Bulletin of Mathematical Biology (2023) 85:24
https://doi.org/10.1007/s11538-023-01129-4

ORIGINAL ARTICLE

Reading Frame Retrieval of Genes: A New Parameter of Codon Usage Based on the Circular Code Theory

Christian J. Michel[1] · Jean-Sébastien Sereni[1]

Received: 25 September 2022 / Accepted: 26 January 2023 / Published online: 24 February 2023

Abstract

Based on the circular code theory, we define a new function f that quantifies the property of reading frame retrieval (RFR) of genes from their codon usage. This RFR function f is computed on a massive scale in genes of genomes of bacteria, eukaryotes and archaea. By expressing f as a function of the mean number $\overline{n}$ of codons per gene, a "universal" property is identified, whatever the kingdom: the reading frame retrieval is enhanced in large genes. By investigating this property according to the theory developed, a Spearman's rank correlation with a strong negative coefficient is observed between the codon usage dispersion d (from the uniform codon distribution $\frac{1}{64}$) and the RFR function f, whatever the kingdom (p-values $< 10^{-180}$ in bacteria, $< 10^{-61}$ in eukaryotes and $< 10^{-159}$ in archaea). Thus, the reading frame retrieval is enhanced with the codon usage dispersion. Furthermore, this approach identifies a "genome centre" from which emerge two distinct "genome arms": an upper arm and a lower arm, respectively, above and below the linear regression. The RFR function by itself or combined with classical methods (alignment, phylogeny) could also be a new approach to classify the genomes in the future.

Keywords Circular code theory · Reading frame retrieval · Codon usage · Gene length · Codon dispersion

1 Introduction

The genetic code is a surjective map between the 64 trinucleotides and the 20 amino acids plus 1 stop signal. Thus, there are codons, called synonymous, coding the same

✉ Christian J. Michel
c.michel@unistra.fr

Jean-Sébastien Sereni
jean-sebastien.sereni@cnrs.fr

[1] Theoretical Bioinformatics, ICube, C.N.R.S., University of Strasbourg, 300 Boulevard Sébastien Brant, 67400 Illkirch, France

amino acid, except methionine and tryptophan that are encoded by a single codon, ATG and TGG, respectively. Synonymous codons have different frequencies between genomes of different species, as well as between genes within a given genome. This statistical property is known as codon usage bias (CUB). CUB influences different aspects of protein production (Grantham et al. 1981) and codon choice has effects at many biological stages, including transcription (Zhou et al. 2016), translation efficiency (Qian et al. 2012), mRNA stability (Presnyak et al. 2015), protein folding (Buhr et al. 2016) and protein function (Bali and Bebok 2015).

A great number of statistical parameters have been defined to analyse CUB, to mention a few: the Effective Number of Codons (ENC) (Wright 1990), the GC content (GC), the GC content at third codon positions (GC3), the Relative Synonymous Codon Usage (RSCU) (Sharp et al. 1986), the Codon Adaptation Index (CAI) (Sharp and Li 1987), the Frequency of Optimal Codons (FOC) (Ikemura 1985), the Relative Codon Bias (RCB) and the Relative Codon Bias Strength (RCBS) (Roymondal et al. 2009), the Relative Codon Adaptation (RCA) (Fox and Erill 2010) and the Codon Deviation Coefficient (CDC) (Zhang et al. 2012). Combination of two parameters allows the creation of 2D plots, for example: ENC plot to investigate codon usage across genes (He et al. 2019), neutrality plot to analyse the effects of natural selection and mutation pressure on codon usage (Yu et al. 2021) and parity rule 2 (PR2) plot to evaluate the effect of mutation pressure and natural selection at the third codon position of the four-codon amino acids (Yu et al. 2021). Factorial statistical methods, such as the Correspondence Analysis (CA) (Grantham et al. 1981) and the Principal Component Analysis (PCA) (Michel 1986) also allow the codon usage to be studied.

The circular code theory has been initiated in 1996 by the identification in genes of bacteria and eukaryotes, of a maximal C^3 self-complementary circular code, a particular set called X of 20 trinucleotides with interesting mathematical properties allowing to retrieve the reading frame and the two shifted frames (Arquès and Michel 1996). In 2017, it has been shown that this circular code X is also found in genes of archaea, plasmids and viruses (Michel 2017). The historical context of this result is described in a recent article (Michel 2020). We also refer the reader to the reviews (Michel 2008; Fimmel and Strüngmann 2018) for the biological context and the main combinatorial studies of circular codes.

This unexpected biological result has led to several mathematical developments since 1996: (i) the flower automaton (Arquès and Michel 1996); (ii) the necklaces LDN (letter diletter necklace) and DLN (diletter letter necklace) (Pirillo 2003; Michel et al. 2008a, b) extended to $(n+1)LDCCN$ (letter diletter continued closed necklaces) (Michel and Pirillo 2010); (iii) the group theory (Fimmel et al. 2015); and (iv) the recent and powerful approach based on graph theory in 2016 (Fimmel et al. 2016). The graph approach has recently led to two important generalisations: mixed circular codes (Fimmel et al. 2019) and k-circular codes (Fimmel et al. 2020; Michel et al. 2022; Michel and Sereni 2022).

These theoretical results have led to biological applications, to name a few recent ones: identification of "universal" circular code motifs in the ribosome leading to a model of genetic code evolution associating codes, translation systems, and peptide products at different stages from the primordial translation building blocks to the ancestor of the modern ribosome present in the Last Universal Common Ancestor

(LUCA) (Dila et al. 2019); identification of a circular code periodicity (modulo 3) in a large region of the 16 S rRNA including the 3' major domain corresponding to the primordial proto-ribosome decoding centre, containing numerous sites that interact with the tRNA and messenger RNA (mRNA) during translation and surrounding the mRNA channel (Michel and Thompson 2020); potential role of the circular code X in the regulation of gene expression (Thompson et al. 2021); and characterisation of accessory genes in coronavirus genomes using the circular code information (Michel et al. 2020).

On the genetic alphabet $\mathcal{B}$, there are $n = 2^{64} - 1 \approx 10^{19}$ (non-empty) trinucleotide codes: 64 codes of cardinality 1: $\{AAA\}, ..., \{TTT\}$, 2016 codes of cardinality 2: $\{AAA, AAC\}, ..., \{TTG, TTT\}$, 41664 codes of cardinality 3: $\{AAA, AAC, AAG\}, ..., \{TTC, TTG, TTT\}$ up to 1 code of cardinality 64 (the genetic code): $\{AAA, ..., TTT\}$. The recent theory of trinucleotide k-circular codes makes it possible to study the property of reading frame retrieval, called circularity property, of any of these $\approx 10^{19}$ codes (Michel et al. 2022; Michel and Sereni 2022). Indeed, these codes can be classified into 3 classes according to their circularity property:

(i) 15 trinucleotide codes with no circularity: no sequence generated by such a trinucleotide code can retrieve the reading frame;
(ii) $n - 15 - 115606988558$ trinucleotide codes with a partial circularity: some sequences generated by such a trinucleotide code cannot retrieve the reading frame, but some some other sequences can retrieve the reading frame;
(iii) 115606988558 trinucleotide codes with a complete circularity (circular codes): any sequence generated by such a trinucleotide circular code can retrieve the reading frame.

The property of reading frame retrieval of trinucleotide codes is analysed with a function defined by Michel and Sereni (2022, Definition 6.1) and recalled below with Definition 2.10. The theoretical work here extends this function to weighted trinucleotide codes with the new Definition 2.12. Such a quantitative parameter to retrieve the reading frame has never been proposed to date. Furthermore, we show here that this RFR function f can be applied in a biological context to the codon usage, of a gene or a set of genes. By computing this RFR function f on a massive scale in genes of genomes of bacteria, eukaryotes and archaea, new properties associated with reading frame retrieval are identified.

This article is organised as follows. The necessary definitions and notations of trinucleotide codes, circular codes and their generalisation to k-circular codes are gathered in Sect. 2.1. The definition of mean number of codons per gene is recalled in Sect. 2.2. Section 2.3 defines the dispersion function of codon usage and states a proposition about its interval. Section 2.4 defines the reading frame retrieval (RFR) function and states several propositions concerning its interval and its particular value 1 associated with a uniform codon usage. Section 2.5 describes the acquisition of codon usage for the genomes of bacteria, eukaryotes and archaea from the codon statistics database (CSD) (Subramanian et al. 2022). As an example, the value of the RFR function for *Homo sapiens* is given in Sect. 2.6.

