ORIGINAL ARTICLE

Genome Galaxy Identified by the Circular Code Theory

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

© The Author(s), under exclusive licence to the Society for Mathematical Biology 2024

Abstract

The genome galaxy identified in bacteria is studied by expressing the reading frame retrieval (RFR) function according to the YZ-content (GC-, AG- and GT-content) of bacterial codons. We have developed a simple probabilistic model for ambiguous sequences in order to show that the RFR function is a measure of the gene reading frame retrieval. Indeed, the RFR function increases with the ratio of ambiguous sequences and the ratio of ambiguous sequences decreases when the codon usage dispersion increases. The classical GC-content is the best parameter for characterizing the upper arm, which is related to bacterial genes with a low GC-content, and the lower arm, which is related to bacterial genes with a high GC-content. The galaxy center has a GC-content around 0.5. Then, these results are confirmed by expressing the GC-content of bacterial codons as a function of the codon usage dispersion. Finally, the bacterial genome galaxy is better described with the $GC3$-content in the 3rd codon site compared to the $GC1$-content and $GC2$-content in the 1st and 2nd codons sites, respectively. Whereas the codon usage is used extensively by biologists, its dispersion, which is an important parameter to reveal this genome galaxy, is surprisingly little known and unused. Therefore, we have developed a mathematical theory of codon usage dispersion by deriving several formulæ. It shows three important parameters in codon usage: the minimum and maximum codon probabilities and the number of codons with high frequency, i.e. with a probability at least 1/64. By applying this theory to the evolution of the genetic code, we see that bacteria have optimised the number of codons with high frequency to maximise the codon dispersion, thus maximising the capacity to retrieve the reading frame in genes. The derived formulæ of dispersion can be easily extended to any weighted code over a finite alphabet.

Keywords Genome galaxy · Circular code theory · Reading frame retrieval · Probabilistic model for ambiguous sequences · Formulas for codon usage · Codon dispersion

✉ 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

Springer

1 Introduction

Based on the circular code theory, a beautiful and intriguing "galaxy" structure has been identified in the genomes of bacteria, as well as of eukaryota and archaea (Michel and Sereni 2023). This genome galaxy has a center and two arms, an upper one and a lower one, a structure that is identified for the three kingdoms (Michel and Sereni 2023, Figures 5, 7 and 8). The aim of this work is to characterise this genome galaxy for bacteria.

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 in genes (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 generalizations: 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 center, containing numerous sites that interact with the tRNA and the 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 characterization of accessory genes in coronavirus genomes using the circular code information (Michel et al. 2020).

On the genetic alphabet, there are $2^{64} - 1 \approx 10^{19}$ (non-empty) trinucleotide codes: 64 codes of cardinality 1 ($\{AAA\}, \ldots, \{TTT\}$); 2016 codes of cardinality 2 ($\{AAA, AAC\}, \ldots, \{TTG, TTT\}$); 41664 codes of cardinality 3 ($\{AAA, AAC, AAG\}, \ldots, \{TTC, TTG, TTT\}$); and so on up to 1 code of cardinality 64 (the genetic code $\{AAA, \ldots, TTT\}$). The recent theory of trinucleotide k-circular codes makes it possible to study the property of reading frame retrieval (RFR), called circularity property, for any of these $\approx 10^{19}$ codes (Michel et al. 2022; Michel and Sereni 2022).

The genome galaxy of bacteria will be analysed by the RFR function f (see Definition 2.13) that can be applied to the codon usage, and two codon parameters: dispersion (see Definition 2.7) and YZ-content (see Definition 2.9). The YZ-content of codon is a simple extension of the GC-content, a main parameter to study the codon usage bias (CUB) that influences different aspects of protein production (Grantham et al. 1981) and 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) (recent review in Parvathy et al. 2022). In addition, from a theoretical point of view, our work puts for the first time the circular code theory with its RFR function f in relation to the codon usage with its GC-content.

This article is organised as follows. The necessary definitions and notation of trinucleotide codes, circular codes and their generalization to k-circular codes are gathered in Sect. 2.1. Section 2.2 defines the dispersion function of codon usage and states a proposition about its range. Section 2.3 defines the YZ-content. Section 2.4 defines the reading frame retrieval (RFR) function and states several propositions concerning its range and its particular value 1 associated with a uniform codon usage. Section 2.5 explains why the RFR function is a measure of the gene reading frame retrieval. Section 2.6 defines the parameters involved in our mathematical theory of codon usage dispersion. Section 2.7 describes the acquisition of codon usage for the genomes of bacteria from the codon statistics database (CSD) (Subramanian et al. 2022).

The results are presented in two main parts parts. Section 3 presents new statistical results of the bacterial genome galaxy. It is divided into three parts. Section 3.1 characterizes the genome galaxy of bacteria with its center, its upper arm and its lower arm. Section 3.2 shows that one parameter, the YZ-content of codon, and mainly the GC-content, allows the identification of the three structures of the genome galaxy. Section 3.3 demonstrates that the dispersion of codon usage is mainly related to the GC-content of codon and that the $GC3$-content in the 3rd codon site is a main factor for the reading frame retrieval (RFR) function of genes.

Section 4 develops a mathematical study of codon usage dispersion, an analysis which, to our knowledge, has never been carried out. It is divided into two parts. Section 4.1 derives several formulæ of codon usage dispersion. Section 4.2 gives three applications of this mathematical study to the evolution of the bacterial genetic code. The maximum dispersion of the current genetic code at 64 codons is analysed as a function of its codon of maximum probability (Sect. 4.2.1) and that of minimum probability (Sect. 4.2.2). Moreover, Sect. 4.2.3 studies the minimum and maximum dispersions of the evolutionary genetic code as functions of its number of codons, from 1 to 64.

2 Method

2.1 Definitions and Notation

For the reader's convenience, and to have this article self-contained, we here recall the most relevant notions. 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, 2023).

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 (RFR).

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 $(\geqslant k)$-*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 $(\geqslant k)$-circular if for every $m \in \{1, \ldots, k\}$ and each m-tuple $(x_1, \ldots, x_m)$ 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 $(\geqslant k)$-circular and not $(\geqslant k+1)$-circular.
- The code X is *circular* if it is $(\geqslant k)$-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:

- $V(X) := \displaystyle\bigcup_{N_1 N_2 N_3 \in X} \{N_1, N_3, N_1 N_2, N_2 N_3\}$; and
- $E(X) := \{N_1 \to N_2 N_3 : N_1 N_2 N_3 \in X\} \cup \{N_1 N_2 \to N_3 : N_1 N_2 N_3 \in X\}$.

The graph $\mathcal{G}(X)$ is the graph *associated* with 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 ($\geqslant 4$)-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 ($\geqslant k$)-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 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 (RFR) of genes. These two functions are also analysed as a function of the mean number of codons per gene in each genome.

2.2 Dispersion of Codon Usage

We recall the definition and the proposition of codon usage introduced in a previous work (Michel and Sereni 2023). A codon usage is *uniform* if every codon has the same occurrence frequency. We shall introduce a function to measure the dispersion of codon usage with respect to the uniform one. We write X_g instead of $\mathcal{B}^3$, the particular code of cardinality 64 containing all trinucleotides, which is the well-known genetic code.

Definition 2.5 (Codon usage) 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 2.6 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 2.7 (Dispersion of codon usage, Michel and Sereni 2023, 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 2.8 (Michel and Sereni 2023, Proposition 2.9) *For every weight function $\omega \colon X_g \to [0, 1]$, we have*

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

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* $x' \neq x$).

2.3 *YZ-Content*

Definition 2.9 Let n_A, n_C, n_G and n_T be the number of the nucleotide A, C, G and T of $\mathcal{B}$ in the 3 codon sites. Let Y and Z be 2 different nucleotides of $\mathcal{B}$. Then the YZ-content is the probability

$$YZ\text{-content} = \frac{n_Y + n_Z}{N} \tag{2.2}$$

where $N = n_A + n_C + n_G + n_T$.

Note that the number of codons is $N/3$. Obviously, YZ-content $= ZY$-content and YZ-content $+ \overline{Y}\,\overline{Z}$-content $= 1$ where the complementary nucleotide $\overline{N}$ of a nucleotide $N \in \mathcal{B}$ is given by $\overline{A} = T, \overline{T} = A, \overline{C} = G$ and $\overline{G} = C$. The classical biological parameter is the GC-content. In this work, we will also study the parameters AG-content and GT-content.

The definition of the YZ-content can easily be generalized to YZk-content associated with the kth codon site where $k \in \{1, 2, 3\}$. Note that, using the above notation, the normalisation factor (i.e., the denominator) is not N but $N/3$, i.e. the number of codons. The classical biological parameter is the $GC3$-content. In this work, we will also study the parameters $GC1$-content and $GC2$-content.

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 (e.g., 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). We will explain this important property in detail in the following Sect. 2.5.

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

$$f(X) := q_8(\mathcal{G}(X)) + \frac{4}{3} q_6(\mathcal{G}(X)) + 2\, q_4(\mathcal{G}(X)) + 4\, q_2(\mathcal{G}(X)) = \sum_{i=1}^{4} \frac{4}{i} \cdot q_{2 \cdot i}(\mathcal{G}(X)) \tag{2.3}$$

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

The next proposition 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 $0 \leqslant f(X) \leqslant 301056$. 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*

$$q_2(X_g) = 64, \qquad q_4(X_g) = 1440, \qquad q_6(X_g) = 26880, \qquad q_8(X_g) = 262080.$$

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

Definition 2.12 (Michel and Sereni 2023, 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 $N_1 N_2 N_3 \in X$ the rational number $\frac{\omega(N_1 N_2 N_3)}{2} \in [0, 1]$.

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

$$\left\{ N_1 \xrightarrow{\omega(x)/2} N_2 N_3 \ : \ x = N_1 N_2 N_3 \in X \right\} \cup \left\{ N_1 N_2 \xrightarrow{\omega(x)/2} N_3 \ : \ x = N_1 N_2 N_3 \in X \right\}.$$

The generalised function f associated with every weighted trinucleotide code that has identified the genome galaxy in bacteria, archaea and eukaryota, has been defined as follows.

Definition 2.13 (Michel and Sereni 2023, 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 (X, ω) is the mapping f given by

$$f((X, \omega)) := \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} (2|X|)^{|c|} \prod_{a \in E(c)} \omega'(a) \tag{2.4}$$

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

For the convenience of the reader, we recall three propositions (without proof).

PROPOSITION 2.14 (Uniform codon usage, Michel and Sereni 2023, Proposition 2.14) *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 $f((X_g, \omega)) = 1$.*

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.

PROPOSITION 2.15 (Circular code, Michel and Sereni 2023, Proposition 2.15) *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, leading to the following observation.

PROPOSITION 2.16 (Michel and Sereni 2023, Proposition 2.16) *We have*

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

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

2.5 Probabilistic Model for Ambiguous Sequences

The RFR function of a weighted trinucleotide code is associated with the ambiguous sequences that do not retrieve the reading frame. In order to demonstrate this property, we describe a simple probabilistic model $\mathcal{M}_1$ to quantify the capacity of a weighted trinucleotide code to retrieve the reading frame.

Let (X_g, ω) be a weighted trinucleotide code, where, as already mentioned, X_g is the genetic code of cardinality $|X_g| = 64$. Real-life genomes usually comprise all 64 codons. So we assume that w is positive, i.e. $w(x) > 0$ for each $x \in X_g$. As a consequence, every sequence of codons is ambiguous, i.e. it can be read in all 3 frames when written on a circle. We use the codon frequency of the sequences to quantify whether a sequence and its circular shifts can be identified, or not, as being in reading frame. Let us formalize this problem.

We fix a positive length m and consider all sequences (concatenations) composed of m trinucleotides of X_g, that is all sequences of $s = 3m$ nucleotides (as defined in Sect. 2.1). We want to determine the probability of detection error between any given sequence $w = N_1 \cdots N_s$ composed of s nucleotides (hence m trinucleotides) and its circular 1-shift $w_1 = N_2 \cdots N_s N_1$. We consider this case appears when at least one of the sequences w and w_1 occurs frequently enough and the discrepancy between their two codon frequencies is not too large. These conditions depend on the weight ω, i.e. the codon usage. This approach is based on the two following observations:

Observation (i)   if both w and w_1 have a low probability of occurring (i.e., under an arbitrarily fixed threshold) then an error on their reading frame does not occur often, and thus could be considered biologically insignificant; and

Observation (ii)   if one of them, say w_1 for instance, occurs much less frequently than w, then it is certain to consider the reading frame of w rather than that of w_1, since this frame will be "almost always" correct.

Consequently, every such couple (w, w_1) in these two cases is considered not to represent a problematic ambiguity regarding the reading frame.

