, then it does not exist such an ambiguous word with $\mathscr{X}'$. Then, $\mathscr{X}'$ is a circular code. The construction frame of a word generated by a concatenation of the words of the 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.

For a set of words with 3 letters (trinucleotides) on a 4-letter alphabet, a circular code contains at most 20 words and is maximal. Indeed, a necessary, but not sufficient, condition for a code to be circular is the absence of 2 permutated words in the code (Section 3.3 for details).

### 1.2. The code $\mathscr{X}_0(P)$ is a $\mathscr{C}^3$ code

Recall the definition of the trinucleotide (left circular) permutation: The (left circular) permutation $\mathscr{P}$ of the trinucleotide $w_0 = l_0l_1l_2$, $l_0l_1l_2 \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, is the permutated trinucleotide $\mathscr{P}(w_0) = w_1 = l_1l_2l_0$, e.g. $\mathscr{P}(\mathrm{AAC}) = \mathrm{ACA}$, and $\mathscr{P}(\mathscr{P}(w_0)) = \mathscr{P}(w_1) = w_2$, e.g. $\mathscr{P}(\mathscr{P}(\mathrm{AAC})) = \mathrm{CAA}$.

This definition is naturally extended to the trinucleotide set permutation: The permutation $\mathscr{P}$ of a set of trinucleotides is the permutated trinucleotide set obtained by the permutation $\mathscr{P}$ of all its trinucleotides.

Each circular code can be deduced from the permutation of another circular code. The code $\mathscr{X}_1(P)$ associated with the frame defined as the (left) permutation of the reading frame, i.e. the frame 1, can be obtained by the permutation of the code $\mathscr{X}_0(P)$. Similarly, the code $\mathscr{X}_2(P)$ can be deduced from the permutation of the code $\mathscr{X}_1(P)$ and the code $\mathscr{X}_0(P)$, from the code $\mathscr{X}_2(P)$. As the code $\mathscr{X}_0(P)$ is coding for the reading frame in genes, it is considered as the main code and then called $\mathscr{C}^3$ code (maximal circular code with 2 permutated maximal circular codes).

*Remark*: The property that $\mathscr{X}_0(P)$ is a circular code, does not necessarily imply that $\mathscr{X}_1(P)$ and $\mathscr{X}_2(P)$ are also circular codes.

Another $\mathscr{C}^3$ code without the complementary property has been found in the mitochondrial genes (Arquès and Michel, 1997).

The main biological property directly related to a circular code is the ability to retrieve the reading frame which might be involved in the transcription and the translation apparatus (Arquès and Michel, 1996, 1997).

### 1.3. The developed method

The method FTF, which was used for determining the $\mathscr{C}^3$ codes in the gene populations of eukaryotes, prokaryotes and mitochondria, presents exceptions for the assignments of preferential frames to some trinucleotides. These exceptions are usually observed with borderline cases existing when the occurrence frequencies of a trinucleotide are similar in the 3 frames. They have been treated "by hand" (Arquès and Michel, 1996, 1997 for details).

The method proposed here, called frame permutated trinucleotide frequency (FPTF), generalizes and automates the statistical approach FTF and increases the detection sensibility of circular codes in genes. As 2 permutated trinucleotides cannot simultaneously belong to a circular code, the 60 trinucleotides (without AAA, CCC, GGG and TTT) are gathered in 20 classes of 3 trinucleotides invariant by permutation. The method FPTF considers both the preferential frame of a trinucleotide by comparing its occurrence frequencies in the 3 frames and the preferential permutated trinucleotide in a frame by comparing the occurrence frequencies of the 3 permutated trinucleotides in this frame. A statistical function, taking into account these 2 parameters, allows to associate each trinucleotide with a unique frame.

Otherwise, in contrast to the method FTF which has been used with gene populations, i.e. genes from different genomes, the method FPTF is applied here to the genomes, precisely to their complete sets of genes in both strands. So far, no circular code has been searched in the archaeal genomes, members of the third domain of life.

### 1.4. The archaeal genomes

Archaea have features that are either unique, typically prokaryotic or typically eukaryotic (e.g. Bernander, 2000; Woese, 2000; Forterre, 2001). They possess a prokaryotic mode of cellular organization, e.g. no nuclear envelope, circular DNA molecules organized similar 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.

Most of them are extremophiles, i.e. adapted to live in extreme environment such as high or low temperature, acid or salinity water, etc. Archaea are often viewed as predominant over prokarytotes in all hostile environments. Their variability in their morphology, metabolic pathway and genetic information makes the construction of their phylogenetic tree difficult.

The actual hypothesis divides the archaeal domain into 3 main phyla:

- The Euryarchaeota (EA) containing the methanogens (strict anaerobes and methane producers), the halophiles (strict aerobes living in high-salt environments), some thermoacidophiles (aerobes living in hot and acidic environments) and some hyperthermophiles (living in hot environments).
- The Crenarchaeota (CA) principally composed of hyperthermophiles.
- The Korarchaeota deduced from small subunit ribosomal DNA sequences and for which no organism has been sequenced yet.

Circular codes are searched in the 16 archaeal complete genomes sequenced at the time of writing this article which comprise 12 Euryarchaeota and 4 Crenarchaeota. Their complete sets of genes are analysed in both strands. The method FPTF identifies several circular codes in archaea. Several properties of these codes are described. In particular, a new definition of a parameter allows to measure some probabilities of words generated by the circular codes. Some biological consequences in archaea are presented in Discussion with respect to the 3 2-letter genetic alphabets (purine/pyrimidine, amino/ceto, strong/weak interaction) and the genetic code.

## 2. Method

The methods developed for identifying circular codes in genes consider the permutated trinucleotides. On the genetic alphabet {A,C,G,T}, there are 60 trinucleotides $w \in \mathscr{T} = \{\mathrm{AAA},\ldots,\mathrm{TTT}\} - \{\mathrm{AAA},\mathrm{CCC},\mathrm{GGG},\mathrm{TTT}\}$ which can be gathered in 20 sets of 3 permutated trinucleotides. In genes, a trinucleotide $w$ can be read in 3 frames $p \in \{0,1,2\}$ and then noted $w^p$ with $p = 0$: reading frame established by the start trinucleotide ATG, and $p = 1$ (resp. $p = 2$): reading frame shifted by 1 (resp. 2) nucleotide in the 5′–3′ direction. Therefore, there are $60 \times 3 = 180$ trinucleotides $w^p$.

The first method developed, called FTF (Arquès and Michel, 1996), computes the 180 occurrence frequencies $o(w^p)$ of the trinucleotides $w^p$ in genes and assigns a preferential frame phasepref to the trinucleotide $w$, noted $w^{\mathrm{phasepref}}$, such as the occurrence frequency of $w$ in this frame is the highest one among the 3 possible frames $p$

$$\mathrm{phase}(w) = k \ \text{ such as } \ o(w^k) = \underset{p=0}{\overset{2}{\mathrm{MAX}}}\{o(w^p)\},$$
$$w^{\mathrm{phasepref}} = w^{\mathrm{phase}(w)}. \qquad (1)$$

However, as the method FTF is based on the simple parameter $o(w^p)$, some trinucleotides are associated with 2 frames $p$ and $p'$, i.e. $\left|o(w^p) - o(w^{p'})\right| \leqslant \varepsilon$ where $\varepsilon = 0.2\%$ with the circular code of eukaryotes/prokaryotes (Arquès and Michel, 1996) and $\varepsilon = 0.4\%$ with the circular code of mitochondria (Arquès and Michel, 1997). Furthermore, in these borderline cases, a different preferential frame cannot be assigned to each permutated trinucleotide in order to verify the hypotheses of a potential $\mathscr{C}^3$ code. These exceptions with the circular codes of eukaryotes/prokaryotes and mitochondria have been treated "by hand" (Arquès and Michel, 1996, 1997).

In order to have a general and automatic approach for the frame assignment to a trinucleotide, we propose here a quantitative method which considers simultaneously the 3 permutated trinucleotides for assigning a different preferential frame to each permutated trinucleotide. This method called FPTF will identify several circular codes in archaeal genes.

As there are 3 permutated trinucleotides $w_i$, $i\in\{0,1,2\}$, in 3 different frames $p\in\{0,1,2\}$, there are 9 possible trinucleotides $w_i^p$ and therefore $\binom{9}{3}=84$ possible sets $\mathscr{S}^j$, $j\in\{1,\ldots,84\}$, of 3 trinucleotides. The 84 sets $\mathscr{S}^j$ are noted $\mathscr{S}84$ and defined as follows:

$$\begin{aligned}&\mathscr{S}84\\&=\big\{\{w_0^0,w_0^1,w_0^2\},\ldots,\{w_0^0,w_0^2,w_1^0\},\ldots,\{w_0^0,w_0^0,w_1^1\},\ldots,\\&\quad\{w_0^0,w_1^1,w_1^2\},\ldots,\{w_0^0,w_1^2,w_2^0\},\ldots,\{w_0^0,w_2^0,w_2^1\},\ldots,\\&\quad\{w_0^1,w_0^2,w_1^0\},\ldots,\{w_0^2,w_1^0,w_1^1\},\ldots,\{w_1^0,w_1^1,w_1^2\},\ldots,\\&\quad\{w_1^1,w_1^2,w_2^0\},\ldots,\{w_1^2,w_2^0,w_2^1\},\ldots,\{w_2^0,w_2^1,w_2^2\}\big\}\\&=\big\{\mathscr{S}^1,\ldots,\mathscr{S}^{84}\big\}.\end{aligned}$$

By convention, the trinucleotides $w_i^p$, $w_{i'}^{p'}$ and $w_{i''}^{p''}$, $i,i',i'',p,p',p''\in\{0,1,2\}$, in a set $\mathscr{S}^j=\left\{w_i^p,w_{i'}^{p'},w_{i''}^{p''}\right\}$ are in the lexicographical order.

Among these 84 sets $\mathscr{S}^j$, the 3 sets of left permutated trinucleotides are:

- the trinucleotide $w_0$ in frame 0, i.e. $w_0^0$, the permutated trinucleotide $w_1$ in frame 1, i.e. $w_1^1$, and the permutated trinucleotide $w_2$ in frame 2, i.e. $w_2^2$, leading to the set $\mathscr{S}^{22}=\{w_0^0,w_1^1,w_2^2\}$;
- the trinucleotide $w_0$ in frame 1, i.e. $w_0^1$, the permutated trinucleotide $w_1$ in frame 2, i.e. $w_1^2$, and the permutated trinucleotide $w_2$ in frame 0, i.e. $w_2^0$, leading to the set $\mathscr{S}^{44}=\{w_0^1,w_1^2,w_2^0\}$;
- the trinucleotide $w_0$ in frame 2, i.e. $w_0^2$, the permutated trinucleotide $w_1$ in frame 0, i.e. $w_1^0$, and the permutated trinucleotide $w_2$ in frame 1, i.e. $w_2^1$, leading to the set $\mathscr{S}^{53}=\{w_0^2,w_1^0,w_2^1\}$.

These 3 sets $\mathscr{S}^{22}$, $\mathscr{S}^{44}$ and $\mathscr{S}^{53}$ associate each trinucleotide with a preferential frame and each frame with a permutated trinucleotide by respecting the definition of the trinucleotide permutation (Section 1). Therefore, in these 3 sets, a relation between a trinucleotide and its frame, allows to deduce (by permutation) the 2 other relations between the permutated trinucleotides and their frames. A statistical analysis will show that these 3 sets occur with the highest value in all the archaeal genomes (see below the Remark concerning the sets $\mathscr{S}^j$).

A statistical function is developed for quantifying these 84 sets $\mathscr{S}^j$ and determining the preferential set among $\mathscr{S}^{22}$, $\mathscr{S}^{44}$ and $\mathscr{S}^{53}$. Let $o(w_i^p)$, $i,p\in\{0,1,2\}$, be the occurrence probability of a trinucleotide $w_i^p$ in genes. In order to weight the probability $o(w_i^p)$ of $w_i^p$ by its probabilities in the 3 frames, the following probability function $P$ allows to associate a preferential frame with a trinucleotide

$$P(w_i^p)=\frac{o(w_i^p)}{\sum_{p=0}^{2}o(w_i^p)}. \tag{2}$$

Similarly, in order to weight the probability $o(w_i^p)$ of $w_i^p$ by its permutated trinucleotide probabilities, the following probability function $Q$ allows to associate a preferential permutated trinucleotide with a frame

$$Q(w_i^p)=\frac{o(w_i^p)}{\sum_{i=0}^{2}o(w_i^p)}. \tag{3}$$

*Remark*: In a genome with hundreds of genes, the denominators $DEN(P)$ and $DEN(Q)$ are different from 0. Indeed, there is no stop codon $\tilde{w}_0\in\{\mathrm{TAA},\mathrm{TAG},\mathrm{TGA}\}$ in frame 0 in genes. Therefore, $o(\tilde{w}_0^0)=0$. As $\tilde{w}_0$ occurs in frames 1 and 2, then $DEN(P)=\sum_{p=0}^{2}o(\tilde{w}_0^p)=\sum_{p=1}^{2}o(\tilde{w}_0^p)>0$. On the other hand, as $\mathscr{P}(\tilde{w}_0)$ and $\mathscr{P}(\mathscr{P}(\tilde{w}_0))$ occur in frame 0, then $DEN(Q)=\sum_{i=0}^{2}o(\tilde{w}_i^0)=\sum_{i=1}^{2}o(\tilde{w}_i^0)>0$. These 2 inequalities could obviously not be verified with one gene of short length.

The statistical function for selecting the set $\mathscr{S}$ of 3 trinucleotides $\left\{w_{i_0}^{p_0},w_{i_1}^{p_1},w_{i_2}^{p_2}\right\}$, which has the strongest weight compared to the frame and the permutated trinucleotide, is defined as follows:

$$\begin{aligned}F(\mathscr{S}^{N(i_0,p_0,i_1,p_1,i_2,p_2)})&=F\left(w_{i_0}^{p_0},w_{i_1}^{p_1},w_{i_2}^{p_2}\right)\\&=\frac{1}{3}\sum_{j=0}^{2}\frac{1}{2}\left(P(w_{i_j}^{p_j})+Q(w_{i_j}^{p_j})\right)\\&=\frac{1}{6}\sum_{j=0}^{2}\left(P(w_{i_j}^{p_j})+Q(w_{i_j}^{p_j})\right)\end{aligned} \tag{4}$$

with

$$\begin{aligned}N(i_0,p_0,i_1,p_1,i_2,p_2)=&\frac{1}{6}I_0(I_0^2-24I_0+191)\\&-\frac{1}{2}I_1(I_1-17)+I_2-45\end{aligned}$$

and $I_0=3i_0+p_0+1$, $I_1=3i_1+p_1+1$, $I_2=3i_2+p_2+1$, $i_0,p_0,i_1,p_1,i_2,p_2\in\{0,1,2\}$ verifying $I_0<I_1<I_2$.

**Property 1.** If the 9 occurrence probabilities $o(w_i^p)$ in a set $\mathscr{S}^j$ are identical, i.e. $o(w_i^p)=o(w_{i'}^{p'})\ \forall i,i',p,p'\in\{0,1,2\}$, then $F(\mathscr{S}^j)=1/3\ \forall j\in\{1,\ldots,84\}$ (proof obvious).