The results are presented in three parts. Section 3.1 demonstrates that the reading frame retrieval is enhanced in large genes, in bacteria, eukaryotes and archaea. Section 3.2 shows that the reading frame retrieval is correlated with the dispersion of codon usage and identifies a "genome centre" from which emerge two distinct "genome arms", also in bacteria, eukaryotes and archaea. The results previously obtained as well as a study (done by accident) of codon usage in bird genomes suggest that the RFR function could be a new approach to classify genomes (Sect. 3.3).

2 Method

2.1 Definitions and Notations

For the reader's convenience we here recall the most relevant notions, in order to have this article self-contained. The theoretical aspects, with computer results, proofs, examples, remarks, illustrations and refinements are found in the articles (Michel et al. 2022; Michel and Sereni 2022).

We work with the *genetic alphabet* $\mathcal{B} := \{A, C, G, T\}$, which has cardinality 4. An element N of $\mathcal{B}$ is called *nucleotide*. A *word* over the genetic alphabet is a sequence of nucleotides. A *trinucleotide* is a sequence of 3 nucleotides, that is, using the standard word-theory notation, an element of $\mathcal{B}^3$. If $w = N_1 \cdots N_s$ and $w' = N'_1 \cdots N'_t$ are two sequences of nucleotides of respective lengths s and t, then the *concatenation* $w \cdot w'$ *of* w *and* w' is the sequence $N_1 \cdots N_s N'_1 \cdots N'_t$ composed of $s + t$ nucleotides.

Given a sequence $w = N_1 N_2 \cdots N_s \in \mathcal{B}^s$ and an integer $j \in \{0, 1, \ldots, s-1\}$, the *circular* j*-shift* of w is the word $N_{j+1} \cdots N_s N_1 \cdots N_j$. Note that the circular 0-shift of w is w itself. A sequence w' of nucleotides is a *circular shift* of w if w' is the circular j-shift of w for some $j \in \{0, 1, \ldots, s-1\}$. The elements in $\mathcal{B}^3$ can thus be partitioned into conjugacy classes, where the *conjugacy class* of a trinucleotide $w \in \mathcal{B}^3$ is the set of all circular shifts of w.

Definition 2.1 Let $\mathcal{B}$ be the genetic alphabet.

- A *trinucleotide code* is a subset of $\mathcal{B}^3$, that is, a set of trinucleotides.
- If X is a trinucleotide code and w is a sequence of nucleotides, then an X*-decomposition* of w is a tuple $(x_1, \ldots, x_t) \in X^t$ of trinucleotides from X such that $w = x_1 \cdot x_2 \cdots x_t$.

We now formally define the notion of circularity of a code, i.e. the property of reading frame retrieval.

Definition 2.2 Let $X \subseteq \mathcal{B}^3$ be a trinucleotide code.

- Let m be a positive integer and let $(x_1, \ldots, x_m) \in X^m$ be an m-tuple of trinucleotides from X. A *circular* X*-decomposition* of the concatenation $c := x_1 \cdots x_m$ is an X-decomposition of a circular shift of c.
- Let k be a non-negative integer. The code X is $(\geq k)$*-circular* if every concatenation of trinucleotides from X that admits more than one circular X-decomposition contains at least $k + 1$ trinucleotides. In other words, X is $(\geq k)$-circular if for

every $m \in \{1, \dots, k\}$ and each m-tuple $(x_1, \dots, x_m)$ of trinucleotides from X, the concatenation $x_1 \cdots x_m$ admits a unique circular X-decomposition.
The code X is *k-circular* if X is $(\geq k)$-circular and not $(\geq k+1)$-circular.

- The code X is *circular* if it is $(\geq k)$-circular for all $k \in \mathbf{N}$.

We recall the definition of the graph associated with a trinucleotide code (Fimmel et al. 2016).

Definition 2.3 Let $X \subseteq \mathcal{B}^3$ be a trinucleotide code. We define a graph $\mathcal{G}(X) = (V(X), E(X))$ with set of vertices $V(X)$ and set of arcs $E(X)$ as follows:

- $V(X) := \bigcup\limits_{N_1 N_2 N_3 \in X} \{N_1, N_3, N_1N_2, N_2N_3\}$; and
- $E(X) := \{N_1 \to N_2N_3 \; : \; N_1N_2N_3 \in X\} \cup \{N_1N_2 \to N_3 \; : \; N_1N_2N_3 \in X\}$.

The graph $\mathcal{G}(X)$ is the graph *associated* to X.

The *length* of a directed cycle in a graph $\mathcal{G}$ is the number of arcs of the cycle. We note that every arc of $\mathcal{G}(X)$ joins a nucleotide and a dinucleotide. Thus, the graph $\mathcal{G}(X)$ cannot contain a directed cycle of odd length. A theorem (Fimmel et al. 2020, Theorem 3.3) implies that a cycle in $\mathcal{G}(X)$, if any, must have length in $\{2, 4, 6, 8\}$ and, in particular, that a trinucleotide (≥ 4)-circular code must be circular. As noted in a previous article (Michel et al. 2022), it follows that all trinucleotide codes over $\mathcal{B}$ can be naturally partitioned into 5 classes using the following definition.

Definition 2.4 We define the *circularity* $\mathrm{cir}(X)$ of a non-empty trinucleotide code X to be the largest integer $k \in \{0, 1, 2, 3, 4\}$ such that X is $(\geq k)$-circular.

Thus, the possible values of $\mathrm{cir}(X)$ for a trinucleotide code X are 0, 1, 2, 3, 4, which determine the 5 classes.

Next we introduce two new functions, which turn out to be correlated. The first one deals with the dispersion of the codon usage, and the second one, which uses the graph, deals with the property of reading frame retrieval of genes. These two functions are also analysed as a function of the mean number of codons per gene in each genome.

2.2 Mean Number of Codons per Gene in a Genome

Definition 2.5 The mean number $\overline{n}$ of codons per gene in a genome is the total number of codons divided by the total number of genes in the genome.

2.3 Dispersion of Codon Usage

A codon usage is *uniform* if every codon has the same occurrence frequency. The following function measures the dispersion of codon usage with respect to the uniform one. We write X_g for the genetic code of cardinality 64 (maximal cardinality in $\mathcal{B}^3$).

Definition 2.6 Given any trinucleotide code X, a *weight function on* X is a function $\omega \colon X \to [0, 1]$ such that $\sum_{x \in X} \omega(x) = 1$.

Definition 2.7 A *weighted trinucleotide code* is a pair (X, ω) where X is a trinucleotide code and ω is a weight function on X.

We can now define the dispersion of codon usage.

Definition 2.8 For every weight function $\omega\colon X_g \to [0, 1]$, the *dispersion of codon usage in* (X_g, ω) is the function d given by

$$d((X_g, \omega)) = \sum_{x \in X_g} \left| \omega(x) - \frac{1}{64} \right|. \tag{2.1}$$

The next proposition gives the extremal values taken by the function d.

Proposition 2.9 *For every weight function* $\omega\colon X_g \to [0, 1]$, *we have*

$$0 \leq d((X_g, \omega)) \leq \frac{63}{32} \approx 1.97.$$

Moreover, $d((X_g, \omega)) = 0$ *if and only if* $\omega(x) = \frac{1}{64}$ *for each trinucleotide* $x \in X_g$. *The upper bound is attained if and only if there is a trinucleotide* $x \in X_g$ *such that* $\omega(x) = 1$ *(and hence* $\omega(x') = 0$ *if* $x' \neq x$*).*

Proof The argument is standard. Consider an arbitrary weight function $\omega\colon X_g \to [0, 1]$ on X_g and recall that $\sum_{x \in X_g} \omega(x) = 1$. We define

$$X^+ = \left\{ x \in X \ : \ \omega(x) > \frac{1}{64} \right\}$$

and

$$X^- = \left\{ x \in X \ : \ \omega(x) \leq \frac{1}{64} \right\}.$$

In particular, $|X^+| + |X^-| = |X_g| = 64$. Now,

$$\begin{aligned} d((X_g, \omega)) &= \sum_{x \in X^+} \left(\omega(x) - \frac{1}{64} \right) + \sum_{x \in X^-} \left(\frac{1}{64} - \omega(x) \right) \\ &= \frac{1}{64} \cdot (|X^-| - |X^+|) + \sum_{x \in X^+} \omega(x) - \sum_{x \in X^-} \omega(x) \\ &= \frac{1}{64} \cdot (64 - 2 \cdot |X^+|) + 2 \cdot \sum_{x \in X^+} \omega(x) - 1 \\ &= 2 \cdot \left(\sum_{x \in X^+} \omega(x) - \frac{|X^+|}{64} \right) \end{aligned}$$

where the third line uses that $|X^-| = 64 - |X^+|$ and $\sum_{x \in X^-} \omega(x) = 1 - \sum_{x \in X^+} \omega(x)$.