We now define the codon frequency of the sequence $w = N_1 \cdots N_s$ composed of the m codons $x_i = N_{3i+1} N_{3i+2} N_{3i+3}$ for each $i \in \{0, \ldots, m-1\}$. The probability of a codon $x \in X_g$ is directly given by the weight function ω. By assuming independence of the codon frequencies within a sequence, then the probability $p(w)$ of w is

$$p(w) = \prod_{i=0}^{m-1} \omega(x_i). \tag{2.5}$$

Fix two reals $\varepsilon > 0$ and $R > 1$, which will be used as a threshold and a ratio bound, respectively. For each such sequence w, we compute its probability $p(w)$ and the probability $p(w_1)$ of its circular 1-shift w_1. We say these two sequences w and w_1 are *ambiguous* if:

$$
\begin{aligned}
&(1) \quad p(w) > \varepsilon \text{ and } p(w_1) > \frac{p(w)}{R}; \text{ or} \\
&(2) \quad p(w_1) > \varepsilon \text{ and } p(w) > \frac{p(w_1)}{R}.
\end{aligned}
\tag{2.6}
$$

When none of these two cases occurs, we consider the sequences to be not ambiguous as explained earlier (see Observation (i) and Observation (ii)).

We additionally note that the condition to declare w and w_1 ambiguous is symmetric with respect to w and w_1. Therefore, it is enough to consider circular 1-shifts. Indeed, with sequences of trinucleotides, if w_2 is the circular 2-shift of w, then w is the circular 1-shift of w_2, so that the pair $\{w, w_2\}$ will be dealt with anyway.

Example 2.17 We use the (average) codon usage of X_g provided in Appendix Table 1 to define ω, so $\omega(AAA) = 0.0237$ and $\omega(TTT) = 0.0161$ for instance. Consider the sequence $w = AAA \cdot AAC \cdot AAG$ composed of $m = 3$ codons. The probability of w is $p(w) = \omega(AAA) \times \omega(AAC) \times \omega(AAG) = 0.0237 \times 0.0180 \times 0.0200 \approx 8.5 \times 10^{-6}$. The circular 1-shift of w is the sequence $w_1 = AAA \cdot ACA \cdot AGA$, and hence $p(w_1) = \omega(AAA) \times \omega(ACA) \times \omega(AGA) = 0.0237 \times 0.0093 \times 0.0054 \approx 1.2 \times 10^{-6}$.

Suppose, for example, that the threshold is $\varepsilon = \frac{1}{|X_g|^3} = \frac{1}{64^3} \approx 3.8 \times 10^{-6}$ and the ratio bound is $R = 10$. We then see that $p(w) > \varepsilon$ and in addition that $p(w_1) > \frac{p(w)}{10}$. Therefore with these parameters the sequences w and w_1 are ambiguous according to (2.6).

We perform the above test (2.6) for all sequences of m trinucleotides, giving us the number a of ambiguous pairs. The normalisation of a by dividing it by the total number $|X_g|^m$ of such sequences, leads to the ratio r (rational number in $[0, 1]$) of ambiguous sequences (of m trinucleotides) for the weighted trinucleotide code (X_g, ω)

$$r = \frac{a}{|X_g|^m} = \frac{a}{64^m}. \tag{2.7}$$

This ratio r can be interpreted as the probability that a sequence of m consecutive trinucleotides is ambiguous, meaning that there is a significant detection error of the reading frame.

Remark 2.18 There is no unique model to quantify the reading frame retrieval capacity of all the bacterial genomes. A 2nd model $\mathcal{M}_2$ can consider a "shifted" codon usage of sequences in a shifted frame. Given a sequence w and its circular 1-shift w_1, the probability that w_1 occurs *in reading frame* 1 can be computed using probabilities obtained by shifting the original (i.e. frame 0) codon usage. Specifically, the weight ω_1 for each trinucleotide $N_1 N_2 N_3 \in X_g$ is defined as follows:

$$\omega_1(N_1 N_2 N_3) = \left(\sum_{N \in \mathcal{B}} \omega(N N_1 N_2) \right) \cdot \left(\sum_{N', N'' \in \mathcal{B}} \omega(N_3 N' N'') \right). \qquad (2.8)$$

Indeed, $N_1 N_2 N_3$ is read in frame 1 if and only if $N N_1 N_2 \cdot N_3 N' N''$ is read in frame 0 for some $N, N', N'' \in \mathcal{B}$. The probability of a given concatenation $w = N_1 \cdots N_s$ of $s = 3m$ nucleotides then depends on the position of the sequence with respect to the actual reading frame: if this sequence occurs in frame 0 then the probability is, as before, $p(w) = \prod_{i=1}^m \omega(x_i)$, while if it occurs in frame 1 then the probability becomes $p_1(w) = \prod_{i=1}^m \omega_1(x_i)$. Using conditions analogous to (2.6), the model $\mathcal{M}_2$ recovers results similar to the model $\mathcal{M}_1$ (not shown to avoid overloading the content).

Finally, a 3rd model $\mathcal{M}_3$ can also be based on $p_2(w)$ by shifting the 1-frame codon usage. Then, w can be considered unambiguous if one of the 3 probabilities $p(w)$, $p_1(w)$ and $p_2(w)$ is "significantly greater" than the other two. But in this case, as almost all sequences are ambiguous, the use of a threshold (similar to Observation (i)) is necessary. The model $\mathcal{M}_3$ finds similar results to the previous models $\mathcal{M}_1$ and $\mathcal{M}_2$ (not shown).

The ratio r (2.7) of ambiguous sequences will be computed in Sects. 3.1.2, 3.1.3 and 3.2 on all sequences of $m = 4$ trinucleotides with the threshold $\varepsilon = \frac{1}{10 \cdot 64^4}$ and the ratio bound $R = 10$. We have also performed computer calculations on shorter and longer sequences (namely $m = 3$ and $m = 5$) with the corresponding threshold (i.e. $\varepsilon = \frac{1}{10 \cdot 64^m}$) and obtained similar results (not shown). Other thresholds and ratio bounds were also analysed (in particular $\varepsilon = \frac{1}{64^m}$), and they led to similar results (similar shapes, only scaled along to the x-axis; not shown).

2.6 Parameters of a Codon Usage

Given a codon usage, or equivalently a weighted trinucleotide code $\mathcal{W} = (X_g, \omega)$ over the genetic code X_g, we retain several parameters that seem to influence in a non-trivial way the gene reading frame retrieval (RFR) function.

First, we discriminate between the trinucleotides that occur, and those that do not — or are so rare that are considered not to occur. Second, over these trinucleotides, we retain the lowest and the highest possible value of ω, and also the number of trinucleotides with "high frequency".

Definition 2.19 Let $\mathcal{W} = (X_g, \omega)$ be a weighted trinucleotide code. We define

(1) $p_M(\mathcal{W}) = \max \left\{ \omega(x) : x \in X_g \right\};$

(2) $p_m(\mathcal{W}) = \min \big\{ \omega(x) \,:\, x \in X_g \text{ and } \omega(x) > 0 \big\}$;

(3) the number $n(\mathcal{W})$ of trinucleotides x that occur, that is such that $\omega(x) > 0$; and

(4) the number $n_h(\mathcal{W})$ of trinucleotides x of *high frequency*, that is such that $\omega(x) \geqslant \frac{1}{n(\mathcal{W})}$.

2.7 Data

From the codon statistics database (CSD, http://codonstatsdb.unr.edu) (Subramanian et al. 2022), we have extracted (July 2022) the codon usage of genomes of bacteria 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). The few exceptional genomes in which the codon usage of the stop codons is not given, are not considered. Thus, the bacterial kingdom contains 8,345 genomes, 34,020,997 genes and 11,087,876,805 codons.

The calculus of the codon usage in this bacterial kingdom is given in Appendix Table 1.

3 Statistical Study of the Bacterial Genome Galaxy

### 3.1 Identification of the Genome Galaxy of Bacteria

3.1.1 Gene Reading Frame Retrieval (RFR) Function According to the Dispersion Function

By expressing the gene reading frame retrieval (RFR) function f (2.4) according to the dispersion function d (2.1) of codon usage in the bacterial genomes, a "genome galaxy" with a center and two arms has been identified in a previous work (Michel and Sereni 2023, Figure 5, Section 3.2). By using the linear regression $y = -1.35881x + 1.37993$ between d and f (with a Spearman rank correlation coefficient $\rho = -0.83$ and p-value $< 10^{-180}$), we characterize in this work these 3 structures. The galaxy center $\mathcal{GC}$ is defined by the bacterial genomes such that

$$\mathcal{GC} := d \leqslant 0.6. \tag{3.1}$$

The upper arm $\mathcal{GUA}$ is defined by the bacterial genomes such that

$$\mathcal{GUA} := \begin{cases} d > 0.6 \\ f > y \\ f > 0.2. \end{cases} \tag{3.2}$$
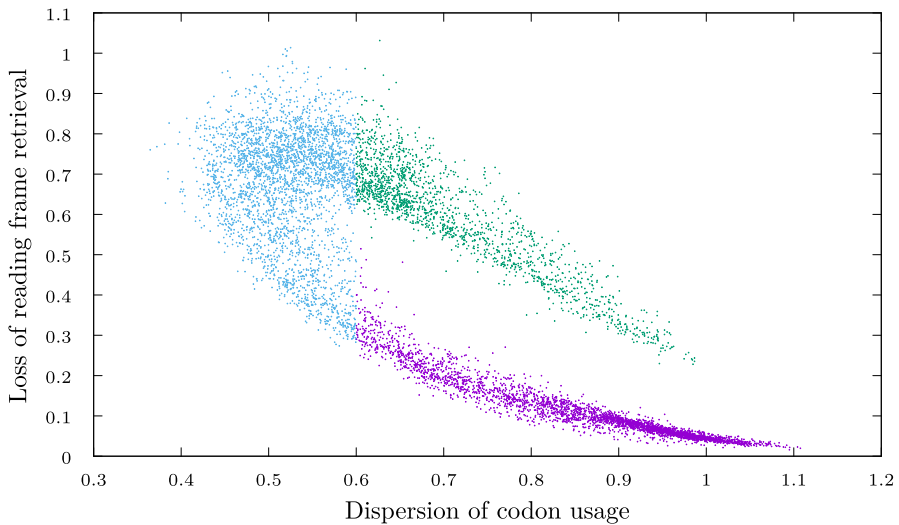
Fig. 1 Genome galaxy of bacteria (8,345 genomes, 34,020,997 genes, 11,087,876,805 codons) with its center $\mathcal{GC}$ (3.1) in blue, its upper arm $\mathcal{GUA}$ (3.2) in green and its lower arm $\mathcal{GLA}$ (3.3) in violet. Each point represents all the genes of a bacterial genome. The x-axis shows the dispersion function d (2.1) of codon usage. The y-axis shows the reading frame retrieval function f (2.4) (Color figure online)

The lower arm $\mathcal{GLA}$ is defined by the bacterial genomes such that

$$\mathcal{GLA} := \begin{cases} f < y & \text{if } d \in]0.6, 0.9] \\ f < 0.2 & \text{if } d > 0.9. \end{cases} \tag{3.3}$$

Figure 1 describes this genome galaxy of bacteria with its center $\mathcal{GC}$ (3.1) in blue, its upper arm $\mathcal{GUA}$ (3.2) in green and its lower arm $\mathcal{GLA}$ (3.3) in violet.

Several parameters are now investigated to analyse this codon dispersion. Surprisingly, in a next section, one parameter, the YZ-content of codon (Sect. 2.3), allows the identification of the three structures of the genome galaxy.

3.1.2 Gene Reading Frame Retrieval (RFR) Function According to the Ratio of Ambiguous Sequences

By expressing the gene reading frame retrieval (RFR) function f (2.4) according to the ratio r (2.7) of ambiguous sequences, Fig. 2 shows that the RFR function increases with the ratio of ambiguous sequences. Figure 2 is a "mirror image" of Fig. 1.