The set $\mathscr{S}^{\max}$, having the highest value with the function $F$ among the 84 sets $\mathscr{S}^j$, is

$$\mathscr{S}^{\max}=\mathscr{S}^k\quad\text{such as } F(\mathscr{S}^k)=\underset{j=1}{\overset{84}{\mathrm{MAX}}}\{F(\mathscr{S}^j)\}. \tag{5}$$

*Remark concerning the sets $\mathscr{S}^j$*: Let $\mathscr{S}_k$, $k \in \{1, \ldots, 20\}$, be a set of 3 permutated trinucleotides in the lexicographical order among the 20 possible ones, i.e. $\mathscr{S}_1 = \{AAC, ACA, CAA\}, \ldots, \mathscr{S}_{20} = \{GTT, TTG, TGT\}$ The 9 trinucleotides $w_i^p$ of a given set $\mathscr{S}_k$ can be partitioned into

$$\frac{1}{3!}\binom{9}{3}\binom{6}{3} = 280$$

different sets of 3 sets of 3 trinucleotides. Such a partition is noted

$$\begin{aligned} E_k^{(j_0,j_1,j_2)} &= \left\{\left\{w_{i_0}^{p_0}, w_{i_1}^{p_1}, w_{i_2}^{p_2}\right\}, \left\{w_{i'_0}^{p'_0}, w_{i'_1}^{p'_1}, w_{i'_2}^{p'_2}\right\}, \right. \\ &\quad \left.\left\{w_{i''_0}^{p''_0}, w_{i''_1}^{p''_1}, w_{i''_2}^{p''_2}\right\}\right\} \\ &= \left\{\mathscr{S}_k^{j_0}, \mathscr{S}_k^{j_1}, \mathscr{S}_k^{j_2}\right\}, \end{aligned}$$

$k \in \{1, \ldots, 20\}$, $j_0, j_1, j_2 \in \{1, \ldots, 84\}$ and $j_0 \neq j_1 \neq j_2$, by associating each set of 3 trinucleotides with a set $\mathscr{S}_k^j$ such as $\mathscr{S}_k^{j_l} \bigcap \mathscr{S}_k^{j_{l'}} = \varnothing$, $l, l' \in \{0, 1, 2\}$ and $l \neq l'$. The particular partition studied here is

$$\begin{aligned} E_k^{(22,44,53)} &= \left\{\mathscr{S}_k^{22}, \mathscr{S}_k^{44}, \mathscr{S}_k^{53}\right\} \\ &= \left\{\left\{w_0^0, w_1^1, w_2^2\right\}, \left\{w_0^1, w_1^2, w_2^0\right\}, \left\{w_0^2, w_1^0, w_2^1\right\}\right\}. \end{aligned}$$

For a given set $\mathscr{S}_k$, the partition $E_k^{(j_0,j_1,j_2)}$ is evaluated by the function $F$ as follows:

$$F\left(E_k^{(j_0,j_1,j_2)}\right) = \underset{l=0}{\overset{2}{\mathrm{MAX}}}\left\{F\left(\mathscr{S}_k^{j_l}\right)\right\} \tag{6}$$

and the partition $E^{(j_0,j_1,j_2)}$ for a genome $\mathscr{G}$ with 20 sets $\mathscr{S}_k$, each corresponding to a trinucleotide permutation class, by the mean $M$

$$M\left(E^{(j_0,j_1,j_2)}, \mathscr{G}\right) = \frac{1}{20}\sum_{k=1}^{20} F\left(E_k^{(j_0,j_1,j_2)}\right). \tag{7}$$

The computation of $M\left(E^{(j_0,j_1,j_2)}, \mathscr{G}\right)$ in each genome $\mathscr{G}$ shows that $M\left(E^{(22,44,53)}, \mathscr{G}\right)$ has the highest value among the 280 partitions $E^{(j_0,j_1,j_2)}$ in the 16 archaeal genomes (data not shown). This strong statistical result for the partition $E^{(22,44,53)}$ explains the choice of the set $\mathscr{S}^j = \left\{w_0^p, w_1^{p'}, w_2^{p''}\right\}$, $p, p', p'' \in \{0, 1, 2\}$ and $p \neq p' \neq p''$, for determining the set $\mathscr{S}^{max}$. Finally, the 3 sets $\mathscr{S}^j$ in the partition $E^{(22,44,53)}$ associate each permutated trinucleotide with a different frame such as the frame of a trinucleotide can be deduced from the frame of a permutated trinucleotide.

Finally, the preferential set $\mathscr{S}^{\mathrm{pref}}$ among $\mathscr{S}^{22} = \left\{w_0^0, w_1^1, w_2^2\right\}$, $\mathscr{S}^{44} = \left\{w_0^1, w_1^2, w_2^0\right\}$ and $\mathscr{S}^{53} = \left\{w_0^2, w_1^0, w_2^1\right\}$ in the partition $E^{(22,44,53)}$, is identified by comparing the sets $\mathscr{S}^{22}$, $\mathscr{S}^{44}$ and $\mathscr{S}^{53}$ with the set $\mathscr{S}^{\max}$

$$\mathscr{S}^{\mathrm{pref}} = \mathscr{S}^k$$

such as

$$F(\mathscr{S}^{\max}) - F(\mathscr{S}^k) = \underset{j=22,44,53}{\mathrm{MIN}}\left\{F(\mathscr{S}^{\max}) - F(\mathscr{S}^j)\right\}. \tag{8}$$

$\mathscr{S}^{\mathrm{pref}}$ is the set with the highest value among the sets $\mathscr{S}^{22}$, $\mathscr{S}^{44}$ and $\mathscr{S}^{53}$. In most cases (see below), $\mathscr{S}^{\max} = \mathscr{S}^{\mathrm{pref}}$ (i.e. $F(\mathscr{S}^{\max}) - F(\mathscr{S}^{\mathrm{pref}}) = 0$).

Sixteen archaeal genomes $\mathscr{G}$ are studied:

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

In all these genomes, the genes beginning obligatory with a start codon are extracted from both strands.

These large populations allow to have stable frequencies leading to significant statistical results.

## 3. Results

### 3.1. Identification of 3 subsets of 20 trinucleotides in the 3 frames of the genes in the archaeal genomes

The 180 occurrence frequencies $o(w_i^p)$ of the 20 sets of 3 permutated trinucleotides $w_i$ in the 3 frames $p$ are computed in all the genes of the 16 archaeal genomes (Table 2).

*Remark*: The frequencies of the 3 stop codons TAA, TAG, and TGA in frame 0 are obviously equal to 0 in all the genomes (Table 2).

Table 2
Occurrence frequencies $o(w_i^p)$ (in % and rounded at 0.1%) of the 20 sets of 3 permutated trinucleotides $w_i$ in the 3 frames $p$ of all the genes in the 16 archaeal genomes

| | *Aeropyrum* (AP) | | | *Archaeoglobus* (AG) | | | *Halobacterium* (HB) | | | *Methanococcus* (MC) | | | *Methanopyrus* (MP) | | | *Methanosarcina acetivorans* (MSA) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| AAC | 1.6 | 0.5 | 1.0 | 2.2 | 0.5 | 1.4 | 2.0 | 1.0 | 0.5 | 1.6 | 1.0 | 2.2 | 1.7 | 1.1 | 0.6 | 2.2 | 1.6 | 1.6 |
| ACA | 1.1 | 2.2 | 0.5 | 1.0 | 2.9 | 0.3 | 0.4 | 2.5 | 0.7 | 2.0 | 1.6 | 0.9 | 0.4 | 1.6 | 0.6 | 1.6 | 2.2 | 0.9 |
| CAA | 0.3 | 1.0 | 1.7 | 0.3 | 2.0 | 2.6 | 0.4 | 0.3 | 1.7 | 0.9 | 3.6 | 1.4 | 0.3 | 0.2 | 1.9 | 0.7 | 2.4 | 2.3 |
| AAG | 2.8 | 1.0 | 1.1 | 4.5 | 1.8 | 1.8 | 1.1 | 0.9 | 0.1 | 3.1 | 4.6 | 2.3 | 3.1 | 1.8 | 0.3 | 2.3 | 3.6 | 1.6 |
| AGA | 1.1 | 2.6 | 2.0 | 2.2 | 3.6 | 2.4 | 0.1 | 1.7 | 0.6 | 2.8 | 2.9 | 6.1 | 0.6 | 2.5 | 1.6 | 1.3 | 2.2 | 3.3 |
| GAA | 1.2 | 0.6 | 2.0 | 2.8 | 2.1 | 3.5 | 1.6 | 0.2 | 1.8 | 5.1 | 2.5 | 2.0 | 2.6 | 0.6 | 2.9 | 5.5 | 1.8 | 2.5 |
| AAT | 0.5 | 0.3 | 0.9 | 1.0 | 0.7 | 2.8 | 0.1 | 0.3 | 0.2 | 3.7 | 2.3 | 5.0 | 0.3 | 0.3 | 0.6 | 2.3 | 1.4 | 2.7 |
| ATA | 3.2 | 1.2 | 0.5 | 2.3 | 0.7 | 0.5 | 0.1 | 0.1 | 0.1 | 4.6 | 4.4 | 1.5 | 1.0 | 0.5 | 0.2 | 2.1 | 2.4 | 0.9 |
| TAA | 0.0 | 1.6 | 1.1 | 0.0 | 2.2 | 1.2 | 0.0 | 0.1 | 0.1 | 0.0 | 5.5 | 4.8 | 0.0 | 0.5 | 0.8 | 0.0 | 2.1 | 2.0 |
| ACC | 1.6 | 2.0 | 0.9 | 1.2 | 1.4 | 0.5 | 3.2 | 3.2 | 0.4 | 0.4 | 0.5 | 0.8 | 1.9 | 2.4 | 1.2 | 1.7 | 1.7 | 0.7 |
| CCA | 0.9 | 2.4 | 1.8 | 0.8 | 1.5 | 0.9 | 0.4 | 2.1 | 2.4 | 2.3 | 0.5 | 0.5 | 0.5 | 1.5 | 1.1 | 0.6 | 1.5 | 1.2 |
| CAC | 1.3 | 0.8 | 1.7 | 1.1 | 0.5 | 1.3 | 2.1 | 0.8 | 3.6 | 0.4 | 0.5 | 0.4 | 1.5 | 0.9 | 1.8 | 0.8 | 0.6 | 1.2 |
| ACG | 1.1 | 2.4 | 0.3 | 1.2 | 2.9 | 0.1 | 3.0 | 7.6 | 0.5 | 0.2 | 0.7 | 0.0 | 2.0 | 4.7 | 1.0 | 0.8 | 1.9 | 0.3 |
| CGA | 0.3 | 1.3 | 3.0 | 0.1 | 1.7 | 3.1 | 0.5 | 2.7 | 10.1 | 0.0 | 0.2 | 0.5 | 1.1 | 2.1 | 6.4 | 0.3 | 0.9 | 2.6 |
| GAC | 2.6 | 0.4 | 1.3 | 2.7 | 0.3 | 0.9 | 8.1 | 0.4 | 2.5 | 1.0 | 0.3 | 0.4 | 4.2 | 1.0 | 1.8 | 2.4 | 0.4 | 1.2 |
| ACT | 0.9 | 1.1 | 1.1 | 0.8 | 1.7 | 0.8 | 0.3 | 1.3 | 0.7 | 1.5 | 1.2 | 0.8 | 0.4 | 1.3 | 1.5 | 1.4 | 1.2 | 1.4 |
| CTA | 1.8 | 1.9 | 1.6 | 0.5 | 0.5 | 1.4 | 0.1 | 0.1 | 1.2 | 0.9 | 1.5 | 0.8 | 0.9 | 0.5 | 0.9 | 0.5 | 1.0 | 1.0 |
| TAC | 2.1 | 1.2 | 0.8 | 2.8 | 0.4 | 0.6 | 2.3 | 0.2 | 0.2 | 1.0 | 0.6 | 1.0 | 2.5 | 1.3 | 0.6 | 1.7 | 0.7 | 1.5 |
| AGC | 2.6 | 2.1 | 2.1 | 1.9 | 2.3 | 1.6 | 1.6 | 2.3 | 0.7 | 0.5 | 0.8 | 2.2 | 0.8 | 2.7 | 1.1 | 1.2 | 1.3 | 1.9 |
| GCA | 1.4 | 1.6 | 1.0 | 2.1 | 1.6 | 1.3 | 0.8 | 1.9 | 1.5 | 2.3 | 0.5 | 0.7 | 0.6 | 0.9 | 1.0 | 2.6 | 1.4 | 1.1 |
| CAG | 1.6 | 1.7 | 2.8 | 1.5 | 1.8 | 2.4 | 2.2 | 0.6 | 1.2 | 0.5 | 2.3 | 0.4 | 1.1 | 0.6 | 1.2 | 1.9 | 2.3 | 1.9 |
| AGG | 4.6 | 4.1 | 1.7 | 3.0 | 4.7 | 1.3 | 0.1 | 3.1 | 0.2 | 1.0 | 2.2 | 2.1 | 1.7 | 4.8 | 1.2 | 1.4 | 2.2 | 2.2 |
| GGA | 1.2 | 2.4 | 3.8 | 2.5 | 1.8 | 4.4 | 0.5 | 1.1 | 4.2 | 3.6 | 1.0 | 2.2 | 1.8 | 2.0 | 6.4 | 2.7 | 1.6 | 3.4 |
| GAG | 5.5 | 1.3 | 3.3 | 6.1 | 1.9 | 2.9 | 5.1 | 0.3 | 0.7 | 3.5 | 2.5 | 1.2 | 7.4 | 1.3 | 1.5 | 2.5 | 1.6 | 1.0 |
| AGT | 0.7 | 1.0 | 1.5 | 0.5 | 1.5 | 1.6 | 0.3 | 1.4 | 0.4 | 1.1 | 1.2 | 2.6 | 0.4 | 1.6 | 1.3 | 1.0 | 1.0 | 1.6 |
| GTA | 1.5 | 0.9 | 0.9 | 1.1 | 0.4 | 0.8 | 0.2 | 0.1 | 1.1 | 1.5 | 1.0 | 0.5 | 1.5 | 1.0 | 1.4 | 1.8 | 0.9 | 0.7 |
| TAG | 0.0 | 3.2 | 1.8 | 0.0 | 1.3 | 0.5 | 0.0 | 0.1 | 0.1 | 0.0 | 3.7 | 1.6 | 0.0 | 1.6 | 0.4 | 0.0 | 1.6 | 0.5 |
| ATC | 1.3 | 0.8 | 0.6 | 2.1 | 0.9 | 0.5 | 3.3 | 0.7 | 0.3 | 1.1 | 1.0 | 1.1 | 3.4 | 1.1 | 0.5 | 2.5 | 1.3 | 0.9 |
| TCA | 0.7 | 2.4 | 0.6 | 0.9 | 3.0 | 0.8 | 0.2 | 2.6 | 0.3 | 1.5 | 0.8 | 0.9 | 0.4 | 2.6 | 0.7 | 1.2 | 2.4 | 1.1 |
| CAT | 0.6 | 0.5 | 2.2 | 0.4 | 0.6 | 2.8 | 0.2 | 0.1 | 2.5 | 1.0 | 1.5 | 1.2 | 0.4 | 0.2 | 1.9 | 0.9 | 0.8 | 2.1 |
| ATG | 2.0 | 1.3 | 0.3 | 2.5 | 2.1 | 0.4 | 1.7 | 0.4 | 0.1 | 2.2 | 4.7 | 0.7 | 1.7 | 0.7 | 0.3 | 2.3 | 3.0 | 0.5 |
| TGA | 0.0 | 2.1 | 1.8 | 0.0 | 3.2 | 3.9 | 0.0 | 1.7 | 0.9 | 0.0 | 1.7 | 5.5 | 0.0 | 2.9 | 1.4 | 0.0 | 2.3 | 4.1 |
| GAT | 1.3 | 0.3 | 2.0 | 2.2 | 0.6 | 2.9 | 0.9 | 0.2 | 2.2 | 4.5 | 1.1 | 2.2 | 1.6 | 0.5 | 3.3 | 2.9 | 0.5 | 2.3 |
| ATT | 0.8 | 0.3 | 0.5 | 2.8 | 0.8 | 0.9 | 0.2 | 0.2 | 0.1 | 4.9 | 2.6 | 4.1 | 0.5 | 0.4 | 0.4 | 2.7 | 1.6 | 1.3 |
| TTA | 0.5 | 1.6 | 0.4 | 0.6 | 1.8 | 1.0 | 0.1 | 0.1 | 0.1 | 5.3 | 4.5 | 1.7 | 0.5 | 0.5 | 0.4 | 0.9 | 2.3 | 1.2 |
| TAT | 1.3 | 0.9 | 1.9 | 0.8 | 0.5 | 1.0 | 0.2 | 0.1 | 0.1 | 3.4 | 2.3 | 4.1 | 0.3 | 0.4 | 0.7 | 2.1 | 0.9 | 2.4 |
| CCG | 1.3 | 2.5 | 0.9 | 1.1 | 1.7 | 0.2 | 2.4 | 7.3 | 2.9 | 0.1 | 0.2 | 0.0 | 2.6 | 4.4 | 2.2 | 1.0 | 2.0 | 0.8 |
| CGC | 0.7 | 1.2 | 2.3 | 0.2 | 0.8 | 1.5 | 3.3 | 3.3 | 7.6 | 0.0 | 0.0 | 0.3 | 1.6 | 1.7 | 3.3 | 0.5 | 0.4 | 0.9 |
| GCC | 3.7 | 2.1 | 2.2 | 2.0 | 0.6 | 1.2 | 5.8 | 2.1 | 2.0 | 0.6 | 0.2 | 0.5 | 3.4 | 1.0 | 1.4 | 1.8 | 0.8 | 0.8 |
| CCT | 1.8 | 1.9 | 4.2 | 0.8 | 1.3 | 2.0 | 0.1 | 0.8 | 2.6 | 0.9 | 0.5 | 0.3 | 0.6 | 1.0 | 2.9 | 1.4 | 1.1 | 2.6 |
| CTC | 3.9 | 1.4 | 2.1 | 2.5 | 1.0 | 1.2 | 4.0 | 0.2 | 1.3 | 0.3 | 0.8 | 0.6 | 3.5 | 0.7 | 1.4 | 1.8 | 1.3 | 1.2 |
| TCC | 1.6 | 2.5 | 1.1 | 0.9 | 1.1 | 1.3 | 1.2 | 2.7 | 0.5 | 0.3 | 0.5 | 1.4 | 1.4 | 3.0 | 1.1 | 1.5 | 1.8 | 1.4 |
| CGG | 0.7 | 2.2 | 2.3 | 0.1 | 1.8 | 2.1 | 2.1 | 5.2 | 5.5 | 0.0 | 0.2 | 0.3 | 2.2 | 4.1 | 3.5 | 0.6 | 1.6 | 2.3 |
| GGC | 3.3 | 2.2 | 3.7 | 1.8 | 1.1 | 2.3 | 5.0 | 1.6 | 4.1 | 0.4 | 0.3 | 1.1 | 1.9 | 1.2 | 3.3 | 1.5 | 1.0 | 1.5 |
| GCG | 2.0 | 2.1 | 0.6 | 1.7 | 1.4 | 0.2 | 6.0 | 5.5 | 2.5 | 0.2 | 0.2 | 0.0 | 3.4 | 2.6 | 1.9 | 0.7 | 1.0 | 0.4 |
| CGT | 0.4 | 0.8 | 2.2 | 0.1 | 0.4 | 2.2 | 0.4 | 2.1 | 5.6 | 0.0 | 0.2 | 0.4 | 1.1 | 1.4 | 4.7 | 0.4 | 0.3 | 1.2 |
| GTC | 2.3 | 0.5 | 1.0 | 1.5 | 0.3 | 0.5 | 5.0 | 0.4 | 1.6 | 0.5 | 0.2 | 0.2 | 3.8 | 1.1 | 1.1 | 1.6 | 0.5 | 0.6 |
| TCG | 1.1 | 2.8 | 0.3 | 0.7 | 3.1 | 0.1 | 1.9 | 8.5 | 0.5 | 0.1 | 0.4 | 0.0 | 1.4 | 6.3 | 1.0 | 0.7 | 2.1 | 0.3 |
| CTG | 2.8 | 2.2 | 1.0 | 2.6 | 2.1 | 0.9 | 3.2 | 0.7 | 0.9 | 0.2 | 2.2 | 0.4 | 3.5 | 0.6 | 0.9 | 2.4 | 2.3 | 1.1 |
| TGC | 0.6 | 1.6 | 1.4 | 0.9 | 1.7 | 2.5 | 0.4 | 2.6 | 0.8 | 0.4 | 0.6 | 2.0 | 0.9 | 1.9 | 0.6 | 0.7 | 1.5 | 2.7 |
| GCT | 2.5 | 1.3 | 3.3 | 2.1 | 1.1 | 3.2 | 0.5 | 0.9 | 3.8 | 2.5 | 0.5 | 0.6 | 0.9 | 0.7 | 3.3 | 1.8 | 0.8 | 1.9 |
| CTT | 1.7 | 0.6 | 1.6 | 2.4 | 0.7 | 2.2 | 0.3 | 0.1 | 1.0 | 0.9 | 1.3 | 1.5 | 0.6 | 0.4 | 1.4 | 3.0 | 1.3 | 1.4 |
| TTC | 2.1 | 1.0 | 0.5 | 2.8 | 2.0 | 0.9 | 2.9 | 0.5 | 0.2 | 0.9 | 1.1 | 1.0 | 2.6 | 0.9 | 0.5 | 2.0 | 1.7 | 2.0 |
| TCT | 0.9 | 1.9 | 1.7 | 0.6 | 1.6 | 1.9 | 0.1 | 1.5 | 0.5 | 1.1 | 1.1 | 0.7 | 0.3 | 1.5 | 0.9 | 1.2 | 1.6 | 1.9 |
| GGT | 1.5 | 1.1 | 2.8 | 1.3 | 0.7 | 2.6 | 0.7 | 1.0 | 3.1 | 1.3 | 0.4 | 1.1 | 2.2 | 1.3 | 3.8 | 1.2 | 0.7 | 1.7 |
| GTG | 2.9 | 1.2 | 0.7 | 2.2 | 1.1 | 0.7 | 3.8 | 1.1 | 0.8 | 0.6 | 1.2 | 0.3 | 4.3 | 1.7 | 1.1 | 1.4 | 1.1 | 0.3 |
| TGG | 1.3 | 3.7 | 1.4 | 1.0 | 2.8 | 1.7 | 1.1 | 3.5 | 0.6 | 0.7 | 1.8 | 3.2 | 1.2 | 4.5 | 0.7 | 1.1 | 2.4 | 1.4 |
| GTT | 2.1 | 0.3 | 1.3 | 3.9 | 0.4 | 1.4 | 0.7 | 0.1 | 2.6 | 4.4 | 0.9 | 1.7 | 1.0 | 0.4 | 2.3 | 2.1 | 0.8 | 1.1 |
| TTG | 0.7 | 2.1 | 0.3 | 1.1 | 5.1 | 0.3 | 0.7 | 0.7 | 0.1 | 1.9 | 5.3 | 0.7 | 1.1 | 0.5 | 0.3 | 0.8 | 4.2 | 0.3 |
| TGT | 0.3 | 1.1 | 2.1 | 0.3 | 0.7 | 2.2 | 0.3 | 1.6 | 0.5 | 0.9 | 0.8 | 2.8 | 0.5 | 1.4 | 0.8 | 0.6 | 0.8 | 2.3 |