Now, if $|X^+| = 0$ then $\sum_{x \in X^+} \omega(x) = 0$, and hence $d((X_g, \omega)) = 0$ and moreover ω is constantly equal to $\frac{1}{64}$ (because $\omega(x) \leq \frac{1}{64}$ for every $x \in X = X^-$ and $\sum_{x \in X} \omega(x) = 1$).

If $|X^+| \neq 0$ then $|X^+| \geq 1$ and therefore $d((X_g, \omega)) \leq 2(1 - \frac{1}{64}) = \frac{63}{32}$, with equality if and only if $|X^+| = 1$ and $\sum_{x \in X^+} \omega(x) = 1$, which ends the proof.

2.4 Gene Reading Frame Retrieval (RFR) Function Associated with a Codon Usage

Theoretical considerations over trinucleotide codes led to the following definition (Michel and Sereni 2022, Definition 6.1) as a measure of the reading frame retrieval of genes. Indeed, the number and length of cycles in the graph are associated with ambiguous sequences that do not retrieve the reading frame. Short cycles are associated with short ambiguous sequences, i.e. the reading frame is lost quickly (after 1 trinucleotide), in contrast to long cycles where the ambiguous sequences are long, i.e. the reading frame is lost after several trinucleotides, up to 4 trinucleotides (see Michel et al. 2022; Michel and Sereni 2022 for details).

Definition 2.10 (Michel and Sereni 2022, Definition 6.1) The *reading frame loss* function f of a trinucleotide code X is the mapping $f : \mathcal{B}^3 \to \mathbf{R}$ given by

$$f(X) := \mathrm{q}_8(\mathcal{G}(X)) + \frac{4}{3}\,\mathrm{q}_6(\mathcal{G}(X)) + 2\,\mathrm{q}_4(\mathcal{G}(X)) + 4\,\mathrm{q}_2(\mathcal{G}(X)) = \sum_{i=1}^{4} \frac{4}{i} \cdot \mathrm{q}_{2 \cdot i}(\mathcal{G}(X)) \tag{2.2}$$

where $\mathrm{q}_i(\mathcal{G})$ is the number of directed cycles of length i in the graph $\mathcal{G}$ for every positive integer i.

Note that $f(X)$ is always a rational number, but not necessarily an integer. The next proposition (Michel and Sereni 2022) gives the minimum and maximum values taken by f over all trinucleotide codes.

Proposition 2.11 (Michel and Sereni 2022, Proposition 6.2) *For every trinucleotide code X, we have $0 \leq f(X) \leq 301056$. Moreover, $f(X) = 0$ if and only if X is a trinucleotide circular code, and $f(X) = 301056$ if and only if X is the genetic code X_g, where*

$$\mathrm{q}_2(X_g) = 64, \qquad \mathrm{q}_4(X_g) = 1440, \qquad \mathrm{q}_6(X_g) = 26880, \qquad \mathrm{q}_8(X_g) = 262080.$$

We generalise the function f to the codon usage, where each trinucleotide x has occurrence frequency $w(x)$.

In view of Definition 2.3, one naturally associates with each weighted code the following weighted graph.

Definition 2.12 Let (X, ω) be a weighted trinucleotide code. The *weighted graph associated with* ω is the pair $(\mathcal{G}(X), \omega')$ where $\mathcal{G}(X)$ is given by Definition 2.3 with respect to X, and ω' is a function assigning to each of the two arcs of $\mathcal{G}(X)$ coming from a trinucleotide $N_1N_2N_3 \in X$ the rational number $\frac{\omega(N_1N_2N_3)}{2} \in [0, 1]$.

In other words, the arcs of the weighted graph $(\mathcal{G}(X), \omega')$ can be written as follows:

$$\left\{N_1 \xrightarrow{\omega(x)/2} N_2N_3 \,:\, x = N_1N_2N_3 \in X\right\} \cup \left\{N_1N_2 \xrightarrow{\omega(x)/2} N_3 \,:\, x = N_1N_2N_3 \in X\right\}.$$

We are now in a position to define the generalised function f associated with every weighted trinucleotide code.

Definition 2.13 Let (X, ω) be a weighted trinucleotide code and $(\mathcal{G}(X), \omega')$ its associated weighted graph. Let $\mathcal{C}$ be the set of all directed cycles of $\mathcal{G}(X)$. The *loss of reading frame retrieval (RFR)* function f of a (X, ω) is the mapping f given by

$$f((X, \omega)) := \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} (2|X|)^{|c|} \prod_{a \in E(c)} \omega'(a) \tag{2.3}$$

where $E(c)$ is the set of arcs of the directed cycle c.

Proposition 2.14 *(Uniform codon usage) Let X_g be the genetic code and let ω the uniform distribution over X_g, that is, $\omega \colon X_g \to [0, 1]$ is constant and equal to $\frac{1}{64}$. Then $f((X_g, \omega)) = 1$.*

Proof We have

$$\begin{aligned} f((X_g, \omega)) &= \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} (2 \cdot 64)^{|c|} \prod_{a \in E(c)} \frac{1}{2 \cdot 64} \\ &= \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} 1 \\ &= 1. \end{aligned}$$

The next proposition implies that for circular codes, the weight function ω has no significance for f, in the sense that all distributions yield the same value as the uniform one, namely 0. More precisely, the equivalence given by Proposition 2.11 between circular codes and f-value 0 generalises to weighted trinucleotide codes.

Proposition 2.15 *(Circular code) Let (X, ω) be a weighted trinucleotide code. Then $f((X, \omega)) = 0$ if and only if X is a circular code.*

The function f seems to be maximised by codes obtained from a circular code of maximal size (20) by adding a periodic trinucleotide x (i.e. AAA, CCC, GGG or TTT), with a weight function tending to 1 on x and 0 on all other trinucleotides. The determination of the theoretical maximal values of f is beyond the scope of this article, but we at least make the following observation, obtained by computing f for the aforementioned family of codes.

Proposition 2.16 *We have*

$$\sup\{f(X,\omega) : (X,\omega) \textit{ weighted trinucleotide code}\} \geq 441.$$

That is, for every $\varepsilon > 0$*, there exists a weighted trinucleotide code* (X,ω) *such that* $f(X,\omega) > 441 - \varepsilon$.

Definition 2.12 can be generalised to codes with words of arbitrary length ℓ on any finite alphabet Σ as a graph can be associated with any ℓ-letter code $X \subseteq \Sigma^\ell$ (Fimmel et al. 2016).

Definition 2.17 Fix $\ell \in \mathbf{N}$ and let $X \subseteq \Sigma^\ell$ be an ℓ-letter code. We define a directed graph $\mathcal{G}(X) = (V(X), E(X))$ with vertex set $V(X)$ and edge set $E(X)$ as follows.

(1) $V(X) := \bigcup_{i=1}^{\ell-1} \{N_1 \cdots N_i, N_{i+1} \cdots N_\ell \ : \ N_1 \cdots N_\ell \in X\}$; and
(2) $E(X) := \bigcup_{i=1}^{\ell-1} \{N_1 \cdots N_i \to N_{i+1} \cdots N_\ell \ : \ N_1 \cdots N_\ell \in X\}$.

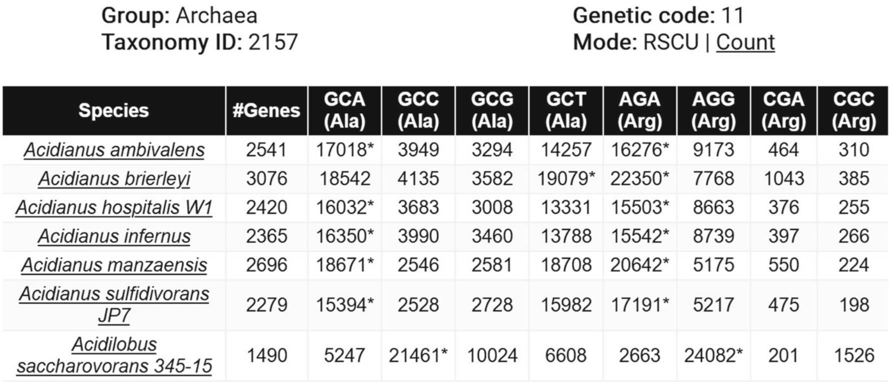
The weight function over the arcs then generalises naturally.