3.1.3 Ambiguous Sequences According to the Codon Usage Dispersion

By expressing the ratio r (2.7) of ambiguous sequences according to the codon usage dispersion d (2.1), Fig. 3 shows that the ratio of ambiguous sequences decreases when the codon usage dispersion increases. The dispersion of codon usage and the ratio of ambiguous sequences vary in opposite direction.
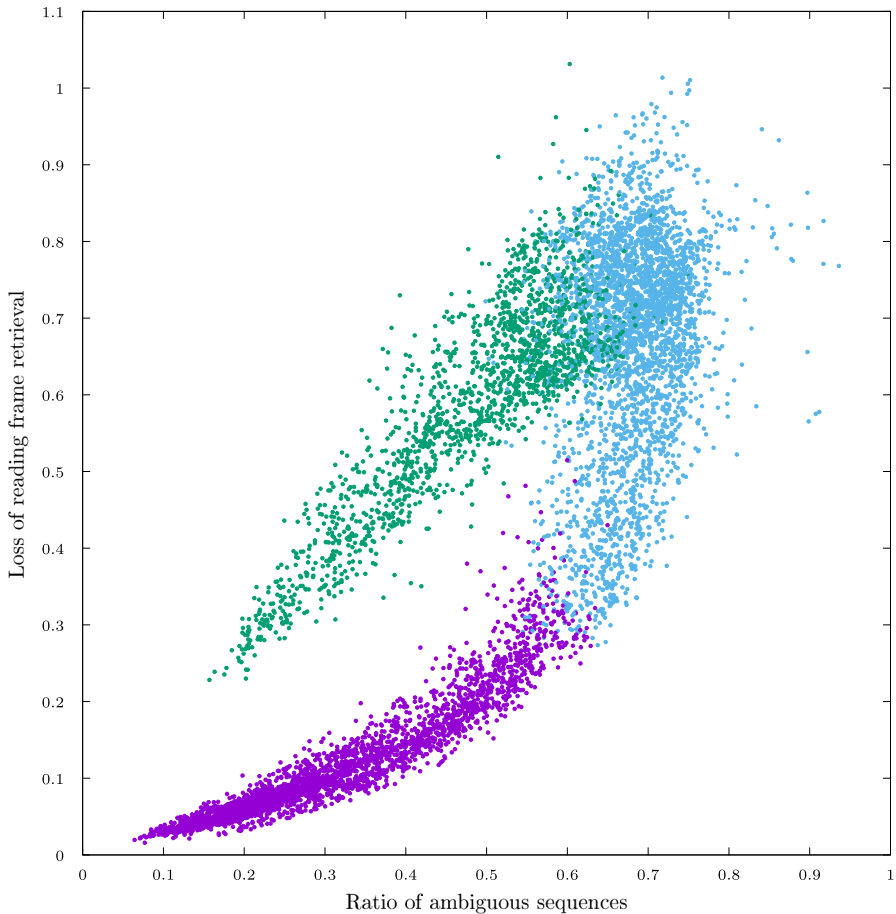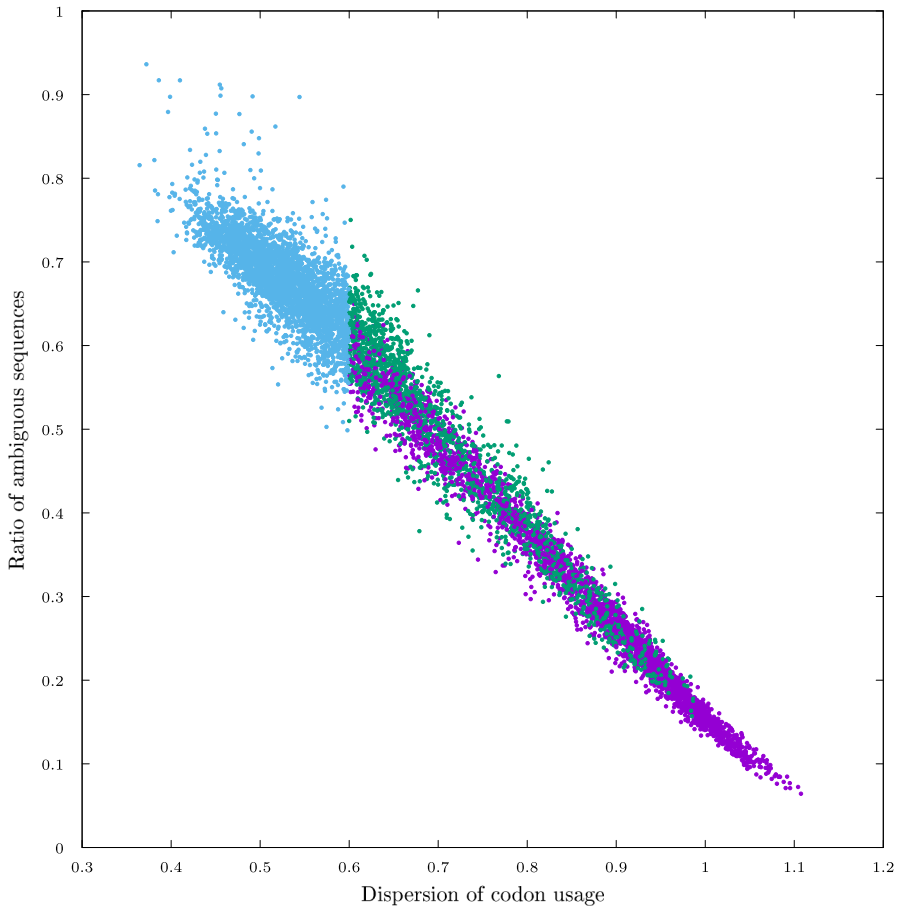
Fig. 2 Genome galaxy of bacteria (8,345 genomes, 34,020,997 genes, 11,087,876,805 codons) with its center $\mathcal{GC}$ (3.1) in blue, its upper arm $\mathcal{GUA}$ (3.2) in green and its lower arm $\mathcal{GLA}$ (3.3) in violet. Each point represents all the genes of a bacterial genome. The x-axis shows the ratio r (2.7) of ambiguous sequences. The y-axis shows the reading frame retrieval function f (2.4) (Color figure online)

Remark 3.1 It is very interesting to stress that the two arms of the bacterial genome galaxy could be identified thanks to the RFR function function f (2.4) but not by the ratio r (2.7) of ambiguous sequences (compare Figs. 1 and 3).

3.2 Genome Galaxy of Bacteria Identified by the *GC*-Content of Codon

By expressing the gene reading frame retrieval (RFR) function f (2.4) according to the GC-content of codon, the 3 structures: center $\mathcal{GC}$ (3.1), upper arm $\mathcal{GUA}$ (3.2) and lower arm $\mathcal{GLA}$ (3.3) are well characterized in Fig. 4A. The variation of the GC-content is
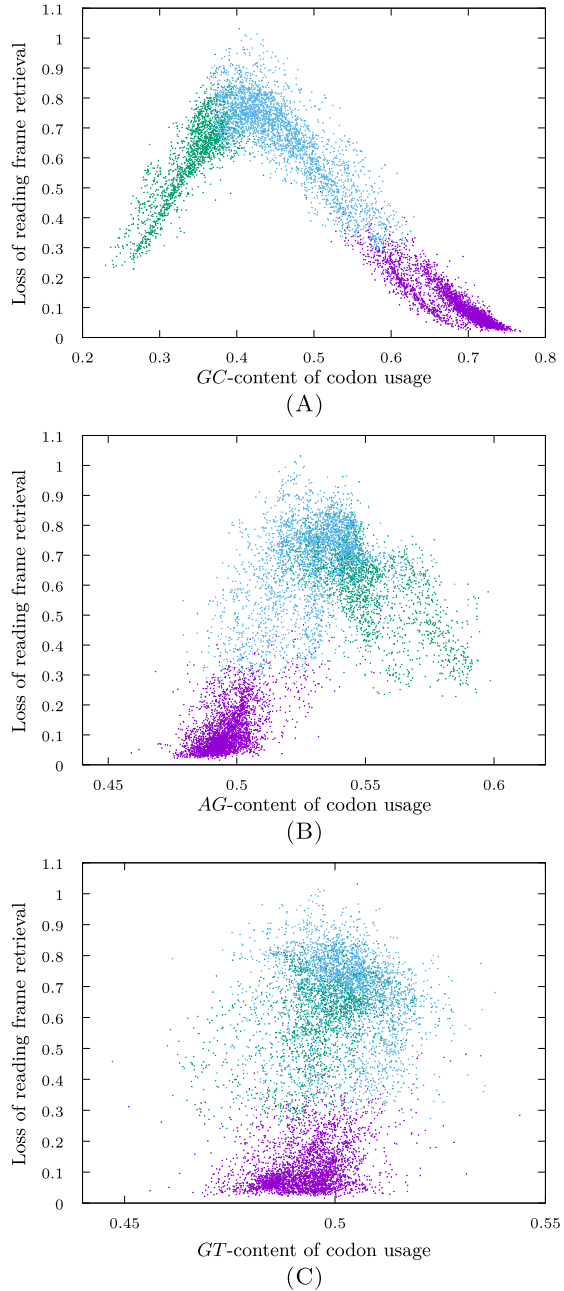
Fig. 3 Genome galaxy of bacteria (8,345 genomes, 34,020,997 genes, 11,087,876,805 codons) with its center $\mathcal{GC}$ (3.1) in blue, its upper arm $\mathcal{GUA}$ (3.2) in green and its lower arm $\mathcal{GLA}$ (3.3) in violet. Each point represents all the genes of a bacterial genome. The x-axis shows the dispersion function d (2.1) of codon usage. The y-axis shows the ratio r (2.7) of ambiguous sequences (Color figure online)

important and in the interval [0.2, 0.8]. Note that, AT being complementary to GC, the AT-content leads to a symmetrical figure with respect to $y = 0.5$ (not shown). Thus, the upper arm is related to genomes with a low GC-content while the lower arm is related to genomes with a high GC-content, the center being related to genomes with a GC-content around 0.5.

The variation of the AG-content (or equivalently GA-content as already mentioned) is restricted to the interval [0.45, 0.60]. The upper and lower arms are still separated but to a lesser extent (Fig. 4B).

The variation of the GT-content is restricted to the interval [0.45, 0.55]. The upper and lower arms are neighbours (Fig. 4C). Thus, the GT-content is not a parameter for characterizing the genome galaxy.

Fig. 4 Genome galaxy of bacteria (8345 genomes, 34,020,997 genes, 11,087,876,805 codons) with its center $\mathcal{GC}$ (3.1) in blue, its upper arm $\mathcal{GUA}$ (3.2) in green and its lower arm $\mathcal{GLA}$ (3.3) in violet, identified by the YZ-content of codon, and mainly by the GC-content. Each point represents all the genes of a bacterial genome. The y-axis shows the reading frame retrieval function f (2.4). The x-axis shows the YZ-content of codon in the 3 cases:
(A) GC-content;
(B) AG-content;
(C) GT-content (Color figure online)

## 3.3 Dispersion of Codon Usage Related to the *GC*-Content of Codon

According to the previous results, it is natural to express the dispersion function d (2.1) of codon usage in the bacterial genomes according to their YZ-content of