Table 2 (*continued*)

| | *Methanosarcina mazei* (MSM) | | | *Methanothermobacter* (MT) | | | *Pyrobaculum* (PB) | | | *Pyrococcus abyssi* (PCA) | | | *Pyrococcus furiosus* (PCF) | | | *Pyrococcus horikoshii* (PCH) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| AAC | 2.0 | 1.5 | 1.5 | 2.2 | 1.5 | 1.4 | 1.6 | 0.8 | 0.9 | 2.0 | 0.8 | 1.8 | 1.7 | 1.0 | 1.9 | 1.7 | 1.0 | 2.0 |
| ACA | 1.7 | 2.1 | 0.9 | 2.1 | 2.9 | 0.8 | 1.1 | 2.0 | 0.9 | 0.8 | 2.3 | 0.4 | 1.6 | 2.2 | 0.6 | 1.1 | 1.7 | 0.5 |
| CAA | 0.5 | 2.5 | 2.0 | 0.1 | 2.2 | 1.9 | 1.0 | 1.4 | 1.9 | 0.7 | 2.2 | 2.2 | 1.0 | 3.1 | 2.0 | 0.8 | 2.6 | 1.7 |
| AAG | 2.2 | 3.9 | 1.7 | 2.8 | 0.9 | 1.9 | 2.8 | 2.0 | 1.2 | 5.4 | 2.9 | 2.2 | 4.3 | 4.0 | 2.5 | 4.6 | 3.2 | 2.4 |
| AGA | 1.4 | 2.2 | 3.8 | 1.2 | 3.3 | 2.3 | 1.9 | 2.3 | 3.2 | 1.7 | 4.0 | 3.5 | 2.8 | 3.2 | 4.8 | 2.0 | 3.4 | 3.6 |
| GAA | 5.7 | 1.8 | 2.4 | 3.2 | 1.2 | 2.6 | 2.0 | 0.9 | 1.9 | 3.6 | 2.1 | 3.5 | 4.9 | 2.7 | 2.6 | 4.1 | 2.2 | 3.0 |
| AAT | 2.4 | 1.4 | 3.0 | 1.1 | 1.0 | 2.1 | 1.0 | 1.0 | 2.6 | 1.3 | 0.9 | 3.4 | 1.8 | 1.2 | 4.4 | 1.9 | 1.1 | 3.7 |
| ATA | 2.4 | 2.5 | 0.9 | 4.3 | 1.2 | 0.7 | 2.9 | 1.6 | 1.1 | 4.9 | 1.5 | 0.7 | 4.1 | 1.8 | 1.1 | 4.5 | 2.2 | 0.9 |
| TAA | 0.0 | 2.4 | 2.3 | 0.0 | 2.3 | 1.4 | 0.0 | 2.7 | 2.0 | 0.0 | 4.0 | 2.4 | 0.0 | 4.2 | 2.3 | 0.0 | 4.3 | 2.9 |
| ACC | 1.4 | 1.7 | 0.7 | 1.9 | 2.4 | 1.3 | 0.8 | 1.2 | 0.4 | 1.1 | 1.1 | 0.8 | 0.8 | 0.9 | 0.8 | 1.1 | 0.9 | 0.9 |
| CCA | 0.7 | 1.4 | 1.2 | 1.1 | 2.3 | 1.7 | 0.8 | 1.8 | 1.1 | 1.7 | 1.3 | 0.9 | 2.0 | 1.1 | 0.9 | 1.8 | 1.2 | 0.8 |
| CAC | 0.8 | 0.7 | 1.1 | 1.1 | 1.3 | 1.8 | 1.1 | 0.3 | 1.5 | 0.9 | 0.5 | 0.8 | 0.8 | 0.7 | 0.8 | 0.7 | 0.6 | 0.7 |
| ACG | 0.6 | 1.8 | 0.3 | 0.5 | 1.7 | 0.3 | 1.2 | 3.6 | 0.3 | 1.2 | 2.4 | 0.1 | 0.6 | 1.5 | 0.1 | 1.0 | 1.5 | 0.1 |
| CGA | 0.2 | 0.7 | 2.1 | 0.1 | 0.4 | 1.8 | 0.2 | 1.0 | 2.7 | 0.1 | 1.4 | 2.6 | 0.1 | 0.6 | 1.5 | 0.1 | 1.1 | 1.7 |
| GAC | 2.4 | 0.4 | 1.1 | 2.9 | 0.7 | 1.2 | 2.8 | 0.4 | 1.2 | 1.9 | 0.4 | 0.9 | 1.6 | 0.5 | 0.8 | 1.2 | 0.4 | 0.8 |
| ACT | 1.4 | 1.1 | 1.4 | 0.5 | 1.4 | 2.2 | 1.4 | 1.4 | 1.0 | 1.1 | 1.5 | 1.4 | 1.5 | 1.4 | 1.5 | 1.4 | 1.4 | 1.5 |
| CTA | 0.4 | 1.0 | 0.8 | 0.5 | 0.2 | 1.4 | 1.3 | 1.4 | 1.2 | 1.9 | 1.4 | 1.1 | 1.8 | 1.3 | 1.0 | 1.8 | 1.7 | 0.9 |
| TAC | 1.5 | 0.7 | 1.5 | 2.3 | 1.2 | 0.5 | 2.6 | 1.0 | 0.8 | 2.4 | 0.9 | 0.7 | 1.8 | 0.8 | 1.0 | 1.9 | 1.0 | 1.0 |
| AGC | 1.3 | 1.3 | 2.0 | 1.1 | 1.1 | 1.4 | 1.2 | 1.7 | 1.5 | 1.4 | 1.9 | 2.0 | 1.1 | 1.5 | 2.3 | 1.2 | 1.6 | 2.3 |
| GCA | 2.9 | 1.4 | 1.1 | 2.9 | 1.4 | 0.7 | 1.5 | 1.4 | 1.0 | 1.7 | 0.8 | 0.8 | 2.5 | 0.7 | 0.7 | 1.9 | 0.6 | 0.7 |
| CAG | 2.0 | 2.7 | 2.1 | 1.8 | 3.4 | 2.9 | 1.1 | 1.8 | 1.5 | 0.9 | 1.6 | 1.2 | 0.8 | 2.3 | 1.0 | 0.8 | 1.7 | 0.9 |
| AGG | 1.7 | 2.1 | 2.5 | 4.1 | 4.3 | 1.8 | 2.7 | 3.5 | 1.4 | 3.7 | 4.2 | 1.8 | 2.1 | 3.0 | 2.2 | 3.0 | 3.4 | 1.9 |
| GGA | 2.8 | 1.6 | 3.1 | 2.1 | 2.4 | 4.6 | 1.3 | 1.7 | 3.3 | 3.0 | 2.3 | 4.0 | 3.7 | 1.7 | 2.9 | 3.2 | 2.1 | 3.5 |
| GAG | 2.4 | 1.7 | 1.0 | 5.0 | 0.8 | 2.2 | 5.0 | 1.6 | 2.4 | 5.3 | 1.6 | 2.5 | 4.0 | 2.5 | 1.9 | 4.2 | 1.8 | 2.1 |
| AGT | 0.9 | 0.9 | 1.7 | 0.8 | 1.0 | 1.1 | 0.4 | 1.4 | 2.5 | 0.7 | 1.5 | 2.5 | 1.0 | 1.3 | 3.1 | 1.0 | 1.3 | 2.5 |
| GTA | 1.9 | 0.8 | 0.6 | 1.0 | 1.0 | 0.6 | 2.1 | 0.6 | 1.1 | 1.4 | 0.7 | 1.0 | 1.9 | 0.7 | 0.8 | 1.8 | 0.9 | 0.8 |
| TAG | 0.0 | 1.7 | 0.4 | 0.0 | 1.6 | 0.3 | 0.0 | 3.2 | 1.1 | 0.0 | 3.7 | 1.7 | 0.0 | 3.6 | 1.7 | 0.0 | 3.6 | 1.9 |
| ATC | 2.2 | 1.2 | 0.9 | 2.2 | 0.7 | 1.5 | 1.1 | 0.8 | 0.6 | 1.5 | 1.1 | 0.6 | 1.4 | 1.3 | 0.6 | 1.7 | 1.4 | 0.8 |
| TCA | 1.4 | 2.3 | 1.1 | 2.1 | 4.1 | 0.6 | 0.7 | 1.6 | 0.6 | 0.9 | 2.2 | 1.0 | 1.0 | 1.9 | 1.1 | 1.2 | 1.9 | 1.1 |
| CAT | 0.9 | 0.8 | 2.0 | 0.8 | 1.3 | 4.1 | 0.4 | 0.5 | 1.9 | 0.6 | 0.8 | 2.3 | 0.7 | 1.0 | 2.0 | 0.8 | 1.0 | 2.0 |
| ATG | 2.4 | 3.0 | 0.5 | 2.9 | 3.4 | 0.6 | 1.8 | 1.2 | 0.3 | 2.3 | 2.4 | 0.4 | 2.2 | 3.1 | 0.5 | 2.3 | 3.0 | 0.5 |
| TGA | 0.0 | 2.2 | 4.5 | 0.0 | 2.8 | 5.4 | 0.0 | 2.1 | 2.1 | 0.0 | 2.3 | 3.5 | 0.0 | 2.0 | 4.1 | 0.0 | 2.1 | 3.9 |
| GAT | 2.9 | 0.5 | 2.2 | 3.1 | 0.7 | 2.9 | 1.5 | 0.5 | 1.5 | 2.8 | 0.7 | 3.2 | 2.8 | 0.9 | 2.2 | 3.1 | 0.8 | 2.8 |
| ATT | 3.0 | 1.5 | 1.3 | 1.2 | 0.6 | 1.3 | 2.3 | 1.1 | 1.3 | 2.1 | 1.0 | 1.7 | 3.3 | 1.2 | 2.2 | 2.6 | 1.3 | 2.1 |
| TTA | 0.9 | 2.6 | 1.3 | 0.4 | 1.3 | 0.5 | 2.1 | 2.4 | 0.9 | 1.7 | 2.9 | 1.1 | 2.0 | 3.1 | 1.1 | 2.2 | 3.4 | 1.2 |
| TAT | 2.0 | 0.9 | 2.7 | 1.0 | 1.1 | 1.4 | 1.7 | 1.4 | 1.9 | 1.4 | 1.1 | 1.6 | 2.2 | 1.2 | 2.0 | 2.0 | 1.4 | 2.3 |
| CCG | 0.8 | 1.6 | 0.7 | 0.6 | 1.8 | 0.7 | 1.4 | 2.7 | 0.6 | 0.7 | 1.3 | 0.1 | 0.3 | 0.8 | 0.1 | 0.5 | 1.2 | 0.1 |
| CGC | 0.5 | 0.4 | 0.9 | 0.4 | 0.2 | 1.2 | 0.9 | 1.9 | 3.1 | 0.1 | 0.4 | 1.0 | 0.1 | 0.2 | 0.6 | 0.1 | 0.3 | 0.7 |
| GCC | 1.6 | 0.8 | 0.8 | 2.7 | 0.9 | 0.8 | 3.9 | 1.2 | 2.3 | 1.9 | 0.5 | 1.0 | 1.4 | 0.4 | 0.8 | 1.6 | 0.4 | 0.7 |
| CCT | 1.7 | 0.9 | 2.5 | 0.9 | 1.4 | 4.1 | 1.0 | 1.4 | 1.8 | 0.9 | 1.0 | 2.3 | 1.1 | 0.9 | 1.7 | 1.2 | 1.2 | 2.1 |
| CTC | 1.7 | 1.4 | 1.1 | 3.5 | 0.3 | 1.3 | 2.0 | 0.9 | 1.3 | 2.1 | 0.9 | 1.0 | 1.7 | 1.3 | 0.9 | 1.8 | 1.1 | 1.2 |
| TCC | 1.3 | 1.8 | 1.4 | 1.3 | 2.6 | 0.7 | 0.8 | 1.3 | 0.9 | 0.7 | 1.5 | 1.6 | 0.6 | 1.3 | 1.8 | 1.0 | 1.5 | 1.9 |
| CGG | 0.4 | 1.3 | 2.1 | 0.5 | 1.3 | 1.9 | 0.6 | 2.3 | 2.5 | 0.1 | 1.1 | 1.2 | 0.1 | 0.6 | 0.7 | 0.1 | 1.0 | 0.9 |
| GGC | 1.5 | 1.0 | 1.4 | 1.7 | 0.9 | 2.4 | 3.7 | 1.8 | 4.2 | 1.1 | 0.9 | 2.0 | 0.8 | 0.7 | 1.6 | 0.7 | 0.8 | 1.5 |
| GCG | 0.6 | 1.0 | 0.4 | 0.7 | 0.9 | 0.3 | 2.9 | 2.7 | 0.9 | 1.0 | 0.9 | 0.1 | 0.6 | 0.5 | 0.1 | 0.7 | 0.6 | 0.1 |
| CGT | 0.4 | 0.3 | 1.0 | 0.5 | 0.1 | 1.4 | 0.3 | 1.2 | 2.9 | 0.1 | 0.5 | 2.2 | 0.1 | 0.3 | 1.2 | 0.1 | 0.4 | 1.6 |
| GTC | 1.5 | 0.5 | 0.6 | 2.1 | 0.6 | 0.6 | 1.8 | 0.3 | 0.9 | 1.5 | 0.4 | 0.3 | 1.1 | 0.4 | 0.3 | 1.1 | 0.4 | 0.3 |
| TCG | 0.5 | 1.8 | 0.3 | 0.3 | 2.0 | 0.1 | 0.8 | 2.3 | 0.2 | 0.5 | 2.4 | 0.1 | 0.3 | 1.3 | 0.1 | 0.5 | 1.7 | 0.1 |
| CTG | 2.3 | 2.7 | 1.1 | 2.4 | 2.3 | 1.0 | 1.9 | 1.8 | 0.7 | 1.1 | 1.6 | 0.5 | 0.8 | 2.0 | 0.5 | 0.9 | 1.7 | 0.5 |
| TGC | 0.7 | 1.4 | 2.9 | 0.8 | 1.1 | 2.3 | 0.5 | 1.5 | 1.2 | 0.3 | 0.9 | 1.6 | 0.2 | 0.8 | 2.1 | 0.3 | 0.8 | 1.8 |
| GCT | 2.0 | 0.8 | 1.9 | 1.1 | 0.7 | 1.4 | 1.6 | 1.0 | 2.8 | 2.0 | 0.7 | 2.2 | 2.1 | 0.6 | 1.7 | 2.2 | 0.7 | 1.8 |
| CTT | 3.5 | 1.4 | 1.2 | 2.5 | 0.2 | 1.6 | 1.5 | 0.9 | 2.1 | 2.1 | 0.9 | 2.4 | 2.4 | 1.1 | 2.1 | 2.6 | 1.2 | 2.4 |
| TTC | 1.8 | 1.8 | 2.1 | 2.6 | 0.9 | 0.8 | 1.7 | 0.9 | 0.6 | 2.7 | 1.8 | 0.9 | 1.8 | 2.1 | 1.1 | 2.2 | 1.9 | 1.3 |
| TCT | 1.3 | 1.4 | 2.1 | 0.6 | 1.8 | 1.1 | 1.0 | 1.5 | 1.2 | 0.7 | 1.7 | 1.4 | 0.9 | 1.5 | 1.9 | 1.0 | 1.7 | 1.6 |
| GGT | 1.3 | 0.6 | 1.6 | 2.5 | 0.8 | 2.8 | 0.5 | 0.8 | 2.1 | 1.5 | 0.9 | 2.2 | 1.2 | 0.7 | 1.5 | 1.5 | 0.7 | 1.9 |
| GTG | 1.3 | 1.2 | 0.3 | 2.3 | 2.1 | 0.4 | 3.8 | 0.3 | 0.9 | 1.4 | 1.0 | 0.4 | 1.4 | 1.2 | 0.4 | 1.2 | 1.1 | 0.4 |
| TGG | 0.9 | 2.3 | 1.4 | 0.8 | 3.5 | 1.9 | 1.5 | 4.0 | 1.1 | 1.2 | 2.2 | 2.5 | 1.2 | 2.2 | 2.8 | 1.2 | 2.0 | 2.5 |
| GTT | 2.3 | 0.8 | 1.0 | 2.4 | 0.6 | 0.8 | 1.9 | 0.4 | 2.2 | 3.9 | 0.5 | 1.9 | 3.6 | 0.5 | 1.6 | 3.6 | 0.6 | 1.7 |
| TTG | 0.6 | 4.6 | 0.2 | 0.2 | 4.2 | 0.1 | 1.8 | 2.9 | 0.4 | 1.5 | 3.7 | 0.4 | 1.4 | 4.8 | 0.4 | 1.2 | 4.1 | 0.5 |
| TGT | 0.6 | 0.7 | 2.6 | 0.5 | 0.6 | 2.4 | 0.4 | 1.7 | 1.8 | 0.3 | 0.8 | 1.3 | 0.4 | 0.8 | 2.2 | 0.3 | 0.7 | 1.6 |

