Acta Biotheoretica (2026) 74:4
https://doi.org/10.1007/s10441-025-09509-0

REGULAR ARTICLE

Construction of Dinucleotide Circular Codes Based on Nucleotide Probabilities

Elena Fimmel[1] · **Christian J. Michel**[2] · **Lutz Strüngmann**[1]

Received: 21 August 2025 / Accepted: 16 October 2025
© Prof. Dr. Jan van der Hoeven stichting voor theoretische biologie 2025

Abstract

The construction of a circular code through a biological process, particularly a primitive one in the absence of the protein world, has remained an open problem since the discovery of a maximal C^3 self-complementary trinucleotide circular code in genes in 1996 (Arquès and Michel, 1996). Circular codes are defined by their ability to recover the correct reading frame of genes at any position. While a class of 216 such trinucleotide codes has been identified, the KL method (Koch and Lehman, 1997), based on nucleotide probability products, generates only a restricted subclass of 88 C^3-codes (Lacan and Michel, 2001). Revisiting this probabilistic framework 25 years later, we demonstrate that various classes of dinucleotide circular codes can be generated using a nucleotide probability product model (called Construction 2). We introduce the concept of transitive dinucleotide codes and prove new theorems characterizing their circularity and comma-free properties. Using codon usage from bacteria, archaea, and eukaryotes, 2 "universal" maximal dinucleotide circular codes are observed: $D_{1,2} = \{AT, CA, CT, GA, GC, GT\}$ in the codon site $1-2$ and $D_{2,3}$ in the codon site $2-3$ which can be deduced from $D_{1,2}$ by 1-letter cyclical permutation α_1 or identically by reversing permutation $D_{2,3} = \alpha_1(D_{1,2}) = \overleftarrow{D_{1,2}}$. Unexpectedly, we then show that, under the independence assumption, the dinucleotide code $E_{1,2}$ through Construction 2 from nucleotide frequencies in the codon sites 1 and 2, is a maximal dinucleotide circular code and is equal to the observed dinucleotide code: $E_{1,2} = D_{1,2}$. These findings support a theoretical model in which dinucleotide circular codes may have originated from statistical properties of primitive nucleotide distributions, providing insights into the possible emergence of the genetic code.

Keywords Dinucleotide circular codes · Construction · Classes · Nucleotide probability · Transitive codes · Genetic code

Extended author information available on the last page of the article

Published online: 29 December 2025

1 Introduction

The construction of a circular code through a biological process, particularly a primitive one, i.e. in the absence of the protein world, has remained an open problem since the discovery of a maximal C^3 self-complementary trinucleotide circular code, denoted X, in genes in 1996 (Arquès and Michel 1996), that is, for nearly 30 years. The defining property of a circular code is its ability to unambiguously retrieve the correct reading frame of genes at any position (see the reviews in Michel (2008); Fimmel and Strüngmann (2018)). The code X is maximal (comprising 20 trinucleotides), self-complementary (10 trinucleotides are complementary of the remaining 10) and C^3 (the codes obtained by shifting each trinucleotide in X by 1 and 2 nucleotides, denoted X_1 and X_2, respectively, are also circular and maximal). Over a 4-letter alphabet and words of length 3, there exist exactly 216 maximal C^3 self-complementary trinucleotide circular codes (Arquès and Michel 1996), and the code X found in genes is one of them.

One year after this discovery, Koch and Lehman proposed in 1997 a simple probabilistic approach, referred to as the KL method, based on the product of nucleotide probabilities, which enables the construction of a particular class of circular codes that are C^3 (Koch and Lehmann 1997). Four years later, in 2001, Lacan and Michel demonstrated that the KL method does not allow the construction of all C^3-codes, but only a subclass of 88 out of the 216 (Lacan and Michel 2001). Furthermore, it does not permit the construction of any non-C^3 circular code, i.e. a classical circular code. As a result, the theoretical contributions of these four authors were subsequently abandoned.