Fig. 5 Genome galaxy of bacteria (8345 genomes, 34,020,997 genes, 11,087,876,805 codons) with its center $\mathcal{GC}$ (3.1) in blue, its upper arm $\mathcal{GUA}$ (3.2) in green and its lower arm $\mathcal{GLA}$ (3.3) in violet, identified by the YZ-content of codon, and mainly by the GC-content. 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 YZ-content of codon in the 3 cases:
(A) GC-content;
(B) AG-content;
(C) GT-content (Color figure online)
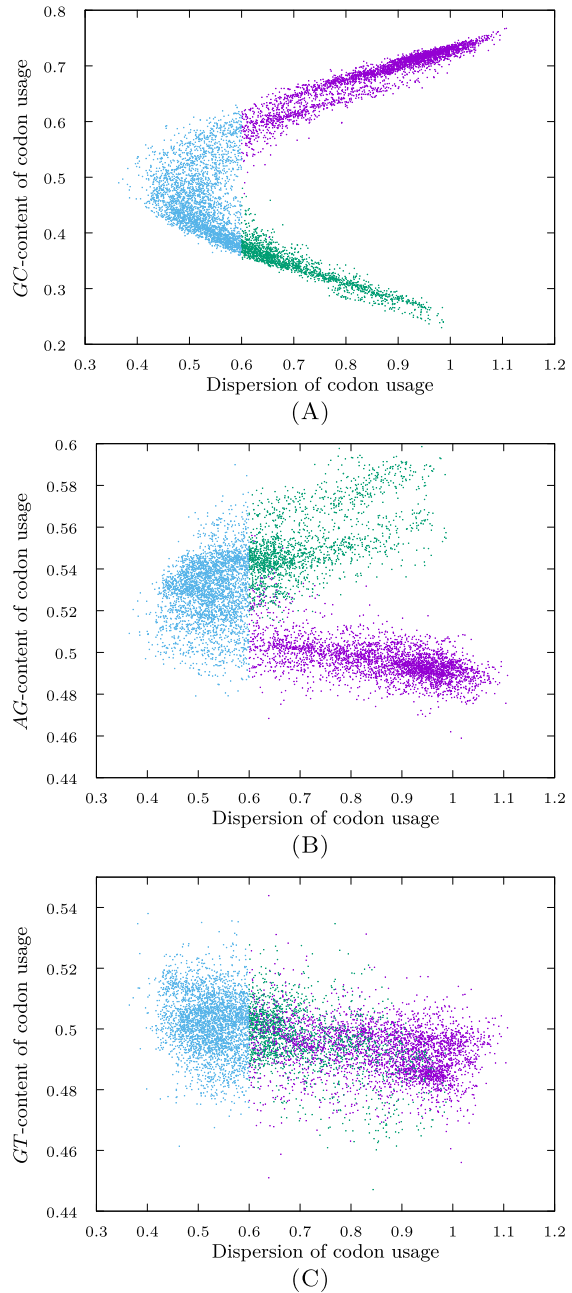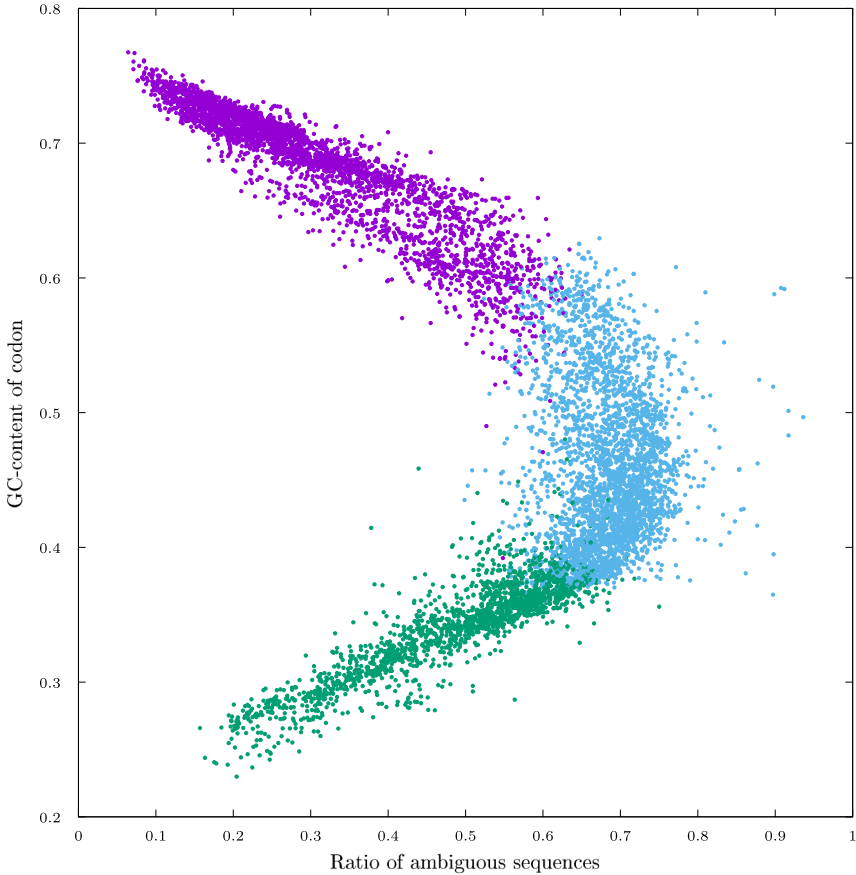
Fig. 6 Genome galaxy of bacteria (8345 genomes, 34,020,997 genes, 11,087,876,805 codons) with its center $\mathcal{GC}$ (3.1) in blue, its upper arm $\mathcal{GUA}$ (3.2) in green and its lower arm $\mathcal{GLA}$ (3.3) in violet. Each point represents all the genes of a bacterial genome. The x-axis shows the ratio r (2.7) of ambiguous sequences. The y-axis shows the GC-content of codon (Color figure online)

codon (Sect. 2.3). Figure 5A confirms that the GC-content better identifies the three structures of the galaxy compared to the AG-content (Fig. 5B) and the GT-content (Fig. 5C).

By expressing the GC-content of codon according to the ratio r (2.7) of ambiguous sequences, Fig. 6 is, as expected, a "mirror image" of Fig. 5A.

In order to further analyse the results with the GC-content, we express the gene reading frame retrieval (RFR) function f (2.4) according to GC-content in each of the 3 codon sites. The three galaxy structures are well characterized with the $GC3$-content in the 3rd codon site (Fig. 7C). The variation of the $GC3$-content covers almost the entire interval [0.1, 1]. The upper arm is related to genomes with a low $GC3$-content while the lower arm is related to genomes with a high $GC3$-content, the center being related to genomes with a $GC3$-content around 0.5.

Fig. 7 Genome galaxy of
bacteria (8345 genomes,
34,020,997 genes,
11,087,876,805 codons) with its
center $\mathcal{GC}$ (3.1) in blue, its upper
arm $\mathcal{GUA}$ (3.2) in green and its
lower arm $\mathcal{GLA}$ (3.3) in violet,
identified by the GC-content in
each of the 3 codon sites, and
mainly by the $GC3$-content.
Each point represents all the
genes of a bacterial genome.
The y-axis shows the reading
frame retrieval function f (2.4).
The x-axis shows the
GC-content in the 3 codon sites:
(A) $GC1$-content in the 1st
codon site;
(B) $GC2$-content in the 2nd
codon site;
(C) $GC3$-content in the 3rd
codon site (Color figure online)

(A)

(B)

(C)
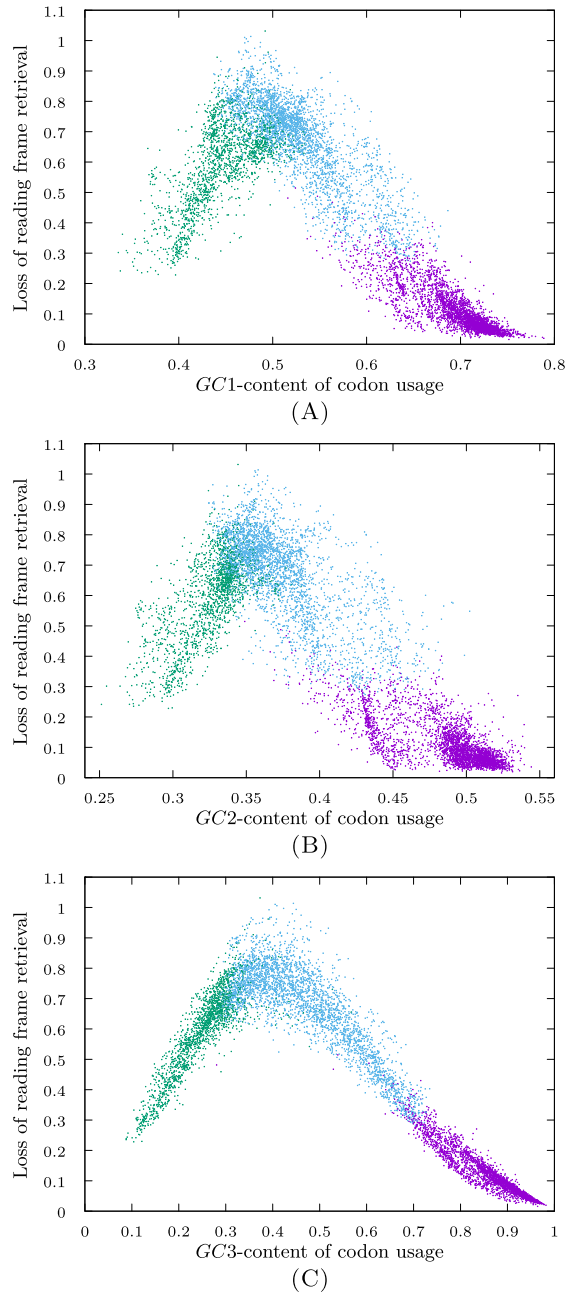
The variation of the $GC1$-content is restricted to the interval [0.30, 0.80]. The upper
and lower arms are still separated but to a lesser extent (Fig. 7A).

The variation of the $GC2$-content is restricted to the interval [0.25, 0.55]. The upper
and lower arms are separated but close (Fig. 7B).

In summary, the genome galaxy of bacteria with its center $\mathcal{GC}$, its upper arm $\mathcal{GUA}$ and its lower arm $\mathcal{GLA}$ is mainly related to the GC-content of codon, compared to the AG-content and the GT-content, and in particular to the $GC3$-content in the 3rd codon site, compared to the $GC1$-content and $GC2$-content.

4 A Mathematical Study of Codon Usage Dispersion

Codon usage, i.e. the association of probabilities (frequencies) (Definition 2.5) with the 64 codons, is a biological parameter that has been the most widely studied and published. Surprisingly, the dispersion of codon usage (Definition 2.7) is a basic statistical parameter which, as far as we know, has been completely ignored in biology. This section is divided into two parts. In the first part, we develop a mathematical theory of dispersion by deriving various formulæ. In the second part, we apply this theory to the evolution of the genetic code. In particular, we address the following problem. In the current genetic code, not all 64 codons code for amino acids. Indeed, there are 3 unused codons, precisely the 3 stop codons $\{TAA, TAG, TGA\}$, with a probability equal to (close to) 0. The hypothesis that the genetic code evolves with the appearance of codons, i.e. with a non-zero probability, over time will therefore be studied.

4.1 Formulæ of Codon Usage Dispersion

We study the dispersion of the weighted trinucleotide codes $\mathcal{W} = (X_g, \omega)$ over the genetic alphabet X_g that satisfy certain properties. According to Definitions 2.6 and 2.19, we must have $p_M(\mathcal{W}) \geqslant \frac{1}{n(\mathcal{W})}$ and $p_m(\mathcal{W}) \leqslant \frac{1}{n(\mathcal{W})}$. To ease the exposition, we consider an arbitrary enumeration $x_1, \ldots, x_{64}$ of X_g, and set $\omega_i = \omega(x_i)$. Accordingly, any weighted trinucleotide code over X_g can be viewed as the data of 64 non-negative real numbers summing to 1. We use this identification in the sequel.

We want to fix the number of trinucleotides that occur, the number of those of high frequency, and bounds on the minimum and maximum frequencies. To this end, we fix integers $n \in \{1, \ldots, 64\}$ and $n_h \in \{1, \ldots, n\}$, along with positive real numbers p_m, p_M such that $p_m \leqslant \frac{1}{n} \leqslant p_M$. Indeed, if each of the n trinucleotides that occur has frequency at most p_M, then $n \cdot p_M \geqslant 1$, since $n \cdot p_M \geqslant \sum_{i=1}^{64} \omega_i = 1$. Similarly, if each of the n trinucleotides that occur has frequency at least p_m, then necessarily $n \cdot p_m \leqslant 1$. Furthermore, at least one trinucleotide must have frequency at least $\frac{1}{n}$, so we always assume that $n_h \geqslant 1$. We note that for every $n_h \in \{1, \ldots, n-1\}$,

$$n \cdot p_m \leqslant 1 \Leftrightarrow p_m \leqslant \frac{1}{n} \Leftrightarrow (n - n_h) \cdot p_m \leqslant \frac{n - n_h}{n} \Leftrightarrow \frac{n_h}{n} + (n - n_h) \cdot p_m \leqslant 1.$$

We consider only trinucleotide codes satisfying the following three conditions:

(1) for every $i \in \{1, \ldots, n_h\}$ we have $\frac{1}{n} \leqslant \omega_i \leqslant p_M$;
(2) for every $i \in \{n_h + 1, \ldots, n\}$ we have $p_m \leqslant \omega_i \leqslant \frac{1}{n}$; and
(3) for every $i \in \{n + 1, \ldots, 64\}$ we have $\omega_i = 0$.

In other words, and referring to Definition 2.19, we require that

$$\mathrm{n}(\mathcal{W}) = n; \quad \mathrm{n_h}(\mathcal{W}) \geqslant n_h; \quad \text{and} \quad p_m \leqslant \mathrm{p_m}(\mathcal{W}) \leqslant \frac{1}{n} \leqslant \mathrm{p_M}(\mathcal{W}) \leqslant p_M.$$

Another observation is that if $p_m = \frac{1}{n}$, then Conditions (2) and (3) totally determine a unique weighted trinucleotide code, having $w_i = \frac{1}{n}$ for $i \in \{1, \dots, n\}$ and $\omega_i = 0$ for $i \in \{n+1, \dots, 64\}$. It is similar if $p_M = \frac{1}{n}$. We thus assume in the rest of this section that

$$p_m < \frac{1}{n} < p_M. \tag{4.1}$$

We are interested in determining the minimum value and the maximum value taken by the dispersion over all weighted trinucleotide codes satisfying Conditions (1)–(3). We derive closed formulæ for these — along with weighted trinucleotide codes attaining them.