Table 2 (*continued*)

| | *Sulfolobus solfataricus* (SLS) | | | *Sulfolobus tokodaii* (SLT) | | | *Thermoplasma acidophilum* (TPA) | | | *Thermoplasma volcanium* (TPV) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| AAC | 1.7 | 1.2 | 1.7 | 1.4 | 1.3 | 1.8 | 2.4 | 0.7 | 1.4 | 2.2 | 1.0 | 1.6 |
| ACA | 1.4 | 1.5 | 1.1 | 1.7 | 1.1 | 1.1 | 1.4 | 3.0 | 0.7 | 1.7 | 2.3 | 0.9 |
| CAA | 1.6 | 2.3 | 1.5 | 1.6 | 2.5 | 1.0 | 0.3 | 2.2 | 2.5 | 0.8 | 2.6 | 2.2 |
| AAG | 3.8 | 3.2 | 2.3 | 2.8 | 4.1 | 2.4 | 3.9 | 1.2 | 1.6 | 3.5 | 2.6 | 2.0 |
| AGA | 2.5 | 2.7 | 4.1 | 2.6 | 2.2 | 4.9 | 1.6 | 3.6 | 2.4 | 1.7 | 2.7 | 3.7 |
| GAA | 3.9 | 2.0 | 2.6 | 4.7 | 2.1 | 2.1 | 2.8 | 1.7 | 3.4 | 4.0 | 1.8 | 2.7 |
| AAT | 3.3 | 1.8 | 4.6 | 3.5 | 2.1 | 4.9 | 1.9 | 1.2 | 2.2 | 2.6 | 1.6 | 3.5 |
| ATA | 5.0 | 3.8 | 1.5 | 5.0 | 4.1 | 1.6 | 5.6 | 1.8 | 1.4 | 5.6 | 2.9 | 1.5 |
| TAA | 0.0 | 5.1 | 4.6 | 0.0 | 5.4 | 4.9 | 0.0 | 2.7 | 1.1 | 0.0 | 3.9 | 2.5 |
| ACC | 0.7 | 0.6 | 1.3 | 0.4 | 0.5 | 1.4 | 1.4 | 1.1 | 0.8 | 0.8 | 1.0 | 0.9 |
| CCA | 1.6 | 0.9 | 0.5 | 1.7 | 0.5 | 0.4 | 1.3 | 2.2 | 0.9 | 1.5 | 1.2 | 0.8 |
| CAC | 0.5 | 0.8 | 0.7 | 0.4 | 0.9 | 0.5 | 0.8 | 0.8 | 1.4 | 0.7 | 0.9 | 0.9 |
| ACG | 0.6 | 1.8 | 0.2 | 0.4 | 1.2 | 0.1 | 1.4 | 2.5 | 0.3 | 0.9 | 2.3 | 0.3 |
| CGA | 0.1 | 0.7 | 1.3 | 0.1 | 0.4 | 0.9 | 0.2 | 1.5 | 2.7 | 0.2 | 1.0 | 2.1 |
| GAC | 1.3 | 0.5 | 0.8 | 1.0 | 0.5 | 0.6 | 2.0 | 0.3 | 1.1 | 1.8 | 0.3 | 0.9 |
| ACT | 2.0 | 1.3 | 1.6 | 2.2 | 1.1 | 1.6 | 0.7 | 1.1 | 1.3 | 1.4 | 1.2 | 1.5 |
| CTA | 1.9 | 2.5 | 0.8 | 1.7 | 2.9 | 0.6 | 0.6 | 0.6 | 1.3 | 1.3 | 1.8 | 1.0 |
| TAC | 1.7 | 1.6 | 1.7 | 1.3 | 1.6 | 1.9 | 2.6 | 1.2 | 0.9 | 2.2 | 1.3 | 1.5 |
| AGC | 0.7 | 0.7 | 2.1 | 0.6 | 0.7 | 2.4 | 1.6 | 1.2 | 1.3 | 1.3 | 1.0 | 1.8 |
| GCA | 1.9 | 0.5 | 0.6 | 2.1 | 0.4 | 0.5 | 2.2 | 1.9 | 1.3 | 2.4 | 1.4 | 1.0 |
| CAG | 0.5 | 1.9 | 0.6 | 0.5 | 2.2 | 0.5 | 1.9 | 2.0 | 2.6 | 1.3 | 2.5 | 1.3 |
| AGG | 1.8 | 2.5 | 2.3 | 1.2 | 1.8 | 2.5 | 2.6 | 3.3 | 1.5 | 2.0 | 2.3 | 2.0 |
| GGA | 2.6 | 1.5 | 2.5 | 2.6 | 1.2 | 1.7 | 2.2 | 1.7 | 3.8 | 2.2 | 1.5 | 2.6 |
| GAG | 3.0 | 1.7 | 1.6 | 2.3 | 1.8 | 1.1 | 3.2 | 0.9 | 1.6 | 2.3 | 1.2 | 1.2 |
| AGT | 1.7 | 1.3 | 3.0 | 1.7 | 1.0 | 3.2 | 0.5 | 0.9 | 1.2 | 0.8 | 1.0 | 2.0 |
| GTA | 2.8 | 1.7 | 0.8 | 3.0 | 1.7 | 0.6 | 1.4 | 0.6 | 1.0 | 2.5 | 0.9 | 0.9 |
| TAG | 0.0 | 4.7 | 2.3 | 0.0 | 4.9 | 2.1 | 0.0 | 2.3 | 0.5 | 0.0 | 3.3 | 1.2 |
| ATC | 1.1 | 1.4 | 1.3 | 0.9 | 1.4 | 1.3 | 2.2 | 2.2 | 2.1 | 1.5 | 1.7 | 1.9 |
| TCA | 1.6 | 1.3 | 1.2 | 1.7 | 1.0 | 1.3 | 1.6 | 3.6 | 0.9 | 1.9 | 2.0 | 0.9 |
| CAT | 0.8 | 1.4 | 1.5 | 1.0 | 1.6 | 1.1 | 0.9 | 1.5 | 4.2 | 0.9 | 1.5 | 2.6 |
| ATG | 2.1 | 3.2 | 0.5 | 2.1 | 3.7 | 0.6 | 3.1 | 3.1 | 0.5 | 2.7 | 3.3 | 0.5 |
| TGA | 0.0 | 2.1 | 3.7 | 0.0 | 1.7 | 4.2 | 0.0 | 3.2 | 2.9 | 0.0 | 2.2 | 3.5 |
| GAT | 3.4 | 1.0 | 2.1 | 3.7 | 1.0 | 1.7 | 3.8 | 1.2 | 3.8 | 3.7 | 1.0 | 2.6 |
| ATT | 3.4 | 2.2 | 3.4 | 4.0 | 2.6 | 3.7 | 1.3 | 1.5 | 1.7 | 2.2 | 1.9 | 2.3 |
| TTA | 4.1 | 3.7 | 1.8 | 4.8 | 4.3 | 2.1 | 0.4 | 1.2 | 1.0 | 1.4 | 2.8 | 1.4 |
| TAT | 3.1 | 2.5 | 3.2 | 3.6 | 2.5 | 4.1 | 2.0 | 1.8 | 1.7 | 2.6 | 2.2 | 3.1 |
| CCG | 0.4 | 0.6 | 0.1 | 0.4 | 0.4 | 0.0 | 1.3 | 1.9 | 0.3 | 0.7 | 1.0 | 0.2 |
| CGC | 0.1 | 0.3 | 0.7 | 0.1 | 0.2 | 0.4 | 0.5 | 1.3 | 1.4 | 0.2 | 0.7 | 0.9 |
| GCC | 0.7 | 0.2 | 0.6 | 0.5 | 0.2 | 0.5 | 2.1 | 1.0 | 1.1 | 1.3 | 0.7 | 1.0 |
| CCT | 1.2 | 0.9 | 0.8 | 1.4 | 0.6 | 0.6 | 0.8 | 0.9 | 1.9 | 1.2 | 0.8 | 1.5 |
| CTC | 0.7 | 1.0 | 0.9 | 0.6 | 1.1 | 0.8 | 1.9 | 0.6 | 1.0 | 1.2 | 0.8 | 0.9 |
| TCC | 0.8 | 0.7 | 1.4 | 0.5 | 0.6 | 1.5 | 1.6 | 1.2 | 1.3 | 1.0 | 1.0 | 1.2 |
| CGG | 0.1 | 0.6 | 0.7 | 0.0 | 0.4 | 0.7 | 0.3 | 1.8 | 2.3 | 0.2 | 1.0 | 1.5 |
| GGC | 0.7 | 0.5 | 1.1 | 0.5 | 0.4 | 0.8 | 2.5 | 1.0 | 2.0 | 1.9 | 0.8 | 1.4 |
| GCG | 0.7 | 0.5 | 0.1 | 0.4 | 0.4 | 0.1 | 1.4 | 1.6 | 0.5 | 0.8 | 1.1 | 0.3 |
| CGT | 0.2 | 0.7 | 1.2 | 0.1 | 0.4 | 0.7 | 0.3 | 0.7 | 2.1 | 0.3 | 0.6 | 1.5 |
| GTC | 0.7 | 0.5 | 0.6 | 0.6 | 0.4 | 0.4 | 1.6 | 0.3 | 0.8 | 1.1 | 0.3 | 0.7 |
| TCG | 0.5 | 1.0 | 0.1 | 0.3 | 0.7 | 0.1 | 1.2 | 2.5 | 0.3 | 0.9 | 1.7 | 0.2 |
| CTG | 0.6 | 2.0 | 0.2 | 0.4 | 2.3 | 0.2 | 2.8 | 1.9 | 0.5 | 1.2 | 2.5 | 0.4 |
| TGC | 0.2 | 0.7 | 1.7 | 0.2 | 0.6 | 2.0 | 0.5 | 1.6 | 2.2 | 0.4 | 1.2 | 2.3 |
| GCT | 2.3 | 0.4 | 0.9 | 2.6 | 0.4 | 0.7 | 1.3 | 0.7 | 2.2 | 1.9 | 0.6 | 1.6 |
| CTT | 1.5 | 1.5 | 1.9 | 1.9 | 1.9 | 1.6 | 2.0 | 0.7 | 1.5 | 2.5 | 1.2 | 1.6 |
| TTC | 1.8 | 1.4 | 1.6 | 1.4 | 1.6 | 1.9 | 3.2 | 1.5 | 1.6 | 2.2 | 1.3 | 2.0 |
| TCT | 1.5 | 1.3 | 1.4 | 1.9 | 1.2 | 1.5 | 1.0 | 1.6 | 2.0 | 1.7 | 1.2 | 1.7 |
| GGT | 2.2 | 0.7 | 1.3 | 2.4 | 0.6 | 0.9 | 1.5 | 0.6 | 2.0 | 1.7 | 0.7 | 1.3 |
| GTG | 1.2 | 1.4 | 0.3 | 0.8 | 1.5 | 0.2 | 2.1 | 1.1 | 0.3 | 1.2 | 1.2 | 0.3 |
| TGG | 1.1 | 1.8 | 2.3 | 1.0 | 1.2 | 2.3 | 0.9 | 2.8 | 2.0 | 0.8 | 1.9 | 2.4 |
| GTT | 2.8 | 1.0 | 1.9 | 3.0 | 1.1 | 1.6 | 2.1 | 0.5 | 1.3 | 2.4 | 0.7 | 1.3 |
| TTG | 1.6 | 3.0 | 0.6 | 1.0 | 3.5 | 0.6 | 0.7 | 3.0 | 0.1 | 1.2 | 3.6 | 0.3 |
| TGT | 0.4 | 0.9 | 2.0 | 0.5 | 0.8 | 2.5 | 0.1 | 1.1 | 1.9 | 0.2 | 1.0 | 2.4 |

Formula (8) using these frequencies $o(w_i^p)$ identifies 20 preferential sets $\mathscr{S}^{\text{pref}}$ of permutated trinucleotides in each archaeal genome (Table 3). The preferential sets $\mathscr{S}^{\text{pref}}$ have 218 out of $20 \times 16 = 320$, i.e. 68%, the highest value (first rank $Rk$) with the function $F$ among the 84 sets $\mathscr{S}^j$ in the 16 archaeal genomes and have 265 out of 320, i.e. 83%, the first 3 values (Table 3).