Definition 2.18 Fix an integer $\ell \geq 2$. Let $X \subseteq \Sigma^\ell$ and let $\omega\colon X \to [0,1]$ be a weight function over X. For each element $N_1 \ldots N_\ell \in X$ and each $i \in \{1, \ldots, \ell - 1\}$, we set

$$w'(N_1 \cdots N_i \to N_{i+1} \cdots N_\ell) = \frac{\omega(x)}{\ell - 1}.$$

2.5 Data

A very interesting codon statistics database (CSD) has recently been developed by the Alvarez-Ponce group (Subramanian et al. 2022) (Fig. 1). It provides the codon usage for all the species with reference or representative genomes in RefSeq. It is free to access without registration at http://codonstatsdb.unr.edu. From this CSD, we extract



Group: Archaea
Taxonomy ID: 2157

Genetic code: 11
Mode: RSCU | Count

Species	#Genes	GCA (Ala)	GCC (Ala)	GCG (Ala)	GCT (Ala)	AGA (Arg)	AGG (Arg)	CGA (Arg)	CGC (Arg)
Acidianus ambivalens	2541	17018*	3949	3294	14257	16276*	9173	464	310
Acidianus brierleyi	3076	18542	4135	3582	19079*	22350*	7768	1043	385
Acidianus hospitalis W1	2420	16032*	3683	3008	13331	15503*	8663	376	255
Acidianus infernus	2365	16350*	3990	3460	13788	15542*	8739	397	266
Acidianus manzaensis	2696	18671*	2546	2581	18708	20642*	5175	550	224
Acidianus sulfidivorans JP7	2279	15394*	2528	2728	15982	17191*	5217	475	198
Acidilobus saccharovorans 345-15	1490	5247	21461*	10024	6608	2663	24082*	201	1526

Fig. 1 Partial screen shot of the codon statistics database (CSD) (Subramanian et al. 2022) showing the initial data of codon usage (Color figure online)

Table 1 Basic statistics of genomes in the three kingdoms studied with the total numbers of genomes, genes and codons

Kingdom	Total number		
	Genomes	Genes	Codons
Bacteria	8345	34020997	11087876805
Eukaryota	1150	20206058	10374305634
Archaea	432	1280890	367937932

(July 2022) the codon usage of genomes of three kingdoms: bacteria, eukaryota and archaea.

The Archaea (Id 2157) can be directly extracted. The Bacteria and Eukaryota cannot be directly obtained from CSD, which is restricted to taxa for which the genetic code is homogeneous, i.e. species with the same genetic code. For example, the bacterial Mycoplasmatales uses a different genetic code with only two stop codons TAA and TAG, TGA coding Trp. Thus, the Bacteria are constructed from the union of the 22 following bacterial classes: Acidobacteria (Id 57723), Actinobacteria (Id 201174), Aquificae (Id 187857), Bacteroidetes (Id 976), Balneolia (Id 1853221), Chlamydiia (Id 204429), Chloroflexi (Id 200795), Cyanobacteria (Id 1117), Deferribacteres (Id 68337), Deinococcus-Thermus (Id 1297), Epsilonproteobacteria (Id 29547), Firmicutes (Id 1239), Fusobacteria (Id 32066), Mycoplasmatales (Id 2085), Nitrospirae (Id 40117), Planctomycetes (Id 203682), Pseudomonadales (Id 72274), Spirochaetes (Id 203691), Synergistetes (Id 508458), Thermodesulfobacteria (Id 200940), Thermotogae (Id 200918) and Verrucomicrobia (Id 74201). In a similar way, the Eukaryota are constructed from the union of the 4 following eukaryotic classes: Metazoa (animals, Id 33208), Plants (Embryophyta, land plants, Id 3193; Chlorophyta, green algae, Id 3041; Rhodophyta, red algae, Id 2763), Fungi (Agaricomycotina Id 5302; Pezizomycotina, Id 147538; Saccharomyces, Id 4930; Ustilaginomycotina, Id 452284; Basidiomycota, Id 5204) and Protists (Apicomplexa, Id 5794; Kinetoplastea, Id 5653). The few exceptional genomes in which the codon usage of the stop codons is not given, are not considered. Table 1 gives the basic statistics of genomes in the three kingdoms studied.

2.6 Gene Reading Frame Retrieval (RFR) of *Homo sapiens*

The codon usage of *Homo sapiens* obtained from the codon statistics database (CSD) (Subramanian et al. 2022) (taxonomy Id: 9606, 19850 genes, 11577026 codons) is given in Appendix Table 3. The gene reading frame retrieval (RFR) function of *H. sapiens* is equal to $f(H.\ sapiens) = 0.792$. This example also allows the reader to easily verify the computation of the RFR function f which does not pose any algorithmic difficulties.

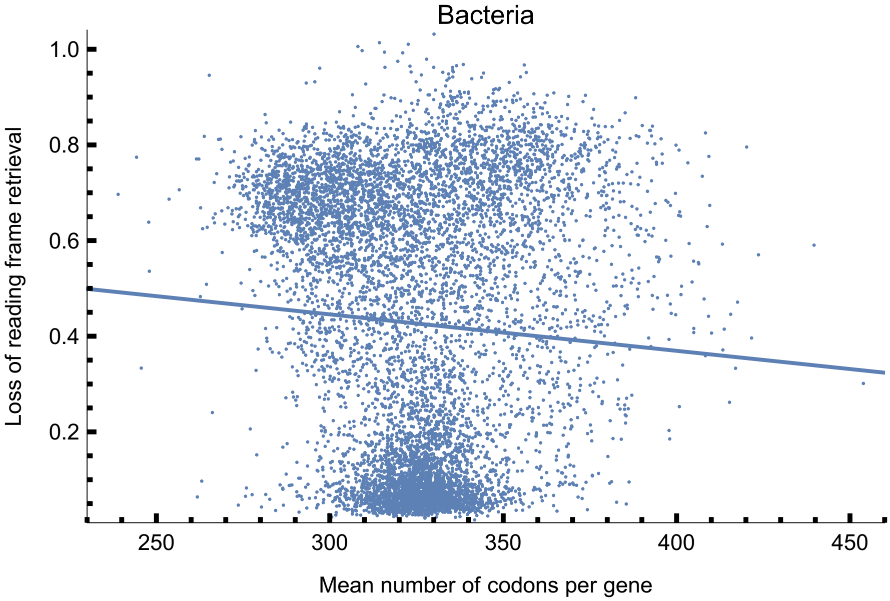


Fig. 2 Loss of reading frame retrieval in genes of 8345 bacterial genomes. Each point represents all the genes of a bacterial genome. The x-axis shows the mean number $\overline{n}$ (Definition 2.5) of codons per gene, i.e. the total number of codons in a genome divided by its total number of genes. The y-axis shows the reading frame retrieval function f (2.3). The linear regression between $\overline{n}$ and f decreases according to the equation $y = -0.000762074x + 0.674293$, with a Spearman's rank correlation coefficient $\rho = -0.07$ and p-value $< 10^{-9}$ (Color figure online)

3 Results

3.1 Reading Frame Retrieval Enhanced in Large Genes

We compute the gene reading frame retrieval (RFR) function f (2.3) as a function of the mean number $\overline{n}$ (Definition 2.5) of codons per gene in the genomes of three kingdoms. A "universal" property is identified: the larger the gene, the lower the reading frame loss, whatever the kingdom. We detail the statistical results for each kingdom.

Figure 2 shows that the mean numbers $\overline{n}$ of codons per gene of the 8345 bacterial genomes all belong to the interval $[230, 460]$. The overall mean number $\overline{\overline{n}}$ of codons per gene, over all these bacterial genomes, is 325 (see Table 2). The RFR function $f(Bacteria)$ decreases according to a Spearman's rank correlation coefficient $\rho = -0.07$ with a very strong significant p-value $< 10^{-9}$.