Definition 4.1 Given Conditions (1)–(3) above and assuming (4.1), we define the maximum dispersion d_M and the minimum dispersion d_m as follows:

$$d_M = \mathrm{d_M}(n, n_h, p_m, p_M) = \sup \left\{ \mathrm{d}(\mathcal{W}) \; : \; \mathcal{W} = (\omega_1, \dots, \omega_{64}) \text{ satisfies } (1) - (3) \right\}; \tag{4.2}$$

and

$$d_m = \mathrm{d_m}(n, n_h, p_m, p_M) = \inf \left\{ \mathrm{d}(\mathcal{W}) \; : \; \mathcal{W} = (\omega_1, \dots, \omega_{64}) \text{ satisfies } (1) - (3) \right\}. \tag{4.3}$$

The determination of the minimum dispersion d_m is quick. It shows in particular that the value of $\mathrm{d_m}(n, n_h, p_m, p_M)$ does not depend on n_h.

PROPOSITION 4.2 *Suppose that Conditions (1)–(3) as above and (4.1) are satisfied. Then*

$$d_m = 2 \left[1 - \frac{n}{64} \right]. \tag{4.4}$$

The proof of Proposition 4.2 is given in Appendix B.

The determination of the maximum dispersion d_M is slightly more technical. We have the following proposition.

PROPOSITION 4.3 *Suppose that Conditions (1)–(3) as above and (4.1) are satisfied. If*

$$n_h \cdot p_M + (n - n_h) \cdot p_m \geqslant 1, \tag{4.5}$$

then

$$
d_M =
\begin{cases}
2\left[\left(p_m - \dfrac{1}{64}\right) \cdot n_h + 1 - n \cdot p_m\right] & \text{if } p_m \leqslant \tfrac{1}{64}, & (4.6a) \\[2ex]
2\left[1 - \dfrac{n}{64}\right] & \text{if } p_m > \tfrac{1}{64}. & (4.6b)
\end{cases}
$$

If (4.5) *does not hold, then with*

$$
\varepsilon = 1 - \left(n_h p_M + (n - n_h) p_m\right) > 0, \quad t_0 = \left\lfloor \frac{\varepsilon}{1/n - p_m} \right\rfloor,
$$

$$
\text{and} \quad z_1 = \varepsilon - t_0 \cdot \left(\frac{1}{n} - p_m\right),
$$

we have

$$
d_M =
\begin{cases}
2\left[\left(p_M - \dfrac{1}{64}\right) \cdot n_h + t_0\left(\dfrac{1}{n} - \dfrac{1}{64}\right)\right] & \text{if } p_m + z_1 \leqslant \tfrac{1}{64}, & (4.7a) \\[2ex]
2\left[\left(p_m - \dfrac{1}{64}\right)(n_h + t_0 + 1) + 1 - n \cdot p_m\right] & \text{if } p_m + z_1 > \tfrac{1}{64} \text{ and } p_m \leqslant \tfrac{1}{64}, & (4.7b) \\[2ex]
2\left[1 - \dfrac{n}{64}\right] & \text{if } p_m > \tfrac{1}{64}. & (4.7c)
\end{cases}
$$

The proof of Proposition 4.3 is given in Appendix C.

Instead of imposing only a lower bound on the number n_h of codons of high frequency, as Condition (2) does, one could rather impose the exact number of such codons. In this case, the optimal values given by Proposition 4.3 are not necessarily always attained by a weighted trinucleotide code anymore, but are still the supremum of the possible values, as we explain in Appendix D.

It is important to stress that this mathematical theory and its formulæ of dispersion, i.e. Equations (4.4), (4.6a)–(4.6b) and (4.7a)–(4.7c), can be easily extended to any weighted codes over a finite alphabet, e.g. the amino acid alphabet, by replacing $1/64$ with the cardinality of the code.

4.2 Properties of the Genetic Code

In order to study the evolution of the bacterial genetic code and to apply the mathematical theory developed in Sect. 4.1, three parameters must be obtained from the bacterial codon usage (see Appendix Table 1): the minimum probability, which is $p_m = 0.06\%$, given by the codon of lowest occurrence, i.e. TAG; the maximum probability, which is $p_M = 4.32\%$, given by the codon of highest occurrence, i.e. GCC; and the number of codons with high frequency, i.e. with a probability greater than $\frac{1}{64}$, which is $n_h = 25$. Note that $\frac{n_h}{n} = \frac{25}{64}$ since there are $n = 64$ occurring codons in the bacterial codon usage.

We start by studying the maximum dispersion of the current genetic code over 64 codons with respect to the frequency of its codon of maximum and minimum probabilities.
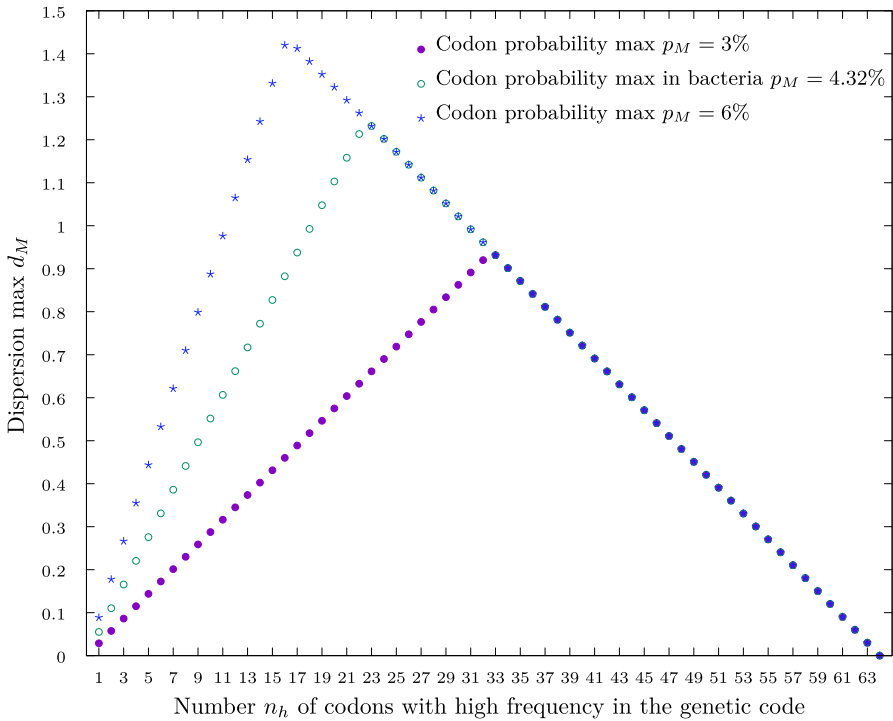
4.2.1 Maximum Dispersion of the Current Genetic Code as a Function of Its Codon of Maximum Probability

Figure 8 gives the maximum dispersion d_M of the current genetic code with 64 codons as a function of its number n_h of codons with high frequency, $n_h \in \{1, \ldots, 64\}$, i.e. codons x with a probability at least $\frac{1}{n} = \frac{1}{64}$ ($\omega(x) \geqslant \frac{1}{64}$). The minimum probability is $p_m = 0.06\%$, given by the codon of lowest occurrence in bacteria, i.e. TAG (see Appendix Table 1). We considered 3 different values for the maximum probability p_M: 4.32% associated with the codon of highest occurrence in bacteria, i.e. GCC (see Appendix Table 1) (curve with green circles) and 2 arbitrarily chosen surrounding values: 3% (curve with violet disks) and 6% (curve with blue stars). Interestingly, for any maximum probability p_M the maximum dispersion d_M increases to a maximum and then decreases to 0 at $n_h = 64$. The curves have a common decreasing slope of equation $y = 1.9232 - 0.03005x$ (see Equation (4.6a)), which does not involve p_M), but different increasing slopes, and thus different maxima. The change occurs for $n_h = \left\lceil \frac{1-64p_m}{p_M - p_m} \right\rceil$ (see Equation (4.5)), and the maximum is thus attained for $n_h = \left\lfloor \frac{1-64p_m}{p_M - p_m} \right\rfloor$ or $n_h = \left\lceil \frac{1-64p_m}{p_M - p_m} \right\rceil$.

From Appendix Table 1, the number of codons with high frequency is $n_h = 25$. The maximum dispersion at $n_h = 25$ is $d_M = 1.17$. Very interestingly, this value $d_M = 1.17$ is very close to the maximum $d_M = 1.23$ at $n_h = 23$ of the curve with $p_M = 4.32\%$ that is associated with bacteria (see Fig. 8). As a biological consequence, bacteria have optimised the number of codons with high frequency in order to have a maximum of dispersion, and thus a maximum capacity to retrieve the reading frame in genes.

4.2.2 Maximum Dispersion of the Current Genetic Code as a Function of Its Codon of Minimum Probability

Figure 9 gives the maximum dispersion d_M of the current genetic code with 64 codons as a function of its number n_h of codons with high frequency. The maximum probability is $p_M = 4.32\%$, given by the codon of highest occurrence in bacteria, i.e. GCC (see Appendix Table 1). We considered 3 different values for the minimum probability p_m: 0.06% associated with the codon of lowest occurrence in bacteria, i.e. TAG (see Appendix Table 1) (curve with green circles) and 2 arbitrarily chosen surrounding values: 0 (curve with violet disks) and 0.5% (curve with blue stars). The value 0 is seen as a sort of limiting case for codons with very small frequencies, that is, when $p_m(\mathcal{W})$ is a very small positive real. In theory, setting $p_m = 0$ amounts to allowing the number of occurring codons to vary (case studied in Sect. 4.2.3), whereas we keep $n = 64$ here. As in the previous case (Sect. 4.2.1), for any minimum probability p_m the maximum dispersion d_M increases to a maximum and then decreases to 0 at $n_h = 64$.
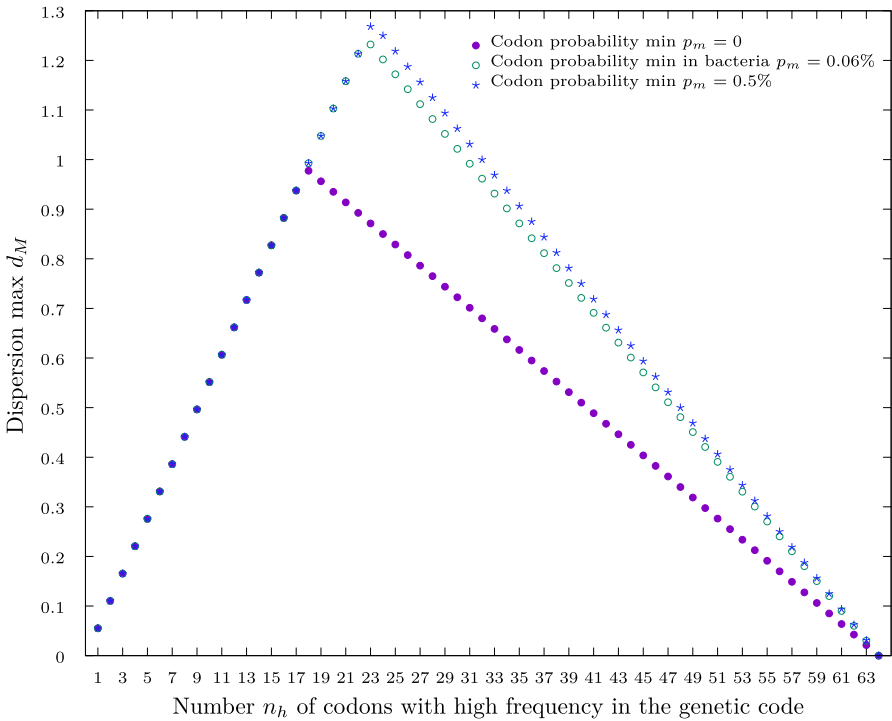
Fig. 8 Maximum dispersion d_M of the current genetic code with 64 codons for three values of the codon of maximum probability p_M, the minimum probability $p_m = 0.06\%$ being that of the codon TAG in bacteria (see Appendix Table 1). The x-axis shows the number n_h of codons with high frequency in the genetic code, $n_h \in \{1, \ldots, 64\}$, i.e. codons x with a probability at least $\frac{1}{64}$ ($\omega(x) \geqslant \frac{1}{64}$). The y-axis shows the maximum dispersion d_M (see Proposition 4.3) of codon usage. Three curves corresponding to different values of the codon of maximum probability are displayed: $p_M = 4.32\%$ associated with the codon GCC in bacteria (see Appendix Table 1) (curve with green circles) and 2 arbitrarily chosen surrounding values, namely $p_M = 3\%$ (curve with violet disks) and $p_M = 6\%$ (curve with blue stars) (Color figure online)

However, the curves have a common increasing slope of equation $y = 0.05515x$ (see Equation (4.7a)), but different decreasing slopes, and thus different maxima.