The 2 strong statistical results, highest value for the partition $E^{(22,44,53)}$ (see above) and more than 2 out of 3 the highest value for the set $\mathscr{S}^{\text{pref}}$, are a consequence of the distribution of the 60 trinucleotides to the 3 frames which is revealed by the sensibility of the method developed.

These preferential sets $\mathscr{S}^{\text{pref}}$ lead to 3 subsets $\mathscr{X}_0(\mathscr{G})$, $\mathscr{X}_1(\mathscr{G})$ and $\mathscr{X}_2(\mathscr{G})$ of 20 trinucleotides associated with the frames 0, 1 and 2, respectively, in the 16 archaeal genomes $\mathscr{G}$. Table 3 indicates directly the subset $\mathscr{X}_0(\mathscr{G})$, e.g.

$$\begin{aligned}\mathscr{X}_0(\text{AP}) = \{&\text{AAC}, \text{AAG}, \text{ATA}, \text{ACC},\\ &\text{GAC}, \text{TAC}, \text{AGC}, \text{GAG},\\ &\text{GTA}, \text{ATC}, \text{ATG}, \text{ATT},\\ &\text{GCC}, \text{CTC}, \text{GCG}, \text{GTC},\\ &\text{CTG}, \text{TTC}, \text{GTG}, \text{GTT}\}\end{aligned}$$

in the genome $\mathscr{G} = \text{AP}$. The 2 subsets $\mathscr{X}_1(\mathscr{G})$ and $\mathscr{X}_2(\mathscr{G})$ are deduced from the subset $\mathscr{X}_0(\mathscr{G})$ by the permutation property (Section 3.2).

The same 3 subsets of 20 trinucleotides are identified with the 2 archaeal genomes $\mathscr{G} = \{\text{AP}, \text{AG}\}$ (Table 3): $\mathscr{X}_i(\text{AP}) = \mathscr{X}_i(\text{AG})$, $i \in \{0, 1, 2\}$. Therefore, 15 subsets $\mathscr{X}_0(\mathscr{G})$ are identified in the archaeal genomes.

### 3.2. Permutation property

$\mathscr{P}(\mathscr{X}_0(\mathscr{G})) = \mathscr{X}_1(\mathscr{G})$ and $\mathscr{P}(\mathscr{X}_1(\mathscr{G})) = \mathscr{X}_2(\mathscr{G})$ in the 16 archaeal genomes $\mathscr{G}$ (by construction of the method). $\mathscr{X}_0(\mathscr{G})$ generates $\mathscr{X}_1(\mathscr{G})$ by 1 permutation and $\mathscr{X}_2(\mathscr{G})$ by 2 permutations.

### 3.3. Circular code property

#### 3.3.1. Definition of a circular code

*Recall of a few notations*: $\mathscr{A}$ being a finite alphabet, $\mathscr{A}^*$ denotes the words on $\mathscr{A}$ of finite length including the empty word of length 0 and $\mathscr{A}^+$, the words on $\mathscr{A}$ of finite length $\geqslant 1$. Let $w_1 w_2$ be the concatenation of the 2 words $w_1$ and $w_2$.

A subset $\mathscr{X}$ of $\mathscr{A}^+$ is a circular code if $\forall n, m \geqslant 1$ and $x_1, x_2, ..., x_n \in \mathscr{X}$, $y_1, y_2, ..., y_m \in \mathscr{X}$ and $r \in \mathscr{A}^*$, $s \in \mathscr{A}^+$, the equalities $s x_2 x_3 \ldots x_n r = y_1 y_2 \ldots y_m$ and $x_1 = rs$ imply $n = m$, $r = 1$ and $x_i = y_i$, $1 \leqslant i \leqslant n$ (Béal, 1993; Berstel and Perrin, 1985) (Fig. 2). In other terms, every word on $\mathscr{A}$ "written on a circle" has at most one decomposition (factorization) on $\mathscr{X}$. Therefore, the construction frame of any word generated by a circular code $\mathscr{X}$ (precisely, generated by any concatenation of words of a circular code) can be retrieved as there is a unique decomposition on $\mathscr{X}$.

In the following, $\mathscr{X}$ will be a set of words of length 3 as a gene is a concatenation of trinucleotides.

By excluding all the trinucleotides $w = lll$ and by gathering the remaining trinucleotides in sets of 3 trinucleotides so that, in each set, the 3 trinucleotides are deduced from each other by permutation, a potential circular code, has at most one trinucleotide per set. This rule applied to the alphabet $\mathscr{A} = \{\text{A}, \text{C}, \text{G}, \text{T}\}$ leads to 20 sets of 3 permutated trinucleotides and therefore, to $3^{20} \sim 3.5$ billions potential circular codes.

A complete study of circular codes with trinucleotides on the reduced alphabet $\mathscr{A} = \{\text{R}, \text{Y}\}$ is presented in Arquès and Michel (1997). It allows to introduce simply the different concepts, e.g. the concept of decomposition and some properties of circular codes. It is done "by hand" as there are only $3^2 = 9$ potential circular codes (2 sets of 3 permutated trinucleotides). Obviously, such a study cannot be realized on the alphabet $\mathscr{A} = \{\text{A}, \text{C}, \text{G}, \text{T}\}$ as there are 12 964 440 circular codes among $3^{20} \sim 3.5$ billions potential ones (Arquès and Michel, 1997). The search of circular codes on the alphabets at 3 (and more) letters needs the introduction of a few classical definitions and results in coding theory, in particular the concept of the flower automaton (Béal, 1993; Berstel and Perrin, 1985).

#### 3.3.2. Identification of circular codes in the archaeal genomes

The subsets $\mathscr{X}_0(\mathscr{G})$, $\mathscr{X}_1(\mathscr{G})$ and $\mathscr{X}_2(\mathscr{G})$ are maximal (i.e. 20 trinucleotides) circular codes in the 16 archaeal genomes $\mathscr{G}$.

A probabilistic model, based on the nucleotide frequencies with a hypothesis of absence of correlation between successive bases on a DNA strand, has been proposed by Koch and Lehmann (1997) for constructing and identifying circular codes. Recently, we have given a mathematical proof that their model can only generate a subset of circular codes which are obtained by the flower automaton method, in particular their model cannot generate the circular code of eukaryotes/prokaryotes (Lacan and Michel, 2001). So far, the flower automaton method is the most simple and complete one for identifying circular codes.

The flower automaton $\mathscr{F}(\mathscr{X})$ associated with a subset $\mathscr{X}$ of $\mathscr{A}^+$ has a particular state (labeled 1 in Fig. 3) and cycles issued from this state 1 and labeled by words of $\mathscr{X}$ (Fig. 3).

Therefore, to prove that "$\mathscr{X}(\mathscr{G})$ is a circular code" is equivalent to prove that $\mathscr{F}(\mathscr{X}(\mathscr{G}))$ does not contain 2 cycles labeled with the same word.

An example with the set $\mathscr{X}_0(\text{AP}) = \mathscr{X}_0(\text{AG})$ in the archaeal genomes AP and AG gives the main steps (without details) for proving that a set of words is a

Table 3
Identification of 20 preferential sets $\mathscr{S}^{\text{pref}}$ of permutated trinucleotides in each archaeal genome by giving the trinucleotide $w^0$ in frame 0 (reading frame) in each set $\mathscr{S}_k$. The preferential sets $\mathscr{S}^{\text{pref}}$ have 218 out of $20 \times 16 = 320$, i.e. 68%, the highest value (first rank $Rk$) with the function $F$ among the 84 sets $\mathscr{S}^j$ in the 16 archaeal genomes and have 265 out of 320, i.e. 83%, the first 3 values. The subsets $\mathscr{X}_0(\mathscr{G})$ in frame 0 in the 16 archaeal genomes $\mathscr{G}$ are directly deduced, e.g. $\mathscr{X}_0(\text{AP}) = \{\text{AAC, AAG, ATA, ACC, GAC, TAC, AGC, GAG, GTA, ATC, ATG, ATT, GCC, CTC, GCG, GTC, CTG, TTC, GTG, GTT}\}$ in the genome $\mathscr{G} = \text{AP}$.

| | AP | | AG | | HB | | MC | | MP | | MSA | | MSM | | MT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathscr{S}_k$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ |
| (AAC,ACA,CAA) | AAC | 1 | AAC | 1 | AAC | 1 | ACA | 1 | AAC | 1 | AAC | 1 | AAC | 6 | AAC | 1 |
| (AAG,AGA,GAA) | AAG | 1 | AAG | 1 | AAG | 2 | GAA | 1 | AAG | 1 | GAA | 1 | GAA | 1 | AAG | 2 |
| (AAT,ATA,TAA) | ATA | 1 | ATA | 1 | AAT | 33 | ATA | 1 | ATA | 3 | ATA | 2 | ATA | 1 | ATA | 1 |
| (ACC,CCA,CAC) | ACC | 2 | ACC | 2 | ACC | 4 | CCA | 1 | ACC | 7 | ACC | 2 | ACC | 7 | ACC | 4 |
| (ACG,CGA,GAC) | GAC | 1 | GAC | 1 | GAC | 1 | GAC | 1 | GAC | 1 | GAC | 1 | GAC | 1 | GAC | 1 |
| (ACT,CTA,TAC) | TAC | 8 | TAC | 1 | TAC | 1 | ACT | 1 | TAC | 6 | TAC | 12 | ACT | 6 | TAC | 3 |
| (AGC,GCA,CAG) | AGC | 2 | AGC | 19 | CAG | 2 | GCA | 1 | CAG | 2 | GCA | 1 | GCA | 1 | GCA | 2 |
| (AGG,GGA,GAG) | GAG | 1 | GAG | 1 | GAG | 1 | GGA | 7 | GAG | 1 | GAG | 1 | GAG | 3 | GAG | 1 |
| (AGT,GTA,TAG) | GTA | 1 | GTA | 1 | TAG | 7 | GTA | 1 | GTA | 3 | GTA | 1 | GTA | 1 | GTA | 1 |
| (ATC,TCA,CAT) | ATC | 1 | ATC | 1 | ATC | 1 | TCA | 2 | ATC | 1 | ATC | 1 | ATC | 1 | ATC | 1 |
| (ATG,TGA,GAT) | ATG | 1 | ATG | 5 | ATG | 1 | GAT | 1 | ATG | 1 | GAT | 1 | GAT | 1 | GAT | 1 |
| (ATT,TTA,TAT) | ATT | 1 | ATT | 1 | TAT | 3 | ATT | 3 | ATT | 3 | ATT | 1 | ATT | 1 | ATT | 1 |
| (CCG,CGC,GCC) | GCC | 1 | GCC | 1 | GCC | 1 | GCC | 2 | GCC | 1 | GCC | 1 | GCC | 1 | GCC | 1 |
| (CCT,CTC,TCC) | CTC | 1 | CTC | 3 | CTC | 1 | CCT | 1 | CTC | 1 | CTC | 1 | CTC | 1 | CTC | 1 |
| (CGG,GGC,GCG) | GCG | 10 | GCG | 4 | GGC | 6 | GCG | 3 | GCG | 1 | GGC | 2 | GGC | 1 | GGC | 7 |
| (CGT,GTC,TCG) | GTC | 1 | GTC | 1 | GTC | 1 | GTC | 1 | GTC | 1 | GTC | 1 | GTC | 1 | GTC | 1 |
| (CTG,TGC,GCT) | CTG | 3 | CTG | 5 | CTG | 1 | GCT | 1 | CTG | 1 | GCT | 3 | GCT | 2 | GCT | 6 |
| (CTT,TTC,TCT) | TTC | 1 | TTC | 9 | TTC | 1 | TTC | 13 | TTC | 1 | CTT | 4 | CTT | 2 | TTC | 2 |
| (GGT,GTG,TGG) | GTG | 1 | GTG | 1 | GTG | 1 | GGT | 1 | GTG | 1 | GTG | 1 | GTG | 1 | GTG | 1 |
| (GTT,TTG,TGT) | GTT | 1 | GTT | 1 | TTG | 1 | GTT | 1 | TTG | 1 | GTT | 1 | GTT | 1 | GTT | 1 |

| | PB | | PCA | | PCF | | PCH | | SLS | | SLT | | TPA | | TPV | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathscr{S}_k$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ | $w^0$ | $Rk$ |
| (AAC,ACA,CAA) | AAC | 1 | AAC | 1 | ACA | 7 | ACA | 8 | ACA | 3 | ACA | 1 | AAC | 1 | AAC | 3 |
| (AAG,AGA,GAA) | GAA | 8 | AAG | 1 | GAA | 1 | GAA | 10 | GAA | 2 | GAA | 1 | AAG | 1 | GAA | 3 |
| (AAT,ATA,TAA) | ATA | 1 | ATA | 1 | ATA | 1 | ATA | 1 | ATA | 1 | ATA | 2 | ATA | 1 | ATA | 1 |
| (ACC,CCA,CAC) | ACC | 6 | ACC | 25 | CCA | 21 | CCA | 18 | CCA | 1 | CCA | 1 | ACC | 1 | CCA | 10 |
| (ACG,CGA,GAC) | GAC | 1 | GAC | 1 | GAC | 1 | GAC | 1 | GAC | 1 | GAC | 1 | GAC | 1 | GAC | 1 |
| (ACT,CTA,TAC) | TAC | 1 | TAC | 7 | TAC | 22 | CTA | 36 | ACT | 1 | ACT | 1 | TAC | 2 | ACT | 16 |
| (AGC,GCA,CAG) | GCA | 3 | GCA | 1 | GCA | 1 | GCA | 1 | GCA | 1 | GCA | 1 | AGC | 6 | GCA | 1 |
| (AGG,GGA,GAG) | GAG | 1 | GAG | 1 | GAG | 2 | GAG | 1 | GAG | 1 | GGA | 2 | GAG | 1 | GAG | 1 |
| (AGT,GTA,TAG) | GTA | 1 | GTA | 1 | GTA | 1 | GTA | 1 | GTA | 1 | GTA | 1 | GTA | 1 | GTA | 1 |
| (ATC,TCA,CAT) | ATC | 1 | ATC | 1 | ATC | 1 | ATC | 1 | TCA | 6 | TCA | 2 | ATC | 1 | ATC | 7 |
| (ATG,TGA,GAT) | ATG | 5 | GAT | 2 | GAT | 1 | GAT | 1 | GAT | 1 | GAT | 1 | ATG | 6 | GAT | 1 |
| (ATT,TTA,TAT) | ATT | 1 | ATT | 1 | ATT | 1 | ATT | 1 | ATT | 9 | ATT | 5 | TAT | 19 | ATT | 2 |
| (CCG,CGC,GCC) | GCC | 1 | GCC | 1 | GCC | 1 | GCC | 1 | GCC | 1 | GCC | 1 | GCC | 1 | GCC | 1 |
| (CCT,CTC,TCC) | CTC | 1 | CTC | 1 | CTC | 4 | CTC | 2 | CCT | 1 | CCT | 1 | CTC | 2 | CTC | 2 |
| (CGG,GGC,GCG) | GCG | 7 | GCG | 2 | GCG | 1 | GCG | 1 | GCG | 2 | GGC | 8 | GGC | 2 | GGC | 1 |
| (CGT,GTC,TCG) | GTC | 1 | GTC | 1 | GTC | 1 | GTC | 1 | GTC | 1 | GTC | 1 | GTC | 1 | GTC | 1 |
| (CTG,TGC,GCT) | CTG | 1 | GCT | 1 | GCT | 1 | GCT | 1 | GCT | 1 | GCT | 1 | CTG | 5 | GCT | 1 |
| (CTT,TTC,TCT) | TTC | 1 | TTC | 1 | CTT | 1 | TTC | 6 | TTC | 7 | TCT | 1 | TTC | 7 | CTT | 18 |
| (GGT,GTG,TGG) | GTG | 1 | GTG | 2 | GTG | 10 | GTG | 9 | GGT | 1 | GGT | 1 | GTG | 1 | GGT | 3 |
| (GTT,TTG,TGT) | GTT | 4 | GTT | 1 | GTT | 1 | GTT | 1 | GTT | 1 | GTT | 1 | GTT | 1 | GTT | 1 |

circular code. The flower automaton $\mathscr{F}(\mathscr{X}_0(\text{AP}))$ of $\mathscr{X}_0(\text{AP})$ is constructed by reading the 20 words of $\mathscr{X}_0(\text{AP})$ in the way given in Fig. 3. As all words of $\mathscr{X}_0(\text{AP})$ are trinucleotides, i.e. of length equal to 3, the states of $\mathscr{F}(\mathscr{X}_0(\text{AP}))$ can be associated with frames: state 1 with frame 0, states 2–5 with frame 1 and states 6–15 with frame 2. The search of 2 cycles labeled with the same word in $\mathscr{F}(\mathscr{X}_0(\text{AP}))$ can be realized by comparing all the words starting from the frames 0 (state 1) and 1 (states 2–5) (Fig. 4). The absence of 2 cycles is verified when all the compared words are "non-extensible" (Fig. 4).