This property is retrieved in the genes of eukaryotic and archaeal genomes. As their genome numbers are significantly smaller than the bacterial one, the statistical significance of their ρ values is obviously smaller compared to the ρ value of bacterial genomes.

Figure 3 shows that the mean numbers $\overline{n}$ of codons per gene of the 1150 eukaryotic genomes all belong to the interval $[260, 860]$. The overall mean number $\overline{\overline{n}}$ of

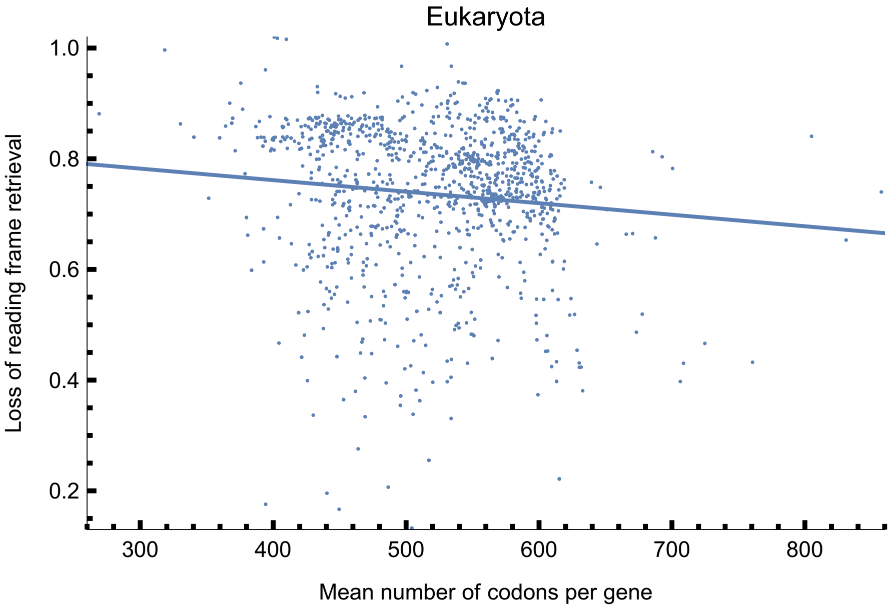


Fig. 3 Loss of reading frame retrieval in genes of 1150 eukaryotic genomes. Each point represents all the genes of an eukaryotic genome. The x-axis shows the mean number $\overline{n}$ (Definition 2.5) of codons per gene, i.e. the total number of codons in a genome divided by its total number of genes. The y-axis shows the reading frame retrieval function f (2.3). The linear regression between $\overline{n}$ and f decreases according to the equation $y = -0.00020828x + 0.844653$, with a Spearman's rank correlation coefficient $\rho = -0.14$ and p-value $< 10^{-6}$ (Color figure online)

codons per gene, over all these eukaryotic genomes, is 519 (see Table 2). Thus, in average, the eukaryotic genes are significantly longer than the bacterial genes. The RFR function $f(Eukaryota)$ decreases according to a Spearman's rank correlation coefficient $\rho = -0.14$ with a strong significant p-value $< 10^{-6}$. *Homo sapiens* has a mean number $\overline{n} = 583$ of codons per gene and, as already mentioned, a RFR function $f(H.\ sapiens) = 0.792$ (thus, above the linear regression).

Figure 4 shows that the mean numbers $\overline{n}$ of codons per gene of the 432 archaeal genomes all belong to the interval [220, 350]. The overall mean number $\overline{\overline{n}}$ of codons per gene, over all these archaeal genomes, is 287 (see Table 2). Thus, in average, the archaeal genes are the shortest in the three kingdoms. This observation could be a consequence of a fundamental structure of archaeal genomes or related to its low number of sequenced genomes. The RFR function $f(Archaea)$ decreases according to a Spearman's rank correlation coefficient $\rho = -0.12$ with a significant p-value $=$ 0.013.

Table 2 gives some additional statistical parameters. Eukaryotic genomes have the highest mean number $\overline{\overline{n}} = 519$ of codons per gene and the archaeal genomes the lowest value $\overline{\overline{n}} = 287$, in agreement with their respective intervals. Concerning the reading frame retrieval function, eukaryotic genomes have the highest value mean$(f) = 0.736$ and the archaeal genomes the lowest value mean$(f) = 0.323$. The highest and lowest

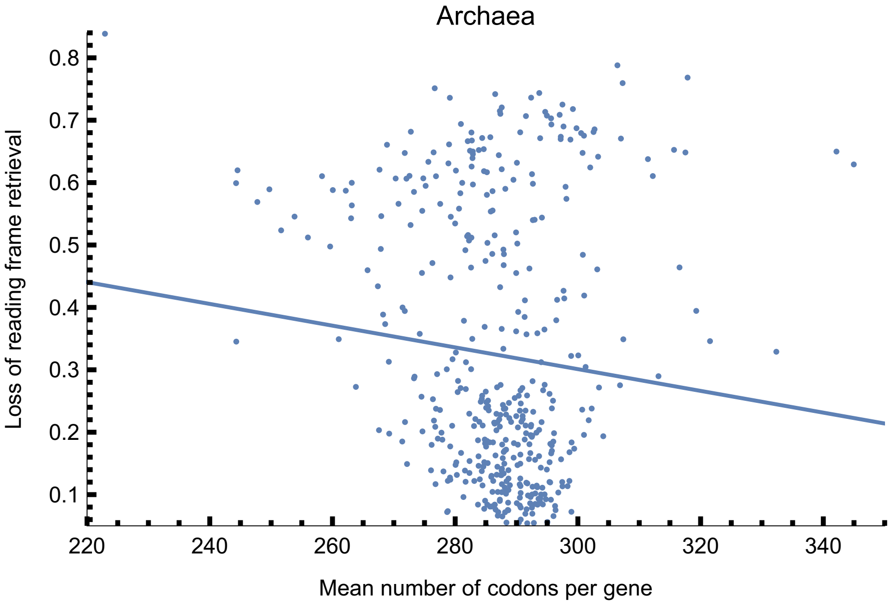


Fig. 4 Loss of reading frame retrieval in genes of 432 archaeal genomes. Each point represents all the genes of an archaeal genome. The x-axis shows the mean number $\overline{n}$ (Definition 2.5) of codons per gene, i.e. the total number of codons in a genome divided by its total number of genes. The y-axis shows the reading frame retrieval function f (2.3). The linear regression between $\overline{n}$ and f decreases according to the equation $y = -0.00174221x + 0.823736$, with a Spearman's rank correlation coefficient $\rho = -0.12$ and p-value $= 0.013$ (Color figure online)

Table 2 Basic statistical parameters of genomes in the three kingdoms studied. The reading frame retrieval function f is given by (2.3)

Kingdom	Basic statistical parameters			
	Mean number $\overline{\overline{n}}$ of codons per gene	mean(f)	min(f)	max(f)
Bacteria	325	0.426	0.016	1.032
Eukaryota	519	0.736	0.133	1.020
Archaea	287	0.323	0.051	0.839

values for the RFR function are observed with the bacterial genomes: $\max(f) = 1.032$ in *Helicobacter pametensis* and $\min(f) = 0.016$ in *Corynebacterium sphenisci DSM*.

3.2 Reading Frame Retrieval Correlated with Dispersion of Codon Usage

We compute the gene reading frame retrieval (RFR) function f (2.3) according to the dispersion function d (2.1) of codon usage in the genomes of the three kingdoms.