The curve with $p_m = 0.06\%$ and $p_M = 4.32\%$ is obviously identical in the two Sects. 4.2.1 and 4.2.2 and the two Figs. 8 and 9. Very interestingly, with $p_M = 4.32\%$, the curve with $p_m = 0.06\%$ that is associated with bacteria is very close to the curve with $p_m = 0$ giving the highest value for d_M. Indeed, at $n_h = 25$ the maximum dispersion is equal to $d_M = 1.17$ with $p_m = 0.06\%$ and to $d_M = 1.22$ with $p_m = 0$. As previously, bacteria have optimised the maximum dispersion, and thus a maximum capacity to retrieve the reading frame in genes.

4.2.3 Minimum and Maximum Dispersions of the Evolutionary Bacterial Genetic Code as a Function of Its Number of Codons

Figure 10 gives the dispersion of the evolutionary bacterial genetic code as a function of its number n of codons, $n \in \{1, \ldots, 64\}$, i.e. codons x with a probability $\omega(x)$ greater

Fig. 9 Maximum dispersion d_M of the current genetic code with 64 codons for three values of the minimum probability p_m, the maximum probability $p_M = 4.32\%$ being that of the codon GCC in bacteria (see Appendix Table 1). The x-axis shows the number n_h of codons with high frequency in the genetic code, $n_h \in \{1, \ldots, 64\}$, i.e. codons x with a probability at least $\frac{1}{64}$ ($\omega(x) \geqslant \frac{1}{64}$). The y-axis shows the maximum dispersion d_M (see Proposition 4.3) of codon usage. Three curves corresponding to different values of the codon of minimum probability are displayed: $p_m = 0.06\%$ associated with the codon TAG in bacteria (see Appendix Table 1) (curve with green circles) and 2 arbitrarily chosen surrounding values, namely $p_m = 0$ (curve with violet disks) and $p_m = 0.5\%$ (curve with blue stars) (Color figure online)

than 0. Since n varies from 1 to 64, and we must have $\frac{1}{n} \leqslant p_M$, the parameter p_M cannot stay equal to its original value of 4.32% from the bacterial genome with 64 codons: we normalise it by setting $p_M = \frac{4.32}{100} \cdot \frac{64}{n}$. It then seems natural to similarly scale the minimum probability p_m instead of keeping it at the bacterial value 0.06%, and we do so by setting $p_m = \frac{0.06}{100} \cdot \frac{n}{64}$. Analogously, we make the parameter n_h vary with n, keeping the ratio $\frac{n_h}{n}$ as close as possible to the bacterial value $\frac{25}{64} \approx 0.39$. Formally, we set n_h in $\{1, \ldots, n\}$ to be either $\left\lfloor \frac{25n}{64} \right\rfloor$ or $\left\lceil \frac{25n}{64} \right\rceil$, depending on which value is closer to $\frac{25n}{64}$ (and discarding the value 0, i.e., setting $n_h = 1$ when $n = 1$, since as reported earlier we can always assume that $n_h \geqslant 1$).

The dispersion function d (2.1) of codon usage ranges in the interval $[0, \frac{63}{32}] \approx [0, 1.97]$ (Proposition 2.8). The minimum dispersion d_m (curve with violet disks in Fig. 10) has the maximum value $\frac{63}{32} \approx 1.97$ at $n = 1$ and the minimum value 0 at $n = 64$. It decreases along the straight line with equation $y = 2 - \frac{x}{32}$ (see Equation (4.4)). The maximum dispersion d_M (curve with green circles in Fig. 10) has the maximum
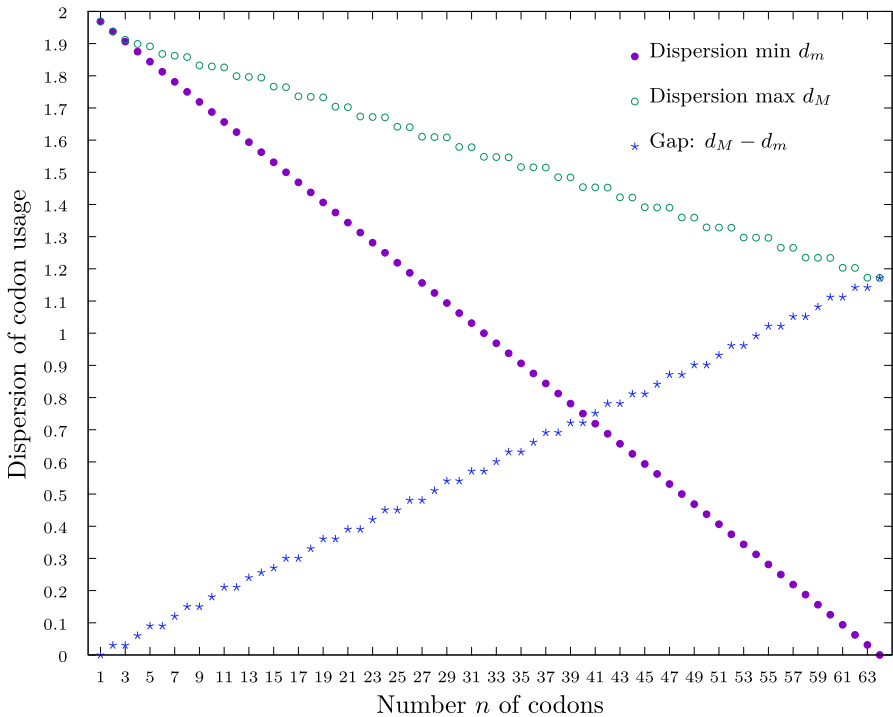
Fig. 10 Dispersion of the evolutionary bacterial genetic code. From the bacterial codon usage (Appendix Table 1), three parameters are used: the (normalised) minimum probability is $p_m = 0.06 \cdot \frac{64}{n} \%$, based on the frequency of the codon of lowest occurrence in the bacterial genome, i.e. TAG; the (normalised) maximum probability is $p_M = 4.32 \cdot \frac{64}{n} \%$, based on the frequency of the codon of highest occurrence, i.e. GCC; and the (normalised) number n_h of codons with high frequency, i.e. with a probability at least $\frac{1}{64}$, defined as the positive integer that is closest to $25 \cdot \frac{n}{64}$. The x-axis shows the number n of codons in the genetic code, $n \in \{1, \ldots, 64\}$, i.e. the codons x with a probability greater than 0 ($\omega(x) > 0$). The y-axis shows the dispersion function d (2.1) of codon usage in the range $[0, \frac{63}{32}] \approx [0, 1.97]$ (Proposition 2.8): the curve with violet disks is the minimum dispersion d_m (see Proposition 4.2), the curve with green circles is the maximum dispersion d_M (see Proposition 4.3) and the curve with blue stars is the dispersion difference $\Delta = d_M - d_m$ (Color figure online)

value $\frac{63}{32} \approx 1.97$ at $n = 1$, as d_m, and the minimum value 1.17195 at $n = 64$. It decreases with n, approximately along the straight line with equation $y = 1.948 - 0.012x$ (see Equation (4.6a)), recalling that the values of p_m and n_h vary with n). Thus, the dispersion difference $\Delta = d_M - d_m$ has the minimum value 0 at $n = 1$ and the maximum value 1.17195 at $n = 64$.

From a biological point of view, this theoretical result quantifies and explains that the more codons a code contains, the greater the dispersion and the greater the capacity of genes to retrieve the reading frame, or equivalently the lower the loss of reading frame retrieval (see the RFR function f (2.4) in Fig. 1).

5 Conclusion

The reading frame retrieval (RFR) function f of genes expressed as a function of the codon usage dispersion d (from the uniform codon distribution 1/64), identifies a genome galaxy in bacteria (Fig. 1). A simple probabilistic model for ambiguous sequences shows that the RFR function is a measure of the gene reading frame retrieval. Indeed, the RFR function increases with the ratio of ambiguous sequences and the ratio of ambiguous sequences decreases when the codon usage dispersion increases. This genome galaxy is studied by expressing the RFR function f according to the YZ-content (GC-, AG- and GT-content) of codon in bacteria (Fig. 4). The classical GC-content is the best parameter for characterizing the upper arm, which is related to genomes with a low GC-content, and the lower arm, which is related to genomes with a high GC-content. The galaxy center has a GC-content around 0.5 (Fig. 4A). Then, these results are confirmed by expressing the GC-content of codon as a function of the codon usage dispersion (Fig. 5A). Finally, the bacterial genome galaxy is better described with the $GC3$-content in the 3rd codon site compared to the $GC1$-content and $GC2$-content in the 1st and 2nd codons sites, respectively (Fig. 7C).

Whereas the codon usage is used extensively by biologists, its dispersion is surprisingly little known and unused, even though it is a classical parameter in basic statistics. With this in mind, and also to study the genome galaxy, we have developed here a mathematical theory of codon usage dispersion by deriving several formulæ. It shows three important parameters that should be considered by biologists: the codon of highest frequency (i.e. the parameter p_M), the codon of lowest frequency (i.e. the parameter p_m) and the number of codons with high frequency, i.e. greater than $100/64 = 1.5625\%$ (i.e. the parameter n_h). The derived formulæ of dispersion can be easily extended to any weighted codes over a finite alphabet, e.g. the amino acid alphabet. The theory developed shows that bacteria have optimised the codon dispersion to be maximal, and thus a maximum capacity to retrieve the reading frame in genes, in two ways: (i) the existence of stop codons with frequencies $p_m \approx 0$, e.g. the bacterial codon TAG of lowest frequency $p_m = 0.06\%$ (Fig. 9); (ii) the number n_h of codons with high frequency, e.g. $n_h = 25$ is almost optimal with the bacterial codon GCC of highest occurrence $p_M = 4.32\%$ (Fig. 8).

Appendix A: Codon Usage of Bacterial Kingdom

See Table 1.

Table 1 Codon usage of 34,020,997 genes of 8345 bacterial genomes (11,087,876,805 codons) obtained from the codon statistics database (CSD) (Subramanian et al. 2022)

AAA	2.37	CAA	1.20	GAA	2.91	TAA	0.12
AAC	1.80	CAC	1.27	GAC	3.42	TAC	1.56
AAG	2.00	CAG	2.16	GAG	3.22	TAG	0.06
AAT	1.64	CAT	0.78	GAT	2.25	TAT	1.35
ACA	0.93	CCA	0.63	GCA	1.43	TCA	0.64
ACC	2.56	CCC	1.39	GCC	4.32	TCC	1.25
ACG	1.51	CCG	2.14	GCG	3.22	TCG	1.23
ACT	0.76	CCT	0.67	GCT	1.30	TCT	0.67
AGA	0.54	CGA	0.43	GGA	1.39	TGA	0.15
AGC	1.28	CGC	2.26	GGC	3.63	TGC	0.53
AGG	0.38	CGG	1.64	GGG	1.43	TGG	1.31
AGT	0.69	CGT	0.80	GGT	1.54	TGT	0.29
ATA	0.82	CTA	0.45	GTA	1.05	TTA	1.25
ATC	2.76	CTC	2.23	GTC	2.66	TTC	2.11
ATG	2.15	CTG	3.86	GTG	2.66	TTG	1.21
ATT	1.91	CTT	1.01	GTT	1.20	TTT	1.61

Appendix B: Proof of Proposition 4.2

We start by defining a weighted trinucleotide code $\mathcal{W} = (\omega_1, \ldots, \omega_{64})$ that satisfies Conditions (1)–(3) by setting $\omega_i = \frac{1}{n}$ if $1 \leqslant i \leqslant n$, and $\omega_i = 0$ if $n_h + 1 \leqslant i \leqslant 64$. Then $\mathcal{W}$ readily satisfies Conditions (1)–(3). Moreover, $d(\mathcal{W}) = 2 - \frac{n}{32}$. Indeed, if $i \in \{1, \ldots, n\}$ then $\omega_i = \frac{1}{n} \geqslant \frac{1}{64}$, so $\sum_{i=1}^{n} \left| \omega_i - \frac{1}{64} \right| = 1 - \frac{n}{64}$; moreover if $i \in \{n+1, \ldots, 64\}$ then $\omega_i = 0$ so $\sum_{i=n+1}^{64} \left| \omega_i - \frac{1}{64} \right| = \frac{64-n}{64} = 1 - \frac{n}{64}$.