The implementation in Java of the flower automaton algorithm (not described here) demonstrates that the 15

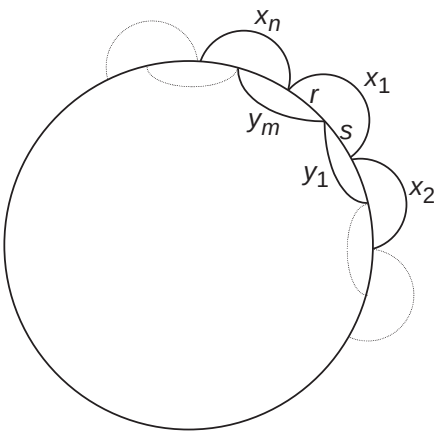


Fig. 2. A representation of the definition of a circular code.

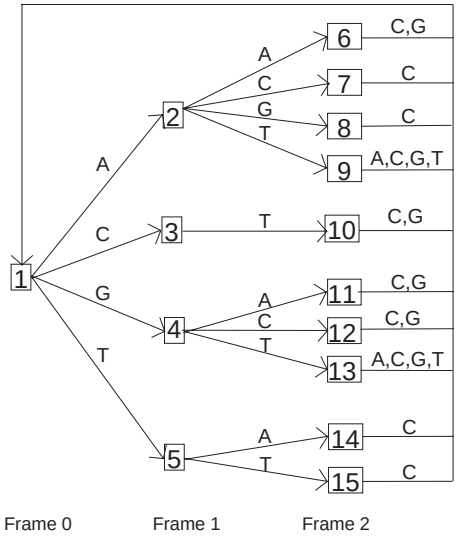


Fig. 3. Flower automaton $\mathscr{F}(\mathscr{X}_0(\mathrm{AP}))$ of the archaeal $\mathscr{C}^3$ code $\mathscr{X}_0(\mathrm{AP})$.

sets $\mathscr{X}_0(\mathscr{G})$ are (maximal) circular code with 2 permutated (maximal) circular codes $\mathscr{X}_1(\mathscr{G})$ and $\mathscr{X}_2(\mathscr{G})$. Therefore, 15 $\mathscr{C}^3$ codes $\mathscr{X}_0(\mathscr{G})$ in reading frame are identified in the archaeal genomes.

*Remark*: Nine $\mathscr{C}^3$ codes $\mathscr{X}_0(\mathscr{G})$ in the 9 genomes $\mathscr{G} = \{\{\mathrm{AP}, \mathrm{AG}\}, \mathrm{MP}, \mathrm{MSA}, \mathrm{MSM}, \mathrm{MT}, \mathrm{PB}, \mathrm{PCA}, \mathrm{PCH}, \mathrm{SLT}\}$ are identified with a strong statistical significance. Indeed, their $20 \times 9 = 180$ preferential sets $\mathscr{S}_k^{\mathrm{pref}}$ of permutated trinucleotides have values $F\left(\mathscr{S}_k^{\mathrm{pref}}\right) \geqslant 1/3$ (data not shown), i.e. different from the random case (Property 1 in the Method).

Five $\mathscr{C}^3$ codes $\mathscr{X}_0(\mathscr{G})$ have one among 20 set $\mathscr{S}_k$, named $\tilde{\mathscr{S}}_k$, which is not preferential as its value $F(\tilde{\mathscr{S}}_k) \approx 1/3$ is close to the random case (Property 1 in Method), precisely $F(\tilde{\mathscr{S}}_k) < 0.37$ (data not shown). In these sets $\tilde{\mathscr{S}}_k$, the occurrence probabilities of the 3 permutated trinucleotides are similar in the 3 frames. These sets are: $\tilde{\mathscr{S}}_4$ in $\mathscr{G} = \mathrm{TPV}$, $\tilde{\mathscr{S}}_6$ in $\mathscr{G} = \mathrm{PCF}$, $\tilde{\mathscr{S}}_{12}$ in $\mathscr{G} = \mathrm{TPA}$ and $\mathscr{S}_{18}$ in $\mathscr{G} = \{\mathrm{MC}, \mathrm{SLS}\}$ (data not shown).

The $\mathscr{C}^3$ code $\mathscr{X}_0(\mathrm{HB})$ has 2 sets $\tilde{\mathscr{S}}_k$: $\tilde{\mathscr{S}}_3 = \{\mathrm{AAT}, \mathrm{ATA}, \mathrm{TAA}\}$ and $\tilde{\mathscr{S}}_{12} = \{\mathrm{ATT}, \mathrm{TTA}, \mathrm{TAT}\}$ (Table 2). $\tilde{\mathscr{S}}_3$ and $\tilde{\mathscr{S}}_{12}$ are the only sets among 20 only made of A and T. As the genome HB has the highest GC content (68%; Ng et al., 2000) compared to the other genomes, its frequencies of A and T are low (data not shown).

These sets $\tilde{\mathscr{S}}_k$ are rare as there are in total 7 sets $\tilde{\mathscr{S}}_k$ among $15 \times 20 = 300$ sets $\tilde{\mathscr{S}}_k^{\mathrm{pref}}$, i.e. about 2%. In these sets $\tilde{\mathscr{S}}_k$, the preferential set $\tilde{\mathscr{S}}_k^{\mathrm{pref}}$ is assigned to one of the 3 sets $\mathscr{S}^{22}$, $\mathscr{S}^{44}$ and $\mathscr{S}^{53}$ leading to a $\mathscr{C}^3$ code $\mathscr{X}_0(\mathscr{G})$. Note that a solution has always been found with these 7 sets $\tilde{\mathscr{S}}_k$ as 15 $\mathscr{C}^3$ codes $\mathscr{X}_0(\mathscr{G})$ have been identified.

### 3.3.3. *Identical circular codes on both DNA strands in the archaeal genomes*

In each archaeal genome, the same $\mathscr{C}^3$ code $\mathscr{X}_0(\mathscr{G})$ is identified in genes of the direct and complementary DNA strands as the occurrence frequencies of the 64 trinucleotides in the 3 frames of genes are almost identical in both strands (data not shown).

### 3.4. *Construction frame determination property in the archaeal circular codes*

The property of the determination of the construction frame (decomposition) of a word generated by any concatenation of words of a circular code (more simply, generated by a circular code), is a consequence of the circular code property. The principle is presented briefly with the previous example of the circular code $\mathscr{X}_0(\mathrm{AP}) = \mathscr{X}_0(\mathrm{AG})$ in the archaeal genomes AP and AG (see Béal, 1993; Berstel and Perrin, 1985 for the details). The unicity of the decomposition of a word into words of a circular code is not obvious.

For example, the word $w' = \mathrm{TAGCGACCT}$ of length 9 can be decomposed into trinucleotides of $\mathscr{X}_0(\mathrm{AP})$ in 2 ways:

- $\{12,4\}\mathrm{T}\{9,13\}\mathrm{A1G4C12G1A2C7C1T5}$ with $l_1\mathrm{TA}, \mathrm{GCG}, \mathrm{ACC}, \mathrm{T}l_2l_3 \in \mathscr{X}_0(\mathrm{AP})$
- $\{9,13\}\mathrm{T1A2G8C1G4A11C1C3T10}$ with $l_1l_2\mathrm{T}, \mathrm{AGC}, \mathrm{GAC}, \mathrm{CT}l_3 \in \mathscr{X}_0(\mathrm{AP})$

where the states of the flower automaton $\mathscr{F}(\mathscr{X}_0(\mathrm{AP}))$ of $\mathscr{X}_0(\mathrm{AP})$ separate the letters of $w'$ (Fig. 3). Therefore, the word $w'$ has 2 construction frames: frame 1 (initial states {2,4} of $\mathscr{F}(\mathscr{X}_0(\mathrm{AP}))$) or frame 2 (initial states {9,13} of $\mathscr{F}(\mathscr{X}_0(\mathrm{AP}))$) (Fig. 3). With the example chosen, the right factor GCGACCT of $w'$ is one of the 4 longest words of length 7 identified in Fig. 4. It begins at the states 1 and 2 in $\mathscr{F}(\mathscr{X}_0(\mathrm{AP}))$ (Fig. 3). The left factor TA of $w'$ can be concatenated with this right factor by ending the letter A of TA at the states 1 and 2 in $\mathscr{F}(\mathscr{X}_0(\mathrm{AP}))$ (Fig. 3).

The unicity of such a decomposition is proved by using the properties of the flower automaton $\mathscr{F}(\mathscr{X}(\mathscr{G}))$

Words starting from the states 1 and 2 in the flower automaton $\mathcal{F}(\mathcal{X}_0(\mathrm{AP}))$

A
AC
ACC
**ACCT**
AG
AGC
**AGCT**
**C**
G
GC
GCC
**GCCT**
GCG
GCGA
GCGAC
GCGACC
**GCGACCT**
GCGAG
GCGAGC
**GCGAGCT**
**GCGC**
GCGT
GCGTA
GCGTAC
**GCGTACT**
GCGTT
GCGTTC
**GCGTTCT**
T
TA
TAC
**TACT**
TT
TTC
**TTCT**

Words starting from the states 1 and 3 in the flower automaton $\mathcal{F}(\mathcal{X}_0(\mathrm{AP}))$

**T**

Words starting from the states 1 and 4 in the flower automaton $\mathcal{F}(\mathcal{X}_0(\mathrm{AP}))$

A
AC
ACC
**ACCT**
AG
AGC
**AGCT**
**C**
T
TA
TAC
**TACT**
TT
TTC
**TTCT**

Words starting from the states 1 and 5 in the flower automaton $\mathcal{F}(\mathcal{X}_0(\mathrm{AP}))$

A
AC
ACC
**ACCT**
**T**

Fig. 4. Absence of 2 cycles labeled with the same word in the flower automaton $\mathcal{F}(\mathcal{X}_0(\mathrm{AP}))$ proving that the set $\mathcal{X}_0(\mathrm{AP})$ in the archaeal genome AP is a circular code. The words in bold are "non-extensible".

of the circular code $\mathcal{X}(\mathcal{G})$. If any letter of a word generated by $\mathcal{X}(\mathcal{G})$ can be associated with a unique state of $\mathcal{F}(\mathcal{X}(\mathcal{G}))$, then a frame can be assigned to this letter as each state of $\mathcal{F}(\mathcal{X}(\mathcal{G}))$ is related to a frame. In the case of a flower automaton $\mathcal{F}$, there exists a number $n(\mathcal{X}(\mathcal{G}))$ so that for a word of length $\geqslant n(\mathcal{X}(\mathcal{G}))$, all paths associated with this word have the same terminal state, i.e. a unique state for a letter of this word is identified. Thus, the construction frame of the word can be determined and the word can be decomposed into

Table 4
Lengths of the minimal windows to retrieve the construction frames 0 (reading frame), 1 and 2 with the 15 archaeal circular codes $\mathscr{X}_0(\mathscr{G})$, $\mathscr{X}_1(\mathscr{G})$ and $\mathscr{X}_2(\mathscr{G})$, respectively

| $\mathscr{G}$ | AP,AG | HB | MC | MP | MSA | MSM | MT | PB | PCA | PCF | PCH | SLS | SLT | TPA | TPV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathscr{X}_0(\mathscr{G})$ | 10 | 9 | 7 | 10 | 10 | 9 | 10 | 7 | 10 | 7 | 10 | 7 | 5 | 7 | 9 |
| $\mathscr{X}_1(\mathscr{G})$ | 10 | 10 | 10 | 8 | 10 | 11 | 10 | 8 | 8 | 6 | 8 | 10 | 7 | 7 | 11 |
| $\mathscr{X}_2(\mathscr{G})$ | 7 | 11 | 7 | 11 | 10 | 7 | 10 | 10 | 10 | 6 | 9 | 10 | 6 | 7 | 10 |

words of the circular code. In other words, the length of the minimal window to retrieve the construction frame is $n(\mathscr{X}(\mathscr{G}))$ letters.

With the previous example, $n(\mathscr{X}_0(\mathrm{AP})) = 10$ (first line and column in Table 4), i.e. the length of the minimal window to retrieve the construction frame with $\mathscr{X}_0(\mathrm{AP})$ is 10 letters. The word $w'$ of length 9 was chosen in the set of the longest words with 2 construction frames. The concatenation of only one letter $l$ to $w'$ leads to the word $w = w'l$ of length 10 with a unique construction frame as 10 is the length of the minimal window. For example with $l = \{\mathrm{A},\mathrm{T}\}$, the construction frame of $w$ is unique and equal to 1 as there is no edge labeled {A,T} leaving the state 10 in $\mathscr{F}(\mathscr{X}_0(\mathrm{AP}))$ (Fig. 3). Then, the unique decomposition of $w$ into trinucleotides of $\mathscr{X}_0(\mathrm{AP})$ is $l_1\mathrm{TA},\mathrm{GCG},\mathrm{ACC},\mathrm{T}\{\mathrm{A},\mathrm{T}\}l_2$.

The program developed in Java testing all possible paths in a flower automaton (not described here) allows to compute the lengths $n(\mathscr{X}(\mathscr{G}))$ of the minimal windows of the archaeal circular codes $\mathscr{X}_0(\mathscr{G})$, $\mathscr{X}_1(\mathscr{G})$ and $\mathscr{X}_2(\mathscr{G})$ to retrieve the construction frames 0, 1 and 2, respectively (Table 4). These lengths vary between 5 for $\mathscr{X}_0(\mathrm{SLT})$ and 11 for $\mathscr{X}_1(\mathrm{MSM})$, $\mathscr{X}_1(\mathrm{TPV})$, $\mathscr{X}_2(\mathrm{HB})$ and $\mathscr{X}_2(\mathrm{MK})$ (Table 4).

### 3.5. Probabilities of words generated by the archaeal circular codes

A new definition of a simple parameter is introduced for measuring some probabilities of words generated by the circular codes. Let $N(\mathscr{L})$ be the number of words of a given length $\mathscr{L}$, i.e. $N(\mathscr{L}) = 4^{\mathscr{L}}$ on the alphabet {A,C,G,T}. Let $\mathscr{F}_p(\mathscr{L},\mathscr{X}(\mathscr{G}))$ be the set of words of a given length $\mathscr{L}$ obtained from all the states of the frame $p \in \{0,1,2\}$, of the flower automaton $\mathscr{F}$ of the circular code $\mathscr{X}(\mathscr{G})$. By using the classical Poincaré formula, the number $N(\mathscr{L},\mathscr{X}(\mathscr{G}))$ of words of a given length $\mathscr{L}$ generated by the circular code $\mathscr{X}(\mathscr{G})$ is equal to

$$N(\mathscr{L},\mathscr{X}(\mathscr{G})) = \mathrm{CARD}\left(\bigcup_{p=0}^{2}\mathscr{F}_p(\mathscr{L},\mathscr{X}(\mathscr{G}))\right) = \sum_{p=0}^{2}(-1)^p S_p \tag{9}$$

with $S_p = \sum_{0\leqslant i_0<\ldots<i_p\leqslant 2}\mathrm{CARD}\left(\mathscr{F}_{i_0}(\mathscr{L},\mathscr{X}(\mathscr{G}))\bigcap\cdots\bigcap\mathscr{F}_{i_p}(\mathscr{L},\mathscr{X}(\mathscr{G}))\right)$.