Figure 5 is somewhat spectacular for the following reasons. An exceptional correlation with a coefficient $\rho = -0.83$ and p-value $< 10^{-180}$ is identified between the codon usage dispersion and the RFR function. As expected from the theory (see

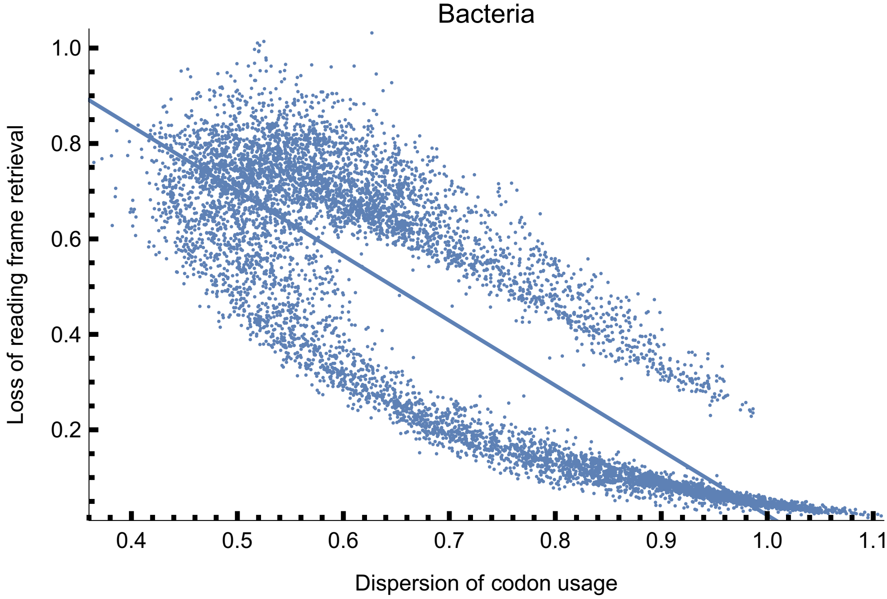


Fig. 5 Reading frame retrieval correlated with dispersion of codon usage in genes of 8345 bacterial genomes. A "genome centre" and two "genome arms" are identified in bacteria. Each point represents all the genes of a bacterial genome. The x-axis shows the dispersion function d (2.1) of codon usage. The y-axis shows the reading frame retrieval function f (2.3). The linear regression between d and f decreases according to the equation $y = -1.35881x + 1.37993$, with a Spearman's rank correlation coefficient $\rho = -0.83$ and p-value $< 10^{-180}$ (Color figure online)

Propositions 2.9 and 2.14), the bacterial RFR function decreases, or equivalently the property of reading frame retrieval increases in bacterial genomes, with the codon usage dispersion ranging from $d = 0.36$ with high RFR values $f > 0.6$ to $d = 1.1$ with low RFR values $f < 0.1$. However, this RFR increase uses two evolutionary processes. From a bacterial "genome centre" ranging approximately from $d = 0.36$ to $d = 0.55$, two bacterial "genome arms" emerge. The upper arm (above the linear regression) is ranging from $d = 0.55$ to $d < 1.0$. The lower arm (below the linear regression) is longer and ranges from $d = 0.55$ to $d = 1.1$.

In order to investigate these two genome arms, we analyse the codon dispersion of each amino acid (AA) with the Spearman's rank correlation coefficient ρ. The 15 AA codons leading to negative ρ values, similarly to the codon usage, are the following ones: Ala codons ($\rho = -0.92$), Arg codons ($\rho = -0.75$), Asp codons ($\rho = -0.66$), Cys codons ($\rho = -0.07$), Gln codons ($\rho = -0.55$), Gly codons ($\rho = -0.84$), His codons ($\rho = -0.44$), Ile codons ($\rho = -0.48$), Leu codons ($\rho = -0.90$), Phe codons ($\rho = -0.31$), Pro codons ($\rho = -0.82$), Ser codons ($\rho = -0.82$, identical value to Pro codons), Thr codons ($\rho = -0.90$), Tyr codons ($\rho = -0.20$) and Val codons ($\rho = -0.86$). The 4 AA codons leading to a positive ρ values, thus not involved in the dispersion of codon usage, are the following ones: Asn codons ($\rho = 0.20$), Lys codons ($\rho = 0.61$), Met codons ($\rho = 0.62$) and Trp codons ($\rho = 0.64$). The Glu codons have a non-significant value $\rho = -0.02$. However, none of the 15 AA codons

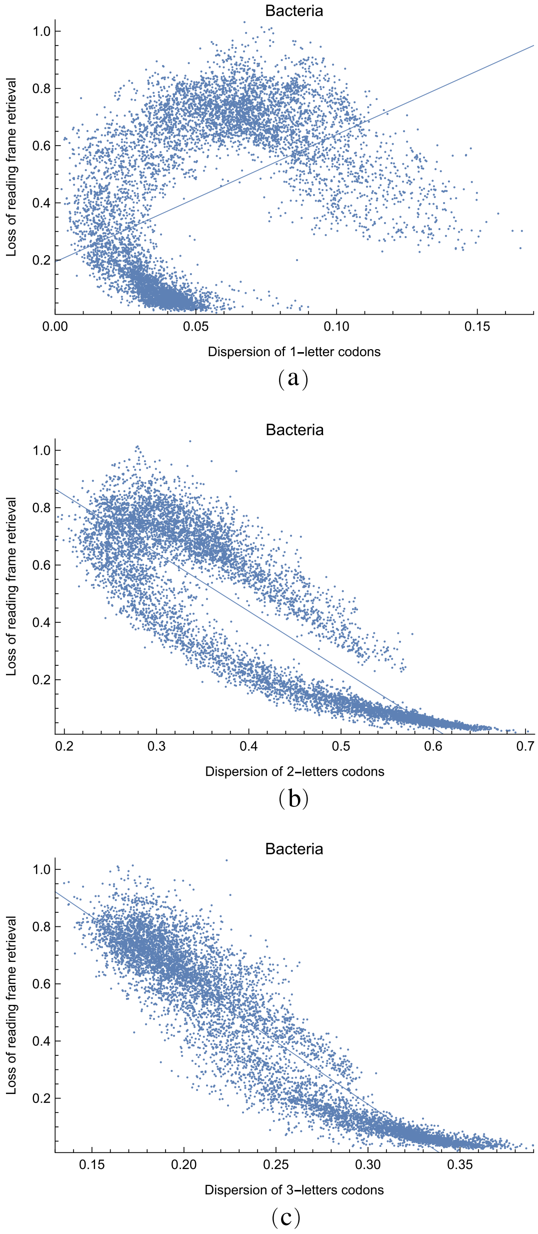


Fig. 6 Reading frame retrieval correlated with codon dispersion in genes of 8345 bacterial genomes. Each point represents all the genes of a bacterial genome. The x-axis shows the codon dispersion function d (2.1). The y-axis shows the reading frame retrieval function f (2.3). The linear regression between d and f, and the Spearman's rank correlation coefficient ρ with its p-value are given for each code. (a) 1-letter codons: $y = 4.45043x + 0.193461$, with $\rho = 0.44$ and p-value $< 10^{-180}$. (b) 2-letter codons: $y = -2.03092x + 1.25098$, with $\rho = -0.85$ and p-value $< 10^{-180}$. (c) 3-letter codons: $y = -4.36971x + 1.49$, with $\rho = -0.94$ and p-value $< 10^{-180}$ (Color figure online)

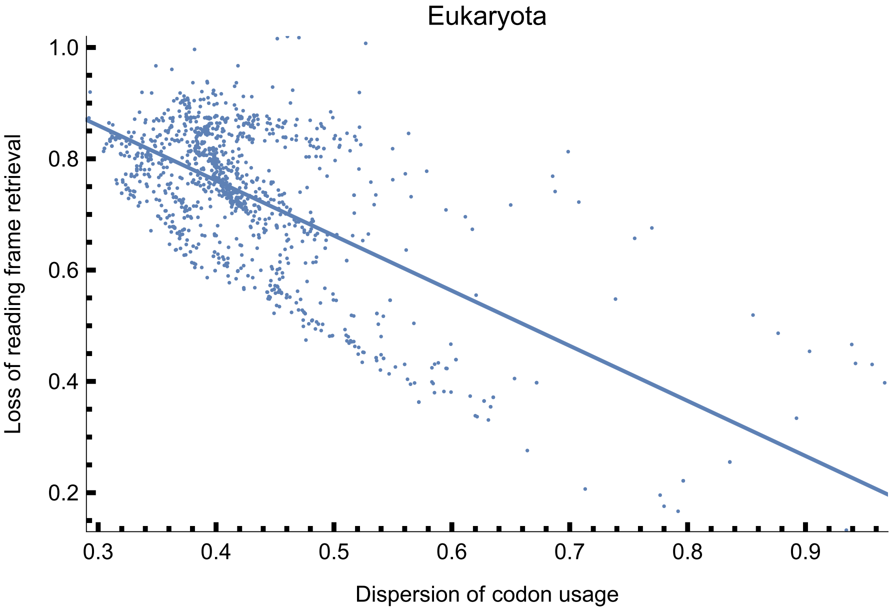


Fig. 7 Reading frame retrieval correlated with dispersion of codon usage in genes of 1150 eukaryotic genomes. A "genome centre" and two "genome arms" are identified in eukaryotes. Each point represents all the genes of an eukaryotic genome. The x-axis shows the dispersion function d (2.1) of codon usage. The y-axis shows the reading frame retrieval function f (2.3). The linear regression between d and f decreases according to the equation $y = -0.989903x + 1.15688$, with a Spearman's rank correlation coefficient $\rho = -0.45$ and p-value $< 10^{-61}$ (Color figure online)

with negative ρ values has a well-differentiated 2-arm structure (results not shown). This observation suggests that dispersion associated with reading frame retrieval is not associated with a particular amino acid.