It remains to show that if a weighted trinucleotide code $\mathcal{W}' = (\omega'_1, \ldots, \omega'_{64})$ satisfies Conditions (1)–(3), then $d(\mathcal{W}') \geqslant d(\mathcal{W})$. First, note that if $n = 64$, then $d(\mathcal{W}) = 0$, which is trivially minimum. So we now assume that $n \leqslant 63$, and we proceed as follows. Since the function d_m is the minimization of a convex function over a convex domain, it suffices to prove that $d(\mathcal{W})$ is a local minimum, that is, if a weighted trinucleotide code $\mathcal{W}'$ still satisfying Conditions (1)–(3) is produced by small enough modifications of the values $\omega_1, \ldots, \omega_{64}$, then $d(\mathcal{W}') \geqslant d(\mathcal{W}) = 2 - \frac{n}{32}$.

Fix $\varepsilon > 0$ small enough that $\frac{1}{64} + \varepsilon < \frac{1}{n}$, which is possible since $n \leqslant 63$ by our assumption. Now, if $\left| \omega'_i - \omega_i \right| \leqslant \varepsilon$ for $i \in \{1, \ldots, 64\}$, then since $\mathcal{W}'$ satisfies Condition (1) we have $\omega'_i \geqslant \omega_i = \frac{1}{n} > \frac{1}{64}$ if $1 \leqslant i \leqslant n_h$, and $\omega'_i \geqslant \omega_i - \varepsilon = \frac{1}{n} - \varepsilon > \frac{1}{64}$ if $n_h + 1 \leqslant i \leqslant n$. So $\left| \omega'_i - \frac{1}{64} \right| = \omega'_i - \frac{1}{64}$ for $i \in \{1, \ldots, n\}$. Consequently,

$$\sum_{i=1}^{n} \left| \omega'_i - \frac{1}{64} \right| = \sum_{i=1}^{n} \left(\omega'_i - \frac{1}{n} \right) + \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{1}{64} \right). \tag{B.1}$$

As a result, using that $\omega_i' = 0 = \omega_i$ if $n + 1 \leqslant i \leqslant 64$ since $\mathcal{W}'$ must satisfy Condition (3), we infer from (B.1) that

$$d(\mathcal{W}') = \sum_{i=1}^{n} \left(\omega_i' - \frac{1}{n} \right) + d(\mathcal{W})$$

$$= d(\mathcal{W}) - 1 + \sum_{i=1}^{n} \omega_i'$$

$$= d(\mathcal{W}).$$

This concludes the proof.   □

Appendix C: Proof of Proposition 4.3

Suppose first that (4.5) **holds.** Let $z_0 = \frac{1 - n \cdot p_m}{n_h}$, and note that $z_0 \geqslant 0$ by (4.1). We set

$$\forall i \in \{1, \ldots, 64\}, \quad \omega_i = \begin{cases} p_m + z_0 & \text{if } 1 \leqslant i \leqslant n_h, \\ p_m & \text{if } n_h + 1 \leqslant i \leqslant n, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathcal{W} = (\omega_1, \ldots, \omega_{64})$ defines a weighted trinucleotide code, since it consists of non-negative real numbers summing to 1. This code $\mathcal{W}$ does satisfy the required properties. The only condition that could be violated is (1). However, $\frac{(n-n_h)p_m}{n_h} \geqslant \frac{1}{n_h} - p_M$ by (4.5), so $p_m + z_0 = \frac{1-(n-n_h) \cdot p_m}{n_h} \leqslant p_M$ on the one hand, and $\frac{(n-n_h)p_m}{n_h} \leqslant \frac{1}{n_h} - \frac{1}{n}$ by (4.1) so $p_m + z_0 \geqslant \frac{1}{n}$ on the other hand. Consequently, $d_M \geqslant d(\mathcal{W})$. Moreover, $d(\mathcal{W})$ is equal to the right side of (4.6a) if $p_m \leqslant \frac{1}{64}$, and to the right side of (4.6b) otherwise. To see this, we define $d_1 = \sum_{i=1}^{n_h} \left| \omega_i - \frac{1}{64} \right|$ and $d_2 = \sum_{i=n_h+1}^{n} \left| \omega_i - \frac{1}{64} \right|$. If $i \in \{1, \ldots, n_h\}$, then $\omega_i = p_m + z_0 \geqslant \frac{1}{n} \geqslant \frac{1}{64}$ so

$$d_1 = n_h \cdot \left(p_m + z_0 - \frac{1}{64} \right) = 1 - (n - n_h) \cdot p_m - \frac{n_h}{64}$$

$$= 1 + n_h \cdot \left(p_m - \frac{1}{64} \right) - n \cdot p_m; \tag{C.1}$$

while if $i \in \{n_h + 1, \ldots, n\}$, then $\omega_i = p_m$, so if $p_m \leqslant \frac{1}{64}$,

$$d_2 = (n - n_h) \cdot \left(\frac{1}{64} - p_m \right) = \frac{n}{64} + n_h \cdot \left(p_m - \frac{1}{64} \right) - n \cdot p_m, \tag{C.2}$$

and if $p_m > \frac{1}{64}$,

$$d_2 = (n - n_h) \cdot \left(p_m - \frac{1}{64} \right) = -\frac{n}{64} - n_h \cdot \left(p_m - \frac{1}{64} \right) + n \cdot p_m; \quad \text{(C.3)}$$

finally, if $i \in \{n + 1, \ldots, 64\}$, then $\omega_i = 0$ and hence $\sum_{i=n+1}^{64} \left| \omega_i - \frac{1}{64} \right| = 1 - \frac{n}{64}$. Summing this with (C.1) and either (C.2) or (C.3) yields (4.6a) or (4.6b), respectively.

It remains to prove that if a weighted trinucleotide code $\mathcal{W}' = (\omega'_1, \ldots, \omega'_{64})$ satisfies Conditions (1)–(3), then $\mathrm{d}(\mathcal{W}') \leqslant \mathrm{d}(\mathcal{W})$. By Condition (3), it suffices to prove that $\sum_{i=1}^n \left| \frac{1}{64} - \omega'_i \right| \leqslant \sum_{i=1}^n \left| \frac{1}{64} - \omega_i \right|$. For convenience we set $S_1 = \sum_{i=1}^{n_h} \omega_i$ and $S_2 = \sum_{i=n_h+1}^n \omega_i$. We similarly define S'_1, S'_2 as well as d'_1 and d'_2 with respect to $\mathcal{W}'$ instead of $\mathcal{W}$. Our goal is thus to show that $d'_1 + d'_2 \leqslant d_1 + d_2$. Note that $S_1 + S_2 = 1 = S'_1 + S'_2$.

As $\omega'_i \geqslant \frac{1}{64}$ for $i \in \{1, \ldots, n_h\}$, we deduce that $d'_1 = S'_1 - \frac{n_h}{64} = 1 - S'_2 - \frac{n_h}{64}$, and hence (C.1) implies that $d'_1 - d_1 = (n - n_h) \cdot p_m - S'_2$.

To express d'_2, let

$$I_2^+ = \left\{ i \in \{n_h + 1, \ldots, n\} : \omega'_i > \frac{1}{64} \right\},$$

and $I_2^- = \{n_h + 1, \ldots, n\} \setminus I_2^+$, thus $\left| I_2^+ \right| = (n - n_h) - \left| I_2^- \right|$ and $\sum_{i \in I_2^+} \omega'_i = S'_2 - \sum_{i \in I_2^-} \omega'_i$. Consequently,

$$\begin{aligned}
d'_2 &= \sum_{i=n_h+1}^n \left| \omega'_i - \frac{1}{64} \right| = \sum_{i \in I_2^+} \left(\omega'_i - \frac{1}{64} \right) + \sum_{i \in I_2^-} \left(\frac{1}{64} - \omega'_i \right) \\
&= -\frac{1}{64} \cdot \left(\left| I_2^+ \right| - \left| I_2^- \right| \right) + \sum_{i \in I_2^+} \omega'_i - \sum_{i \in I_2^-} \omega'_i \\
&= -\frac{1}{64} \cdot \left((n - n_h) - 2 \left| I_2^- \right| \right) + S'_2 - 2 \sum_{i \in I_2^-} \omega'_i \\
&\leqslant -\frac{1}{64} \cdot \left((n - n_h) - 2 \left| I_2^- \right| \right) + S'_2 - 2 \sum_{i \in I_2^-} p_m \\
&= 2 \left| I_2^- \right| \left(\frac{1}{64} - p_m \right) + S'_2 - \frac{n - n_h}{64},
\end{aligned}$$

where the inequality holds because $\omega'_i \geqslant p_m$ for each $i \in I_2^-$. Furthermore, if $p_m > \frac{1}{64}$, then $d_2 = (n - n_h) \cdot (p_m - \frac{1}{64})$, and hence

$$\begin{aligned}
d'_1 + d'_2 - (d_1 + d_2) &= d'_1 - d_1 + d'_2 - d_2 \\
&= 2 \left| I_2^- \right| \left(\frac{1}{64} - p_m \right) \leqslant 0,
\end{aligned}$$

since $p_m > \frac{1}{64}$.

Finally, if $p_m \leqslant \frac{1}{64}$, then $d_2 = (n - n_h) \cdot (\frac{1}{64} - p_m)$, and therefore,

$$
\begin{aligned}
d_1' + d_2' - (d_1 + d_2) &= d_1' - d_1 + d_2' - d_2 \\
&\leqslant 2(n - n_h) \cdot \left(p_m - \frac{1}{64} \right) + 2 \left| I_2^- \right| \cdot \left(\frac{1}{64} - p_m \right) \\
&= 2 \left(p_m - \frac{1}{64} \right) \cdot \left((n - n_h) - \left| I_2^- \right| \right).
\end{aligned}
$$

It follows that $d_1' + d_2' \leqslant d_1 + d_2$, because $p_m - \frac{1}{64} \leqslant 0$ on the one hand, and $(n - n_h) - \left| I_2^- \right| \geqslant 0$ on the other hand since I_2^- is a subset of $\{n_h + 1, \ldots, n\}$.

Suppose now that (4.5) **does not hold, so**

$$
n_h \cdot p_M + (n - n_h) \cdot p_m < 1. \tag{C.4}
$$

This implies that $p_m < \frac{1}{n}$. As in the statement of the proposition, we set $\varepsilon = 1 - (n_h \cdot p_M + (n - n_h) \cdot p_m)$ and $t = \frac{\varepsilon}{1/n - p_m}$, so ε and t are positive; and we also set $t_0 = \lfloor t \rfloor$ and $z_1 = \varepsilon - t_0 \cdot (\frac{1}{n} - p_m)$, so $z_1 \geqslant 0$. We note that $n_h + t_0 + 1 \leqslant n$, for otherwise we infer that $\frac{1}{n} \geqslant p_M$, contrary to (4.1). Indeed, if $n_h + t_0 \geqslant n$, then $n_h + t \geqslant n$, that is,

$$
n_h + \frac{1 - n_h p_M - n p_m + n_h p_m}{\frac{1}{n} - p_m} \geqslant n
$$

so

$$
\frac{n_h}{n} - n_h p_m + 1 - n_h p_M - n p_m + n_h p_m \geqslant 1 - n p_m,
$$

which yields that $\frac{n_h}{n} - n_h p_M \geqslant 0$ and hence $\frac{1}{n} \geqslant p_M$ since $n_h > 0$.