Then, its probability $\mathscr{P}(\mathscr{L},\mathscr{X}(\mathscr{G}))$ is equal to

$$\mathscr{P}(\mathscr{L},\mathscr{X}(\mathscr{G})) = \frac{N(\mathscr{L},\mathscr{X}(\mathscr{G}))}{N(\mathscr{L})}. \tag{10}$$

The words with a unique construction frame generated by the circular code $\mathscr{X}(\mathscr{G})$ are the words obtained from all the states of a unique frame, i.e. the words with an empty intersection in the 3 sets $\mathscr{F}_p(\mathscr{L},\mathscr{X}(\mathscr{G}))$. Therefore, the number $M(\mathscr{L},\mathscr{X}(\mathscr{G}))$ of words of a given length $\mathscr{L}$ with a unique construction frame generated by the circular code $\mathscr{X}(\mathscr{G})$, is equal to

$$\begin{aligned} &M(\mathscr{L},\mathscr{X}(\mathscr{G})) \\ &= N(\mathscr{L},\mathscr{X}(\mathscr{G})) \\ &\quad -\mathrm{CARD}\left(\bigcup_{0\leqslant i_0<i_1\leqslant 2}\left(\mathscr{F}_{i_0}(\mathscr{L},\mathscr{X}(\mathscr{G}))\bigcap\mathscr{F}_{i_1}(\mathscr{L},\mathscr{X}(\mathscr{G}))\right)\right) \\ &= \sum_{p=0}^{2}(-1)^p(p+1)S_p \end{aligned} \tag{11}$$

and its probability $\mathscr{Q}(\mathscr{L},\mathscr{X}(\mathscr{G}))$, to

$$\mathscr{Q}(\mathscr{L},\mathscr{X}(\mathscr{G})) = \frac{M(\mathscr{L},\mathscr{X}(\mathscr{G}))}{N(\mathscr{L})}. \tag{12}$$

With the previous example of the $\mathscr{C}^3$ code $\mathscr{X}_0(\mathrm{AP}) = \mathscr{X}_0(\mathrm{AG})$ in the archaeal genomes AP and AG, Table 5 shows the probabilities $\mathscr{P}(\mathscr{L},\mathscr{X}_0(\mathrm{AP}))$ and $\mathscr{Q}(\mathscr{L},\mathscr{X}_0(\mathrm{AP}))$ by varying the word length $\mathscr{L} \in \{3,11\}$. With $\mathscr{L} = 3$, only 1 word, GGG, cannot be generated by the flower automaton $\mathscr{F}(\mathscr{X}_0(\mathrm{AP}))$ (Fig. 3). Therefore, $\mathscr{P}(3,\mathscr{X}_0(\mathrm{AP})) = (4^3-1)/4^3 \approx 0.984$. With $\mathscr{L} = 9$, 4 words has 2 construction frames (Fig. 3): TAGCGACCT, TAGCGAGCT, TAGCGTACT and TAGCGTTCT with the frames 1 (initial states {2,4} in $\mathscr{F}(\mathscr{X}_0(AP))$) and 2 (initial states {9,13} in $\mathscr{F}(\mathscr{X}_0(AP))$). Note that the suffixes of length 7 of these 4 words are the 4 words with the greatest length in Fig. 4. No word has 2 construction frames with $\mathscr{L} = 10$ which is the length of the minimal window to retrieve the construction frame 0 with $\mathscr{X}_0(\mathrm{AP})$ (Table 4).

The 2 probabilities $\mathscr{P}(\mathscr{L},\mathscr{X}_0(\mathscr{G}))$ and $\mathscr{Q}(\mathscr{L},\mathscr{X}_0(\mathscr{G}))$ with the 15 archaeal $\mathscr{C}^3$ codes $\mathscr{X}_0(\mathscr{G})$ (in reading frame) by varying the word length $\mathscr{L} \in \{3,11\}$ are given in

Table 5
Probability $\mathscr{P}(\mathscr{L},\mathscr{X}_0(\mathrm{AP}))$ (%) of words generated by the archaeal $\mathscr{C}^3$ code $\mathscr{X}_0(\mathrm{AP})$ (in reading frame) and probability $\mathscr{Q}(\mathscr{L},\mathscr{X}_0(\mathrm{AP}))$ (%) of words with a unique construction frame also generated by the same code $\mathscr{X}_0(\mathrm{AP})$ by varying the word length between 3 and 11

| Length $\mathscr{L}$ | Number of words not generated | Number of words in 1 frame | Number of words in 2 frames | Number of words in 3 frames | Sum | Probability $\mathscr{P}(\mathscr{L},\mathscr{X}_0(\mathrm{AP}))$ (%) | Probability $\mathscr{Q}(\mathscr{L},\mathscr{X}_0(\mathrm{AP}))$ (%) |
|---|---|---|---|---|---|---|---|
| 3 | 1 | 32 | 29 | 2 | 64 | 98.4 | 50.0 |
| 4 | 43 | 176 | 37 | 0 | 256 | 83.2 | 68.8 |
| 5 | 360 | 628 | 36 | 0 | 1024 | 64.8 | 61.3 |
| 6 | 2205 | 1862 | 29 | 0 | 4096 | 46.2 | 45.5 |
| 7 | 11 396 | 4976 | 12 | 0 | 16 384 | 30.4 | 30.4 |
| 8 | 51 544 | 13 984 | 8 | 0 | 65 536 | 21.4 | 21.3 |
| 9 | 223 748 | 38 392 | 4 | 0 | 262 144 | 14.6 | 14.6 |
| 10 | 948 576 | 100 000 | 0 | 0 | 1 048 576 | 9.5 | 9.5 |
| 11 | 3 914 304 | 280 000 | 0 | 0 | 4 194 304 | 6.7 | 6.7 |

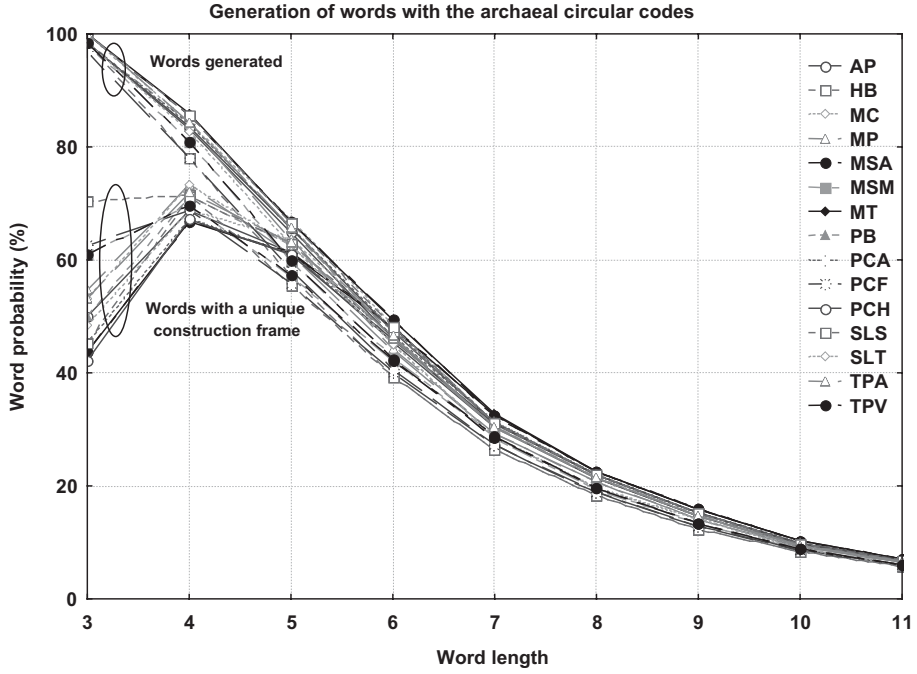


Fig. 5. Probability $\mathscr{P}(\mathscr{L},\mathscr{X}_0(\mathscr{G}))$ (%) of words generated by the 15 archaeal $\mathscr{C}^3$ codes $\mathscr{X}_0(\mathscr{G})$ (in reading frame) and probability $\mathscr{Q}(\mathscr{L},\mathscr{X}_0(\mathscr{G}))$ (%) of words with a unique construction frame also generated by the same codes $\mathscr{X}_0(\mathscr{G})$ by varying the word length between 3 and 11.

Fig. 5. For all the codes, the probabilities $\mathscr{P}(\mathscr{L},\mathscr{X}_0(\mathscr{G}))$ decrease from a mean value of $98.9\pm0.8\%$, where 0.8 represents the average deviation, at $\mathscr{L}=3$, to a mean value of $6.5\pm0.4\%$ at $\mathscr{L}=11$. Five codes in the archaeal genomes $\mathscr{G}=\{\mathrm{MC},\mathrm{MSA},\mathrm{MT},\mathrm{SLS},\mathrm{TPA}\}$ generate all the 64 trinucleotides (100% at $\mathscr{L}=3$). For all the codes, the probabilities $\mathscr{Q}(\mathscr{L},\mathscr{X}_0(\mathscr{G}))$ increase from a mean value of $52.3\pm6.6\%$ at $\mathscr{L}=3$, to a maximum with a mean value of $69.7\pm1.9\%$ at $\mathscr{L}=4$, then decrease and begin to merge with the probabilities $\mathscr{P}(\mathscr{L},\mathscr{X}_0(\mathscr{G}))$ at $\mathscr{L}=5$. Therefore, almost all the words of lengths $\geqslant5$ generated by the 15 archaeal $\mathscr{C}^3$ codes allow to retrieve the reading frame in genes. The number of words with a unique construction frame also differs between the different codes. For example at $\mathscr{L}=5$, $\mathscr{Q}(5,\mathscr{X}_0(\mathrm{MC}))=64\%$ while $\mathscr{Q}(5,\mathscr{X}_0(\mathrm{HB}))=55\%$, i.e. the code $\mathscr{X}_0(\mathrm{MC})$ generates 84 more words of length 5 with a unique construction frame compared to the code $\mathscr{X}_0(\mathrm{HB})$. The code $\mathscr{X}_0(\mathrm{MC})$ may lead to a greater gene diversity. Otherwise, an analytical model of gene evolution based on an independent mixing of the trinucleotides of the $\mathscr{C}^3$ code of eukaryotes/prokaryotes allows to simulate the actual genes of eukaryotes/prokaryotes (Arquès et al., 1998). Therefore, the probabilities $\mathscr{P}(\mathscr{L},\mathscr{X}(\mathscr{G}))$ and $\mathscr{Q}(\mathscr{L},\mathscr{X}(\mathscr{G}))$ may represent evolutionary parameters of the word diversity of the

Table 6
Preferential frame (line "MEAN") for the 8 2-letter trinucleotides on the alphabet {R,Y} (R=purine={A,G}, Y=pyrimidine={C,T}) in the 16 archaeal genomes. This preferential frame is deduced from the average frame associated with the 8 4-letter trinucleotides specified on the 2-letter trinucleotide and belonging to the archaeal circular code $\mathscr{X}_0(\mathscr{G})$ (frame 0 in bold) $\mathscr{X}_1(\mathscr{G})$ (frame 1) and $\mathscr{X}_2(\mathscr{G})$ (frame 2). For example, RRY with $\mathscr{G}$ = AG is in frame 2 (3 trinucleotides of $\mathscr{X}_0(\mathscr{G})$ and 5 trinucleotides of $\mathscr{X}_2(\mathscr{G})$), RRY with $\mathscr{G}$ = MSA is in the 2 frames 0 and 2 (4 trinucleotides of $\mathscr{X}_0(\mathscr{G})$ and $\mathscr{X}_2(\mathscr{G})$)

| | | AG, AP | HB | MC | MP | MSA | MSM | MT | PB | PCA | PCF | PCH | SLS | SLT | TPA | TPV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RRR | AAA | — | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| | AAG | **0** | **0** | 1 | **0** | 1 | 1 | **0** | 1 | **0** | 1 | 1 | 1 | 1 | **0** | 1 |
| | AGA | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 2 |
| | AGG | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
| | GAA | 2 | 2 | **0** | 2 | **0** | **0** | 2 | **0** | 2 | **0** | **0** | **0** | **0** | 2 | **0** |
| | GAG | **0** | **0** | 1 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | 1 | **0** | **0** |
| | GGA | 2 | 2 | **0** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | **0** | 2 | 2 |
| | GGG | — | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| | MEAN | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 |
| RRY | AAC | **0** | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | 2 | 2 | 2 | 2 | **0** | **0** |
| | AAT | 2 | **0** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | AGC | **0** | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | **0** | 2 |
| | AGT | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | GAC | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | GAT | 2 | 2 | **0** | 2 | **0** | **0** | **0** | 2 | **0** | **0** | **0** | **0** | **0** | 2 | **0** |
| | GGC | 2 | **0** | 2 | 2 | **0** | **0** | **0** | 2 | 2 | 2 | 2 | 2 | **0** | **0** | **0** |
| | GGT | 2 | 2 | **0** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | **0** | **0** | 2 | **0** |
| | MEAN | 2 | **0** | 2 | 2 | **0** 2 | **0** 2 | **0** 2 | 2 | 2 | 2 | 2 | 2 | **0** 2 | **0** 2 | **0** |
| RYR | ACA | 1 | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | **0** | **0** | **0** | **0** | 1 | 1 |
| | ACG | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | ATA | **0** | 1 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | ATG | **0** | **0** | 1 | **0** | 1 | 1 | 1 | **0** | 1 | 1 | 1 | 1 | 1 | **0** | 1 |
| | GCA | 1 | 2 | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | 1 | **0** |
| | GCG | **0** | 1 | **0** | **0** | 1 | 1 | 1 | **0** | **0** | **0** | **0** | **0** | 1 | 1 | 1 |
| | GTA | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | GTG | **0** | **0** | 1 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | 1 | 1 | **0** | 1 |
| | MEAN | **0** | 1 | **0** | **0** | **0** 1 | **0** 1 | **0** 1 | **0** | **0** | **0** | **0** | **0** | **0** 1 | **0** 1 | 1 |
| RYY | ACC | **0** | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | 2 | 2 | 2 | 2 | **0** | **0** |
| | ACT | 1 | 1 | **0** | 1 | 1 | **0** | 1 | 1 | 1 | 2 | 2 | **0** | **0** | 1 | **0** |
| | ATC | **0** | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | 2 | 2 | **0** | **0** |
| | ATT | **0** | 1 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | GCC | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | GCT | 2 | 2 | **0** | 2 | **0** | **0** | **0** | 2 | **0** | **0** | **0** | **0** | **0** | 2 | **0** |
| | GTC | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | GTT | **0** | 2 | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | MEAN | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| YRR | CAA | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 |
| | CAG | 2 | **0** | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |
| | CGA | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | CGG | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 |
| | TAA | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | TAG | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | TGA | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |
| | TGG | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 2 |
| | MEAN | 1 | 2 | 1 | 1 | 1 2 | 1 2 | 1 2 | 1 | 1 | 1 | 1 | 1 | 1 2 | 1 2 | 2 |
| YRY | CAC | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 |
| | CAT | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| | CGC | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | CGT | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | TAC | **0** | **0** | 2 | **0** | **0** | 2 | **0** | **0** | **0** | 1 | 1 | 2 | 2 | **0** | 2 |
| | TAT | 2 | **0** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | TGC | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |
| | TGT | 2 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | MEAN | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

Table 6 (*continued*)

| | | AG, AP | HB | MC | MP | MSA | MSM | MT | PB | PCA | PCF | PCH | SLS | SLT | TPA | TPV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| YYR | CCA | 1 | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | **0** | **0** | **0** | **0** | 1 | 1 |
| | CCG | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | CTA | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | **0** | **0** | 1 | 1 | 2 | 1 |
| | CTG | **0** | **0** | 1 | **0** | 1 | 1 | 1 | **0** | 1 | 1 | 1 | 1 | 1 | **0** | 1 |
| | TCA | 1 | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | **0** | **0** | 1 | 1 |
| | TCG | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | TTA | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | TTG | 1 | **0** | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | MEAN | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| YYY | CCC | — | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| | CCT | 2 | 2 | **0** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | **0** | **0** | 2 | 2 |
| | CTC | **0** | **0** | 1 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | 1 | 1 | **0** | **0** |
| | CTT | 2 | 2 | 1 | 2 | **0** | **0** | 2 | 2 | 2 | **0** | 2 | 1 | 1 | 2 | **0** |
| | TCC | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 |
| | TCT | 1 | 1 | **0** | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | **0** | **0** | 1 | 2 |
| | TTC | **0** | **0** | 2 | **0** | 1 | 1 | **0** | **0** | **0** | 1 | **0** | 2 | 2 | **0** | 1 |
| | TTT | — | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| | MEAN | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 | **0** 1 2 |

different $\mathscr{C}^3$ codes, and in particular of the trinucleotide diversity.