We continue our dispersion analysis by considering the 3 codes related to the number of different nucleotides in the codons. The set of 1-letter codons, of cardinality 4, is

$$\{AAA, CCC, GGG, TTT\},$$

the set of 2-letter codons, of cardinality 36, is

$$\{AAC, AAG, AAT, ACA, ACC, AGA, AGG, ATA, ATT, CAA, CAC, CCA, \\ CCG, CCT, CGC, CGG, CTC, CTT, GAA, GAG, GCC, GCG, GGA, GGC, \\ GGT, GTG, GTT, TAA, TAT, TCC, TCT, TGG, TGT, TTA, TTC, TTG\},$$

and the set of 3-letter codons, of cardinality 24, is

$$\{ACG, ACT, AGC, AGT, ATC, ATG, CAG, CAT, CGA, CGT, CTA, CTG, \\ GAC, GAT, GCA, GCT, GTA, GTC, TAC, TAG, TCA, TCG, TGA, TGC\}.$$

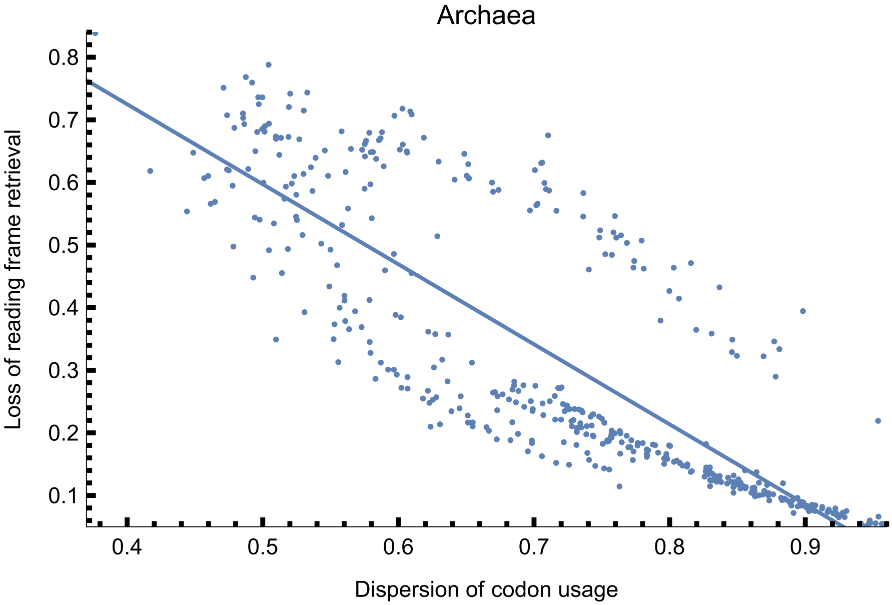


Fig. 8 Reading frame retrieval correlated with dispersion of codon usage in genes of 432 archaeal genomes. A “genome centre” and two “genome arms” are identified in archaea. Each point represents all the genes of an archaeal genome. The x-axis shows the dispersion function d (2.1) of codon usage. The y-axis shows the reading frame retrieval function f (2.3). The linear regression between d and f decreases according to the equation $y = -1.27828x + 1.23627$, with a Spearman’s rank correlation coefficient $\rho = -0.86$ and p-value $< 10^{-159}$ (Color figure online)

The 1-letter codons with a positive ρ value do not contribute to the dispersion associated with reading frame retrieval (Fig. 6a). In contrast, the 2-letter codons lead to a well-differentiated 2-arm structure (Fig. 6b, similarly to Fig. 5). These two arms are less structured with the 3-letter codons (Fig. 6c). In summary, the dispersion of codon usage associated with reading frame retrieval in genes is strongly associated with the 2-letter codons and also, to a lesser extent, with the 3-letter codons, i.e. the 60 codons (except the 4 periodic codons).

This correlation between the codon usage dispersion in the genes of bacterial genomes and the RFR function is retrieved in the genes of eukaryotic and archaeal genomes. Figure 7 shows that the eukaryotic RFR function decreases with the codon usage dispersion ranging from $d = 0.3$ with high RFR values $f > 0.8$ to $d = 0.96$ with low RFR values $f < 0.4$. Two eukaryotic genome arms are observable but the upper arm is very sparse. *Homo sapiens* has a codon usage dispersion $d = 0.40$ and, as already mentioned, a RFR function $f(H.\,sapiens) = 0.792$. He is thus located in the eukaryotic genome centre (see Fig. 7).

Figure 8 shows that the archaeal RFR function decreases with the codon usage dispersion ranging from $d = 0.38$ with high RFR values $f > 0.8$ to $d = 0.96$ with low RFR values $f < 0.1$. Two archaeal genome arms are significantly observable.

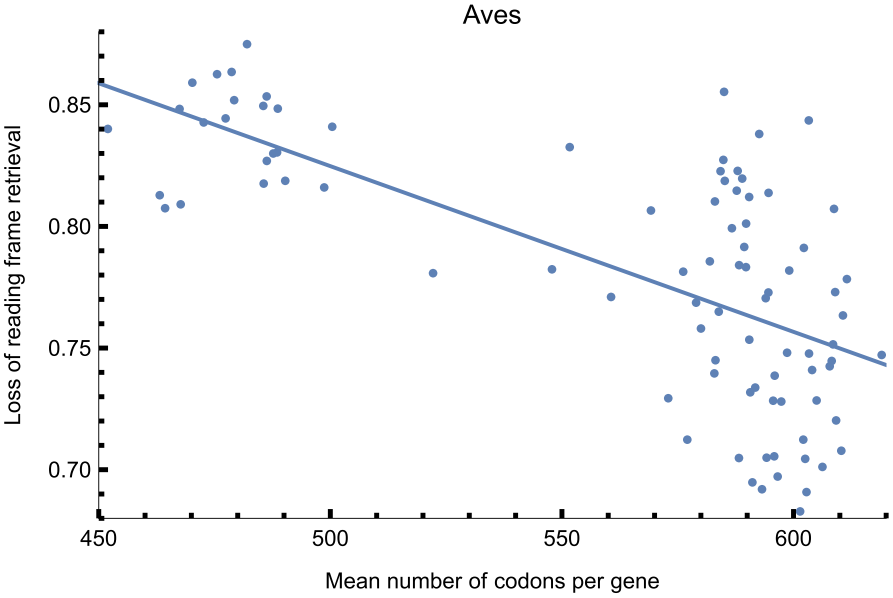


Fig. 9 Loss of reading frame retrieval in genes of 88 bird genomes (Aves, Id 8782, 88 genomes, 1318882 genes, 746444944 codons) identifying 2 groups of genomes. Each point represents all the genes of a bird genome. The x-axis shows the mean number $\overline{n}$ (Definition 2.5) of codons per gene, i.e. the total number of codons in a genome divided by its total number of genes. The y-axis shows the reading frame retrieval function f (2.3). The linear regression between $\overline{n}$ and f decreases according to the equation $y = -0.000681184x + 1.16537$, with a Spearman's rank correlation coefficient $\rho = -0.66$ and p-value $< 10^{-12}$ (Color figure online)

3.3 A Potential Evolutionary Classification of Genomes Based on Reading Frame Retrieval

As we have seen in the previous sections with the RFR function according to the codon usage dispersion, the genome centre and its two arms could be useful for a new and global classification of genomes. Furthermore, at a lower level, by analysing the bird genomes (Aves, Id 8782, 88 genomes, 1318882 genes, 746444944 codons) during the construction process of eukaryotic genomes, the reading frame retrieval function f (2.3) surprisingly identifies 2 groups of genomes (Fig. 9). Thus, according to all these observations, this RFR function by itself or combined with classical methods (alignment, phylogeny) could also classify the genomes, a new approach that could be investigated in the future.