Then setting

$$
\forall i \in \{1, \ldots, 64\}, \quad \omega_i = \begin{cases} p_M & \text{if } 1 \leqslant i \leqslant n_h, \\ \frac{1}{n} & \text{if } n_h + 1 \leqslant i \leqslant n_h + t_0, \\ p_m + z_1 & \text{if } i = n_h + t_0 + 1, \\ p_m & \text{if } n_h + t_0 + 2 \leqslant i \leqslant n, \\ 0 & \text{otherwise,} \end{cases}
$$

defines a weighted trinucleotide code, since the sequence $\mathcal{W} = (\omega_1, \ldots, \omega_{64})$ consists of non-negative real numbers summing to 1. Conditions (1)–(3) can be violated only if $p_m + z_1 > \frac{1}{n}$. However, since $t_0 = \lfloor t \rfloor > t - 1$,

$$
p_m + z_1 = p_m + \varepsilon - t_0 \left(\frac{1}{n} - p_m \right) < \varepsilon - t \left(\frac{1}{n} - p_m \right) + \frac{1}{n} = \frac{1}{n}. \tag{C.5}
$$

Moreover, if $p_m \leqslant \frac{1}{64}$ then $d(\mathcal{W})$ equals the right side of (4.7a) or of (4.7b), depending on whether $p_m + z_1 \leqslant \frac{1}{64}$ or not, while if $p_m > \frac{1}{64}$ then $d(\mathcal{W})$ equals the right side of (4.7c). This follows from the definitions by a direct computation, recalling that $n_h \, p_M = 1 - (n - n_h) p_m - \varepsilon$. Indeed, if $i \in \{1, \ldots, n_h + t_0\}$ then $\left| \omega_i - \frac{1}{64} \right| = \omega_i - \frac{1}{64}$ and hence

$$\sum_{i=1}^{n_h+t_0} \left| \omega_i - \frac{1}{64} \right| = \left(p_M - \frac{1}{64} \right) \cdot n_h + \left(\frac{1}{n} - \frac{1}{64} \right) \cdot t_0. \tag{C.6}$$

If $p_m \leqslant \frac{1}{64}$, then for $i \in \{n_h + t_0 + 2, \ldots, 64\}$ we have $\left| \omega_i - \frac{1}{64} \right| = \frac{1}{64} - \omega_i$, and hence

$$\sum_{i=n_h+t_0+2}^{64} \left| \omega_i - \frac{1}{64} \right| = (n_h + t_0 + 1 - n) \left(p_m - \frac{1}{64} \right) + 1 - \frac{n}{64}, \tag{C.7}$$

while if $p_m > \frac{1}{64}$, then $\left| \omega_i - \frac{1}{64} \right| = p_m - \frac{1}{64}$ for $i \in \{n_h + t_0 + 2, \cdots, n\}$, and $\left| \omega_i - \frac{1}{64} \right| = \frac{1}{64}$ for $i \in \{n + 1, \ldots, 64\}$, so that

$$\sum_{i=n_h+t_0+2}^{64} \left| \omega_i - \frac{1}{64} \right| = (n_h + t_0 + 1 - n) \left(\frac{1}{64} - p_m \right) + 1 - \frac{n}{64}. \tag{C.8}$$

Now if $p_m + z_1 > \frac{1}{64}$ then

$$\left| \omega_{n_h+t_0+1} - \frac{1}{64} \right| = (n_h + t_0 + 1 - n) \cdot p_m - n_h \cdot p_M - \frac{t_0}{n} + 1 - \frac{1}{64}, \tag{C.9}$$

so the sum of (C.6), (C.7) and (C.9) shows that $d(\mathcal{W})$ equals the right side of (4.7b) if $p_m \leqslant \frac{1}{64}$, while the sum of (C.6), (C.8) and (C.9) shows that $d(\mathcal{W})$ equals the right side of (4.7c) if $p_m > \frac{1}{64}$. Moreover, if on the contrary $p_m + z_1 \leqslant \frac{1}{64}$, and in particular $p_m \leqslant \frac{1}{64}$ as $z_1 \geqslant 0$, then

$$\left| \omega_{n_h+t_0+1} - \frac{1}{64} \right| = -(n_h + t_0 + 1 - n) \cdot p_m + n_h \cdot p_M + \frac{t_0}{n} - 1 + \frac{1}{64}, \tag{C.10}$$

so the sum of (C.6), (C.7) and (C.10) shows that $d(\mathcal{W})$ equals the right side of (4.7a).

We now prove that $\mathcal{W}$ has the largest dispersion among all weighted trinucleotide codes satisfying Conditions (1)–(3). To this end, let $\mathcal{W}' = (\omega_1', \ldots, \omega_{64}')$ be such a code, that moreover maximises the dispersion over all weighted trinucleotide codes satisfying Conditions (1)–(3).

We first show that we can assume that $\omega_i' = p_M$ for each $i \in \{1, \ldots, n_h\}$. Indeed, suppose that there exists $i \in \{1, \ldots, n_h\}$ such that $\omega_i' = p_M - \delta$ for some positive

real δ. Because $\omega'_k \leqslant p_M = \omega_k$ for any $k \in \{1, \ldots, n_h\}$ and $\sum_{i=1}^{n} \omega'_i = 1 = \sum_{i=1}^{n} \omega_i$, we infer the existence of a set $J \subseteq \{n_h + 1, \ldots, n\}$ such that

- $\sum_{j \in J} \omega'_j \geqslant \delta + \sum_{j \in J} \omega_j$; and
- $\omega'_j > \omega_j$ for each $j \in J$.

Since $p_m \leqslant \omega_j < \omega'_j \leqslant \frac{1}{n}$ for $j \in J$, we can define a weighted trinucleotide code $\mathcal{Z} = (\zeta_1, \ldots, \zeta_{64})$ satisfying Conditions (1)–(3) by setting $\zeta_i = p_M$, next $\zeta_k = \omega'_k$ if $k \notin J \cup \{i\}$, and finally defining ζ_j for $j \in J$ such that $\zeta_j \leqslant \omega'_j$ and $\sum_{j \in J} \zeta_j = \left(\sum_{j \in J} \omega'_j\right) - \delta$. Now, $\sum_{k \notin J} \left|\zeta_k - \frac{1}{64}\right| = \delta + \sum_{k \notin J} \left|\omega'_k - \frac{1}{64}\right|$ and $\sum_{j \in J} \left|\zeta_j - \frac{1}{64}\right| \geqslant -\delta + \sum_{j \in J} \left|\omega'_j - \frac{1}{64}\right|$, so that $\mathrm{d}(\mathcal{Z})$ is at least, and hence equal to, $\mathrm{d}(\mathcal{W}')$. Consequently, we can now assume that $\omega'_i = p_M$ for $i \in \{1, \ldots, n_h\}$.

We similarly show that we can assume that $\omega'_i = \frac{1}{n}$ if $n_h + 1 \leqslant i \leqslant n_h + t_0$. Indeed, suppose that there exists $i \in \{n_h + 1, \ldots, n_h + t_0\}$ such that $\omega'_i = \frac{1}{n} - \delta$ for some positive real δ. Because $\omega'_k \leqslant \frac{1}{n} = \omega_k$ for $k \in \{n_h + 1, \ldots, n_h + t_0\}$, and $\omega'_k = p_M = \omega_k$ if $1 \leqslant k \leqslant n_h$, we infer the existence of a set $J \subseteq \{n_h + t_0 + 1, \ldots, n\}$ such that

- $\sum_{j \in J} \omega'_j \geqslant \delta + \sum_{j \in J} \omega_j$; and
- $\omega'_j > \omega_j$ for each $j \in J$.

As before, we can then define a weighted trinucleotide code $\mathcal{Z} = (\zeta_1, \ldots, \zeta_{64})$ satisfying Conditions (1)–(3) and with dispersion at least $\mathrm{d}(\mathcal{W}')$ by setting $\zeta_i = \frac{1}{n}$, next $\zeta_k = \omega'_k$ if $k \notin J \cup \{i\}$, and finally defining ζ_j for $j \in J$ such that $\zeta_j \leqslant \omega'_j$ and $\sum_{j \in J} \zeta_j = \left(\sum_{j \in J} \omega'_j\right) - \delta$. In total, we can thus assume that $\omega'_i = \omega_i$ if $1 \leqslant i \leqslant n_h + t_0$.

We now conclude that $\mathrm{d}(\mathcal{W}') = \mathrm{d}(\mathcal{W})$. This is true if $\omega'_k = \omega_k$ for $k \in \{n_h + t_0 + 2, \ldots, n\}$, since then $\mathcal{W}$ and $\mathcal{W}'$ must be equal as both sum to 1. Recalling that $\omega'_k \geqslant p_m = \omega_k$ for each $k \in \{n_h + t_0 + 2, \ldots, n\}$, we can thus suppose that $\sum_{j=n_h+t_0+2}^{n} \omega'_k = \delta + \sum_{j=n_h+t_0+2}^{n} \omega_k$ for some positive real δ. Because, in addition $\omega'_k = \omega_k$ if $1 \leqslant k \leqslant n_h + t_0$, we infer that $\omega'_{n_h+t_0+1} = \omega_{n_h+t_0+1} - \delta$. Therefore, $\mathrm{d}(\mathcal{W}')$ is most, and hence equal to, $\mathrm{d}(\mathcal{W})$. We have proved that $\mathcal{W}$ indeed has the largest dispersion among all weighted trinucleotide codes satisfying Conditions (1)–(3). This concludes the proof. $\qquad\square$

Appendix D: Case with a Given Number of Codons of High Frequency

One could want to insist that the number of codons of high frequency is precisely n_h rather than at least n_h, thereby replacing Condition (2) from Sect. 4.1, namely

(2) for every $i \in \{n_h + 1, \ldots, n\}$, we have $p_m \leqslant \omega_i \leqslant \frac{1}{n}$,

by the following condition

(2') for every $i \in \{n_h + 1, \ldots, n\}$, we have $p_m \leqslant \omega_i < \frac{1}{n}$.

The extremal values proved in Propositions 4.2 and 4.3 are left unchanged under our assumption that $p_m < \frac{1}{n} < p_M$ (for otherwise, as reported above, all n occurring

codons must have frequency $\frac{1}{n}$). This clearly holds for (4.6a) and (4.6b), since the weighted trinucleotide code provided in the proof already satisfies Condition (2'), as $p_m < \frac{1}{n}$.

The other weighted trinucleotide codes provided in the proofs of Propositions 4.2 and 4.3 can be slightly modified to satisfy Condition (2') instead of Condition (2), and have dispersion arbitrarily close to the original one. Indeed, for Proposition 4.2, we can choose $\delta > 0$ such that $f_m := \frac{1}{n} - \delta \geqslant p_m$ and $f_M := \frac{1}{n} + \frac{n-n_h}{n_h}\delta \leqslant p_M$. We then change the value of ω_i to f_M if $1 \leqslant i \leqslant n_h$, and to f_m if $n_h + 1 \leqslant i \leqslant n$. Note that the total sum remains 1, as

$$(n - n_h)f_m + n_h f_M = \frac{n - n_h}{n} - (n - n_h)\delta + \frac{n_h}{n} + (n - n_h)\delta = 1.$$

Furthermore, $\left| f_M - \frac{1}{64} \right| = \left| \frac{1}{n} - \frac{1}{64} \right| + \frac{n-n_h}{n_h}\delta$ and $\left| f_m - \frac{1}{64} \right| \leqslant \left| \frac{1}{n} - \frac{1}{64} \right| + \delta$. Therefore, the obtained weighted trinucleotide code has dispersion at most $2 - \frac{n}{32} + 2(n - n_h)\delta$, which can be made arbitrarily close to $2 - \frac{n}{32}$ by choosing δ arbitrarily close to 0 (and positive). As a side remark, if $n \leqslant 63$ then δ can also be chosen small enough that $f_m \geqslant \frac{1}{64}$, and then the dispersion of the newly defined weighted trinucleotide code is again exactly $2 - \frac{n}{32}$. Indeed, we have $\left| f_m - \frac{1}{64} \right| = \left| \frac{1}{n} - 64 \right| - \delta$, so that

$$(n - n_h)\left| f_m - \frac{1}{64} \right| + n_h \left| f_M - \frac{1}{64} \right| = (n - n_h)\left| \frac{1}{n} - \frac{1}{64} \right| + n_h \left| \frac{1}{n} - \frac{1}{64} \right|.$$

For the weighted trinucleotide code $\mathcal{W} = (\omega_1, \ldots, \omega_{64})$ provided to establish (4.7a)–(4.7c), first recall that $p_m + z_1 < \frac{1}{n}$ by (C.5), and hence $\mathcal{W}$ satisfies Condition (2') unless $t_0 \geqslant 1$, in which case the values $\omega_{n_h+1}, \ldots, \omega_{n_h+t_0}$ are all equal to $\frac{1}{n}$. As already observed in the proof of Proposition 4.3, it holds that $n_h + t_0 \leqslant n - 1$. We can choose $\delta > 0$ such that $f_m := \frac{1}{n} - \delta > p_m$ and $p_m + z_1 + t_0\delta < \frac{1}{n}$, recalling that $p_m + z_1 < \frac{1}{n}$. If $p_m + z_1 < \frac{1}{64}$, then up to further reducing the value of δ, we can moreover suppose that $p_m + z_1 + t_0\delta < \frac{1}{64}$. We now change the value of ω_i to f_m for $i \in \{n_h + 1, \ldots, n_h + t_0\}$, and to $p_m + z_1 + t_0\delta$ if $i = n_h + t_0 + 1$. We note that the obtained code satisfies Condition (2') as well as Conditions (1) and (3), and one sees similarly as before that its dispersion differs from that of the original code by at most $2t_0\delta$, which can be made arbitrarily close to 0 by choosing δ arbitrarily close to 0 (and positive). As a side remark, if $p_m + z_1 \geqslant \frac{1}{64}$, then the dispersion of the newly defined code remains unchanged, thus attaining the right side of (4.7b).

Declarations

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

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

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 (2008a) Varieties of comma free codes. Comput Math Appl 55:989–996

Michel CJ, Pirillo G, Pirillo MA (2008b) 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, Sereni J-S (2023) Reading frame retrieval of genes: a new parameter of codon usage based on the circular code theory. Bull Math Biol 85(24):1–21

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

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

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. Proc Natl Acad Sci USA 113:E6117–E6125