### 3.6. Types of nucleotides in the trinucleotide sites of the archaeal circular codes

The 4 types of nucleotides in the 15 archaeal $\mathscr{C}^3$ codes $\mathscr{X}_0(\mathscr{G})$ (in reading frame) occur in:

- the first trinucleotide site except T with the 3 genomes $\mathscr{G} = \{\text{MSM}, \text{PCF}, \text{TPV}\}$,
- the second trinucleotide site except G with the 5 genomes $\mathscr{G} = \{\text{MP}, \text{PB}, \text{PCA}, \text{PCF}, \text{PCH}\}$,
- the third trinucleotide site except A with the genome $\mathscr{G} = \text{HB}$ and G with the genome $\mathscr{G} = \text{SLT}$.

Therefore, only one type of nucleotide T (resp. G) may not occur in the first (resp. second) site, but 2 types of nucleotides A and G may not occur in the third site. In other words, G may not occur in 2 sites, the second and the third ones, and C always occurs in the 3 sites. The code $\mathscr{X}_0$(PCF) has no trinucleotides T and G in the first and second sites, respectively. It codes for the smallest number of amino acids among all the archaeal codes (see Discussion).

Similar rules can be deduced obviously with the 15 archaeal codes $\mathscr{X}_1(\mathscr{G})$ and $\mathscr{X}_2(\mathscr{G})$ by permutation.

## 4. Discussion

### 4.1. Common codons in the archaeal circular codes

The common and rare codons in the 15 archaeal $\mathscr{C}^3$ codes $\mathscr{X}_0(\mathscr{G})$ (in reading frame) are (from Table 3):

- 20 codons are absent (occurrence number $Nb = 0$ of codons in the 15 codes)
- 11 codons are very rare ($1 \leqslant Nb \leqslant 3$)
- 9 codons are rare ($4 \leqslant Nb \leqslant 7$)
- 10 codons are common ($8 \leqslant Nb \leqslant 11$): AAC, ACC, GAA, GAT, GCA, GCG, GCT, GTG, TAC, TTC
- 7 codons are very common ($12 \leqslant Nb \leqslant 14$): ATA, ATC, ATT, CTC, GAG, GTA, GTT
- 3 codons are always present ($Nb = 15$): GAC, GCC, GTC which code for Asp, Ala and Val, respectively.

### 4.2. Consequences on the 2-letter alphabets of the archaeal genomes

The 3 circular codes $\mathscr{X}_0(\mathscr{G})$, $\mathscr{X}_1(\mathscr{G})$ and $\mathscr{X}_2(\mathscr{G})$ in an archaeal genome $\mathscr{G}$ have 20 trinucleotides in the frames 0, 1 and 2, respectively. Therefore, a preferential frame for the 8 2-letter trinucleotides on a given 2-letter alphabet can be deduced by considering for each 2-letter trinucleotide, the average frame associated with the 8 4-letter trinucleotides specified on the 2-letter trinucleotide and belonging to the code $\mathscr{X}_0(\mathscr{G})$ (frame 0), $\mathscr{X}_1(\mathscr{G})$ (frame 1) and $\mathscr{X}_2(\mathscr{G})$ (frame 2).

On the alphabet {R,Y} (R = purine = {A,G}, Y = pyrimidine = {C,T}), the trinucleotide RYY occurs in frame 0 with all the 15 archaeal codes (Table 6). With the 2 genomes $\mathscr{G} = \{\text{MSM}, \text{TPV}\}$, all the 8 4-letter trinucleotides specified on RYY are in frame 0. The pattern RYY is also deduced from the circular code of mitochondria (Arquès and Michel, 1997). Its permutated trinucleotides YYR and YRY occur obviously in frames 1 and 2, respectively (Table 6). The trinucleotide RYR also occurs in frame 0 with 13 archaeal codes and its permutated trinucleotides YRR and RRY in frames 1

Table 7
Amino acids coded by the 15 archaeal circular codes $\mathscr{X}_0(\mathscr{G})$, $\mathscr{X}_1(\mathscr{G})$ and $\mathscr{X}_2(\mathscr{G})$ according to the universal genetic code. The numbers 0 (in bold), 1 and 2 are associated with the frames 0 (reading frame), 1 and 2 of the codes $\mathscr{X}_0(\mathscr{G})$, $\mathscr{X}_1(\mathscr{G})$ and $\mathscr{X}_2(\mathscr{G})$, respectively

| | | AG, AP | HB | MC | MP | MSA | MSM | MT | PB | PCA | PCF | PCH | SLS | SLT | TPA | TPV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lysine, Lys, K | AAG | **0** | **0** | 1 | **0** | 1 | 1 | **0** | 1 | **0** | 1 | 1 | 1 | 1 | **0** | 1 |
| Methionine, Met, M | ATG | **0** | **0** | 1 | **0** | 1 | 1 | 1 | **0** | 1 | 1 | 1 | 1 | 1 | **0** | 1 |
| Phenylalanine, Phe, F | TTC | **0** | **0** | 2 | **0** | 1 | 1 | **0** | **0** | **0** | 1 | **0** | 2 | 2 | **0** | 1 |
| Tryptophan, Trp, W | TGG | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 2 |
| Asparagine, Asn, N | AAC | **0** | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | 2 | 2 | 2 | 2 | **0** | **0** |
| | AAT | 2 | **0** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Aspartic acid, Asp, D | GAC | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | GAT | 2 | 2 | **0** | 2 | **0** | **0** | **0** | 2 | **0** | **0** | **0** | **0** | **0** | 2 | **0** |
| Cysteine, Cys, C | TGC | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |
| | TGT | 2 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Glutamic acid, Glu, E | GAA | 2 | 2 | **0** | 2 | **0** | **0** | 2 | **0** | 2 | **0** | **0** | **0** | **0** | 2 | **0** |
| | GAG | **0** | **0** | 1 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | 1 | **0** | **0** |
| Glutamine, Gln, Q | CAA | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 |
| | CAG | 2 | **0** | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |
| Histidine, His, H | CAC | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 |
| | CAT | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| Tyrosine, Tyr, Y | TAC | **0** | **0** | 2 | **0** | **0** | 2 | **0** | **0** | **0** | 1 | 1 | 2 | 2 | **0** | 2 |
| | TAT | 2 | **0** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Glycine, Gly, G | GGA | 2 | 2 | **0** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | **0** | 2 | 2 |
| | GGC | 2 | **0** | 2 | 2 | **0** | **0** | **0** | 2 | 2 | 2 | 2 | 2 | **0** | **0** | **0** |
| | GGT | 2 | 2 | **0** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | **0** | **0** | 2 | **0** |
| Isoleucine, Ile, I | ATA | **0** | 1 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | ATC | **0** | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | 2 | 2 | **0** | **0** |
| | ATT | **0** | 1 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| Proline, Pro, P | CCA | 1 | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | **0** | **0** | **0** | **0** | 1 | 1 |
| | CCG | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | CCT | 2 | 2 | **0** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | **0** | **0** | 2 | 2 |
| Alanine, Ala, A | GCA | 1 | 2 | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | 1 | **0** |
| | GCC | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | GCG | **0** | 1 | **0** | **0** | 1 | 1 | 1 | **0** | **0** | **0** | **0** | **0** | 1 | 1 | 1 |
| | GCT | 2 | 2 | **0** | 2 | **0** | **0** | **0** | 2 | **0** | **0** | **0** | **0** | **0** | 2 | **0** |
| Threonine, Thr, T | ACA | 1 | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | **0** | **0** | **0** | **0** | 1 | 1 |
| | ACC | **0** | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | 2 | 2 | 2 | 2 | **0** | **0** |
| | ACG | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | ACT | 1 | 1 | **0** | 1 | 1 | **0** | 1 | 1 | 1 | 2 | 2 | **0** | **0** | 1 | **0** |
| Valine, Val, V | GTA | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | GTC | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | GTG | **0** | **0** | 1 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | 1 | 1 | **0** | 1 |
| | GTT | **0** | 2 | **0** | 2 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| Arginine, Arg, R | AGA | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 2 |
| | AGG | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
| | CGA | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | CGC | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | CGG | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 |
| | CGT | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Leucine, Leu, L | CTA | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | **0** | **0** | 1 | 1 | 2 | 1 |
| | CTC | **0** | **0** | 1 | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | 1 | 1 | **0** | **0** |
| | CTG | **0** | **0** | 1 | **0** | 1 | 1 | 1 | **0** | 1 | 1 | 1 | 1 | 1 | **0** | 1 |
| | CTT | 2 | 2 | 1 | 2 | **0** | **0** | 2 | 2 | 2 | **0** | 2 | 1 | 1 | 2 | **0** |
| | TTA | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | TTG | 1 | **0** | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Serine, Ser, S | AGC | **0** | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | **0** | 2 |
| | AGT | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | TCA | 1 | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | **0** | **0** | 1 | 1 |
| | TCC | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 |
| | TCG | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | TCT | 1 | 1 | **0** | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | **0** | **0** | 1 | 2 |
| Stop codon, ochre | TAA | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Stop codon, amber | TAG | 1 | **0** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Stop codon, opal | TGA | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |

Table 8
Classification of the archaeal genomes according to the number of amino acids coded by the 15 $\mathscr{C}^3$ codes $\mathscr{X}_0(\mathscr{G})$ (in reading frame) (deduced from Table 7). The top part gives the 6 amino acids always coded by these codes

| | PCF | PCH | MC, SLS, SLT | MSM, TPV | MSA | PB | PCA | MT | AG, AP | MP | HB | TPA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aspartic acid, Asp, D | + | + | + | + | + | + | + | + | + | + | + | + |
| Glutamic acid, Glu, E | + | + | + | + | + | + | + | + | + | + | + | + |
| Isoleucine, Ile, I | + | + | + | + | + | + | + | + | + | + | + | + |
| Alanine, Ala, A | + | + | + | + | + | + | + | + | + | + | + | + |
| Threonine, Thr, T | + | + | + | + | + | + | + | + | + | + | + | + |
| Valine, Val, V | + | + | + | + | + | + | + | + | + | + | + | + |
| Lysine, Lys, K | – | – | – | – | – | – | + | + | + | + | + | + |
| Methionine, Met, M | – | – | – | – | – | + | – | – | + | + | + | + |
| Phenylalanine, Phe, F | – | + | – | – | – | + | + | + | + | + | + | + |
| Asparagine, Asn, N | – | – | – | + | + | + | + | + | + | + | + | + |
| Glutamine, Gln, Q | – | – | – | – | – | – | – | – | – | + | + | – |
| Tyrosine, Tyr, Y | – | – | – | – | + | + | + | + | + | + | + | + |
| Glycine, Gly, G | – | – | + | + | + | – | – | + | – | – | + | + |
| Proline, Pro, P | + | + | + | – | – | – | – | – | – | – | – | – |
| Leucine, Leu, L | + | + | – | + | + | + | + | + | + | + | + | + |
| Serine, Ser, S | – | – | + | – | – | – | – | – | + | – | – | + |
| Total number of AA | 8 | 9 | 9 | 9 | 10 | 11 | 11 | 12 | 13 | 13 | 14 | 14 |

and 2, respectively (Table 6). There is no preferential frame for RRR and YYY. Therefore, the archaeal genes follow the pattern RYN = {RYR,RYY} (N = {R,Y}) contrary to the classical one RNY deduced from the circular code of eukaryotes/prokaryotes (Arquès and Michel, 1996; Eigen and Schuster, 1978). Otherwise, the pattern RYN is a maximal circular code on the alphabet {R,Y} but not a self-complementary one.

This analysis on the 2 other 2-letter alphabets {K,M} (K = ceto = {G,T}, M = amino = {A,C}) and {S,W} (S = strong interaction = {C,G}, W = weak interaction = {A,T}) does not reveal a preferential 2-letter pattern as the 4-letter trinucleotides in frame 0 are spread on several 2-letter trinucleotides (data not shown). Therefore, in archaeal genes, the 2-letter alphabet close to the alphabet {A,C,G,T} is {R,Y}.

### 4.3. Consequences on the genetic code of the archaeal genomes

Four amino acids (AA) are never coded by the 15 archaeal $\mathscr{C}^3$ codes $\mathscr{X}_0(\mathscr{G})$ (in reading frame): Arg, Cys, His and Trp (Table 7). These 4 AA have a complex chemical structure in terms of their numbers of atoms or cycles. Indeed, Trp is the single AA with 2 cycles. Arg and His are the 2 most complex positively charged (basic) polar AA: Arg has the greatest number of atoms in the side chain (without cycle) and His has a cycle which leads, under certain conditions, to a dual acid/base property allowing catalytic reactions in the active sites of enzymes. Cys can form disulfide linkages by reaction with another Cys.

Six amino acids are always coded by the 15 archaeal codes: Ala, Asp, Glu, Ile, Thr and Val (Table 7). Ala and Asp represent the complete group of negatively charged (acidic) polar AA. These 6 AA are equally represented in the 2 classes of aminoacyl-tRNA synthetases with a class I containing Glu, Ile and Val, and a class II holding Ala, Asp and Thr (reviewed in Schimmel et al., 1993; Hartman, 1995; Saks and Sampson, 1995). Leu close to Ile is coded by 12 archaeal codes.

These archaeal codes code for a number of AA varying from 8 AA with $\mathscr{G}$ = PCF to 14 AA with $\mathscr{G}$ = {HB, TPA} (Table 8). Circular codes coding a small number of AA may be more "primitive". Pro, which is based on a ring structure that includes the central carbon atom, is only coded by the primitive archaeal codes (number of AA $\leqslant 9$ with $\mathscr{G} = \{\text{PCF}, \text{PCH}, \text{MC}, \text{SLS}, \text{SLT}\}$). Two AA, Gly and Ser, are coded both by primitive and evolved circular codes. Lys is the single basic AA encoded by archaeal codes, and furthermore, only by the evolved ones (number of AA $\geqslant 11$ with $\mathscr{G} = \{\text{PCA}, \text{MT}, \text{AG}, \text{AP}, \text{MP}, \text{HB}, \text{TPA}\}$). Met, Phe, Asn, Gln and Tyr are mainly coded by the evolved archaeal codes.

## Acknowledgements

We thank the referee for his advice.

## References

Andersson, S.G.E., Kurland, C.G., 1990. Codon preferences in free living microorganisms. Microbiol. Rev. 54, 198–210.

Antezana, M.A., Kreitman, M., 1999. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J. Mol. Evol. 49, 36–43.

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., Michel, C.J., 1997. A circular code in the protein coding genes of mitochondria. J. theor. Biol. 189, 273–290.

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.

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 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.

Eigen, M., Schuster, P., 1978. The hypercycle. A principle of natural self-organization. Part C: the realistic hypercycle. Naturwissenschaften 65, 341–369.

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.

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.

Gu, W.J., Ma, J.M., Zhou, T., Sun, X., Lu, Z.H., 2002. Studies on the codon usage bias of genes coding for proteins with different tertiary structure. Acta Biophys. Sin. 18, 81–86.

Gutman, G.A., Hatfield, G.W., 1989. Nonrandom utilization of codons pairs in *Escherichia coli*. Proc. Natl Acad. Sci. USA 86, 3699–3703.

Hanai, R., Wada, A., 1988. The effects of guanine and cytosine variation on dinucleotide frequency and amino acid composition in the human genome. J. Mol. Evol. 27, 321–325.

Hartman, H., 1995. Speculations on the origin of the genetic code. J. Mol. Evol. 40, 541–544.

Ikemura, T., 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. 151, 389–409.

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.

Koch, A.J., Lehmann, J., 1997. About a symmetry of the genetic code. J. theor. Biol. 189, 171–174.

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

Ma, J., Zhou, T., Gu, W., Sun, X., Lu, Z., 2002. Cluster analysis of the codon use frequency of MHC genes from different species. BioSystems 65, 199–207.

Ng, W.V., et al., 2000. Genome sequence of *Halobacterium* species NRC-1. Proc. Natl Acad. Sci. USA 97, 12176–12181.

Saks, M.E., Sampson, J.R., 1995. Evolution of tRNA recognition systems and tRNA gene sequences. J. Mol. Evol. 40, 509–518.

Schimmel, P., Giegé, R., Moras, D., Yokoyama, S., 1993. An operational RNA code for amino acids and possible relationship to genetic code. Proc. Natl Acad. Sci. USA 90, 8763–8768.

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.

Smith, T.F., Ralph, W.W., Goodman, M., Czelusniak, J., 1985. Codon usage in the vertebrate haemoglobins and its implications. Mol. Biol. Evol. 2, 390–398.

Sueoka, N., 1992. Directional mutation pressure, selection constraints, and genetic equilibria. J. Mol. Evol. 34, 95–114.

Tazi, J., Bird, A., 1990. Alternative chromatin structure at CpG islands. Cell 60, 909–920.

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

Yarus, M., Folley, L.S., 1984. Sense codons are found in specific contexts. J. Mol. Biol. 182, 529–540.