4 Conclusion

Codon usage bias depends on a great number of biological factors (described in a recent review, see Parvathy et al. 2022). The statistical parameters analysing CUB were mainly developed by considering the coding of amino acids (see Introduction). In contrast to these classical methods and by using the circular code theory, we have defined here a new function f that quantifies the property of reading frame retrieval (RFR) of genes from their codon usage. Furthermore, by expressing f as a function of the mean number $\overline{n}$ of codons per gene, a “universal” property is identified in genes of genomes of bacteria, eukaryotes and archaea: the reading frame retrieval is enhanced in large genes. Then, by expressing f as a function of the codon usage dispersion d (from the uniform codon distribution $\frac{1}{64}$), another property with a strong statistical significance is found whatever the kingdom: the reading frame retrieval is enhanced with the codon usage dispersion. Surprisingly, this approach has revealed a genome centre from which emerge two distinct genome arms: an upper arm and a lower arm, respectively, above and below the linear regression. As CUB could have evolved through reading frame retrieval, the RFR function by itself or combined with classical methods (alignment, phylogeny) could also be a new approach to classify the genes and the genomes in the future.

Appendix: Codon Usage of *Homo sapiens*

See the Table 3.

Table 3 Codon usage of *Homo sapiens* (Id 9606, 19850 genes, 11577026 codons). The occurrence number and the frequency of each codon are given

AAA	292377	2.53	CAA	148603	1.28	GAA	352949	3.05	TAA	5636	0.05
AAC	212987	1.84	CAC	174799	1.51	GAC	287974	2.49	TAC	166083	1.43
AAG	365013	3.15	CAG	401364	3.47	GAG	464765	4.01	TAG	4436	0.04
AAT	197831	1.71	CAT	129622	1.12	GAT	255933	2.21	TAT	137604	1.19
ACA	177798	1.54	CCA	204198	1.76	GCA	187108	1.62	TCA	150051	1.30
ACC	213847	1.85	CCC	237781	2.05	GCC	323249	2.79	TCC	206550	1.78
ACG	68754	0.59	CCG	86427	0.75	GCG	89097	0.77	TCG	53015	0.46
ACT	156904	1.36	CCT	211342	1.83	GCT	213559	1.84	TCT	182288	1.57
AGA	142934	1.23	CGA	70319	0.61	GGA	193721	1.67	TGA	9773	0.08
AGC	231063	2.00	CGC	119972	1.04	GGC	258040	2.23	TGC	142859	1.23
AGG	140481	1.21	CGG	132275	1.14	GGG	191094	1.65	TGG	142049	1.23
AGT	147388	1.27	CGT	52129	0.45	GGT	123881	1.07	TGT	124438	1.07
ATA	87790	0.76	CTA	83043	0.72	GTA	83020	0.72	TTA	91638	0.79
ATC	224965	1.94	CTC	220601	1.91	GTC	161378	1.39	TTC	222373	1.92
ATG	244092	2.11	CTG	449485	3.88	GTG	314223	2.71	TTG	150213	1.30
ATT	182397	1.58	CTT	155373	1.34	GTT	127368	1.10	TTT	196707	1.70

References

Arquès DG, Michel CJ (1996) A complementary circular code in the protein coding genes. J Theor Biol 182:45–58

Bali V, Bebok Z (2015) Decoding mechanisms by which silent codon changes influence protein biogenesis and function. Int J Biochem Cell Biol 64:58–74

Buhr F, Jha S, Thommen M, Mittelstaet J, Kutz F, Schwalbe H, Rodnina M, Komar AA (2016) Synonymous codons direct cotranslational folding toward different protein conformations. Mol Cell 61:341–351

Dila G, Ripp R, Mayer C, Poch O, Michel CJ, Thompson JD (2019) Circular code motifs in the ribosome: a missing link in the evolution of translation? RNA 25:1714–1730

Fimmel E, Giannerini S, Gonzalez DL, Strüngmann L (2015) Circular codes, symmetries and transformations. J Math Biol 70:1623–1644

Fimmel E, Michel CJ, Pirot F, Sereni J-S, Strüngmann L (2019) Mixed circular codes. Math Biosci 317(108231):1–14

Fimmel E, Michel CJ, Pirot F, Sereni J-S, Starman M, Strüngmann L (2020) The relation between k-circularity and circularity of codes. Bull Math Biol 82(105):1–34

Fimmel E, Michel CJ, Strüngmann L (2016) n-Nucleotide circular codes in graph theory. Phil Trans R Soc A 374(20150058):1–19

Fimmel E, Strüngmann L (2018) Mathematical Fundamentals for the noise immunity of the genetic code. Biosystems 164:186–198

Fox JM, Erill I (2010) Relative codon adaptation: a generic codon bias index for prediction of gene expression. DNA Res 17:185–196

Grantham R, Gautier C, Gouy M, Mercier M, Gautier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r431–r74

He Z, Gan H, Liang X (2019) Analysis of synonymous codon usage bias in Potato Virus M and its adaption to hosts. Viruses 11(752):1–17

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Michel CJ (1986) New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its evaluation. J Theor Biol 120:223–236

Michel CJ (2008) A 2006 review of circular codes in genes. Comput Math Appl 55:984–988

Michel CJ (2017) The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, archaea, eukaryotes, plasmids and viruses. Life 7(2):1–16

Michel CJ (2020) The maximality of circular codes in genes statistically verified. Biosystems 197(104201):1–7

Michel CJ, Mayer C, Poch O, Thompson JD (2020) Characterization of accessory genes in coronavirus genomes. Virol J 17(131):1–13

Michel CJ, Mouillon B, Sereni J-S (2022) Trinucleotide k-circular codes I: theory. Biosystems 217(104667):1–11

Michel CJ, Pirillo G (2010) Identification of all trinucleotide circular codes. Comput Biol Chem 34:122–125

Michel CJ, Pirillo G, Pirillo MA (2008) Varieties of comma free codes. Computer and mathematics with applications 55:989–996

Michel CJ, Pirillo G, Pirillo MA (2008) A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theoret Comput Sci 401:17–26

Michel CJ, Sereni J-S (2022) Trinucleotide k-circular codes II: biology. Biosystems 217(104668):1–18

Michel CJ, Thompson JD (2020) Identification of a circular code periodicity in the bacterial ribosome: origin of codon periodicity in genes? RNA Biol 17:571–583

Parvathy ST, Udayasuriyan V, Bhadana V (2022) Codon usage bias. Mol Biol Rep 49:539–565

Pirillo G (2003) A characterization for a set of trinucleotides to be a circular code, by C. Pellegrini, P. Cerrai, P. Freguglia, V. Benci, G. Israel Determinism, Holism, and Complexity, Kluwer

Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR, Coller J (2015) Codon optimality is a major determinant of mRNA stability. Cell 160:1111–1124

Qian W, Yang JR, Pearson NM, Maclean C, Zhang J (2012) Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet 8:e1002603

Roymondal U, Das S, Sahoo S (2009) Predicting gene expression level from relative codon usage bias: an application to Escherichia coli genome. DNA Res 16:13–30

Sharp PM, Li W-H (1987) The codon adaptation index: a measure of directional synonymous codon usage, and its potential applications. Nucleic Acids Res 15:1281–1295

Sharp PM, Tuohy TMF, Mosurski KR (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14:5125–5143

Subramanian K, Payne B, Feyertag F, Alvarez-Ponce D (2022) The codon statistics database: a database of codon usage bias. Mol Biol Evol 39(8):1–3

Thompson JD, Ripp R, Mayer C, Poch O, Michel CJ (2021) Potential role of the *X* circular code in the regulation of gene expression. Biosystems 203(104368):1–15

Wright F (1990) The "effective number of codons" used in a gene. Gene 87:23–29

Yu X, Liu J, Li H, Liu B, Zhao B, Ning Z (2021) Comprehensive analysis of synonymous codon usage patterns and influencing factors of porcine epidemic diarrhea virus. Adv Virol 166:157–165

Zhang Z, Li J, Cui P, Ding F, Li A, Townsend JP, Yu J (2012) Codon deviation coefficient: a novel measure for estimating codon usage bias and its statistical significance. BMC Bioinf 13:1–10

Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J, Chen S, Liu Y (2016) Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proceed Nat Academy Sci USA 113:E6117–E6125

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.