Twenty-five years later, we revisit the idea of constructing circular codes from nucleotide probabilities. This approach is particularly relevant in the context of prebiotic nucleotide soups, where nucleotides may have been present in unequal proportions. Certain combinations of nucleotide probabilities could therefore give rise to dinucleotide circular codes, at least to specific classes of them.

This article is organized as follows. The necessary definitions, notations, and classical properties of circular codes are presented in Sect. 2. In Sect. 3, we revisit and rephrase key historical results obtained by Koch and Lehman in 1997 (Koch and Lehmann 1997) (Sect. 3.1) and by Lacan and Michel in 2001 (Lacan and Michel 2001) (Sect. 3.2). In Sect. 4, we extend this concept to dinucleotide circular codes using a new probabilistic construction (called Construction 2) based on nucleotide probability distribution tables. Various classes of dinucleotide circular codes have previously been identified, and their mathematical properties studied (Michel and Pirillo 2013; Michel et al. 2016; Fimmel et al. 2015, 2017b). In Sect. 4.1, we demonstrate that these different dinucleotide classes can be constructed using products of nucleotide probabilities. In Sect. 4.2, we introduce the mathematical concept of transitive dinucleotide codes, which leads to several new theorems. In particular:

- Any dinucleotide code $X \subseteq \mathcal{B}^2$ generated by Construction 2 is transitive.
- A transitive dinucleotide code $X \subseteq \mathcal{B}^2 \backslash \{AA, CC, GG, TT\}$ of size 6 is circular.
- A transitive dinucleotide code $X \subseteq \mathcal{B}^2 \backslash \{AA, CC, GG, TT\}$ of size at most 5 is comma-free.

- A dinucleotide code $X \subseteq \mathcal{B}^2 \backslash \{AA, CC, GG, TT\}$ generated by Construction 2 of size at most 4 is strong comma-free.

Section 5 presents a biological application of dinucleotide circular codes. The preparatory Sect. 5.1 details the acquisition of gene codon usage data for genomes of archaea, bacteria and eukaryotes from the Codon Statistics Database (CSD) (Subramanian et al. 2022) (see Appendix A). Codon usage was computed across 34,020,997 genes from 8,345 bacterial genomes Bac, 1,280,890 genes from 432 archaeal genomes Arc, and 20,206,058 genes from 1,150 eukaryotic genomes Euk (summarized in Table 3 of Appendix B). In Sect. 5.2, after computing the 16 dinucleotide frequencies $f_{1,2}$ and $f_{2,3}$ in the gene codon sites $1-2$ and $2-3$, respectively, (Appendix C), we identify 2 maximal dinucleotide circular codes $D_{1,2}$ and $D_{2,3}$ in the codon sites $1-2$ and $2-3$, respectively, across all three kingdoms of bacteria, archaea and eukaryotes, without exception. Section 5.3 is divided into two parts. Section 5.3.1 computes the 4 nucleotide frequencies at each of the 3 codon sites for bacterial, archaeal and eukaryotic genes using codon usage. Section 5.3.2, under an independence assumption, derives the 16 dinucleotide frequencies $p_{1,2}$ in the codon site $1-2$ by applying a product of nucleotide probabilities. Applying Construction 2, the constructed dinucleotide code $E_{1,2}$ in the codon site $1-2$ is shown to be a maximal circular code and is exactly equal to the observed dinucleotide code $D_{1,2}$ found in genes across all three kingdoms. Finally, Sect. 6 proposes a theoretical model of a possible precursor to the genetic code, grounded in dinucleotide circular codes.

2 Definitions and Notations

Let us denote the nucleotide 4-letter alphabet by $\mathcal{B} := \{A, C, G, T\}$ where A stands for Adenine, C stands for Cytosine, G stands for Guanine and T stands for Thymine. Recall that the symmetric group is defined as

$$S_{\mathcal{B}} = \{\pi : \mathcal{B} \to \mathcal{B} \mid \pi \; is \; bijective\}$$

with the usual group operation given by composition of functions. The group $S_{\mathcal{B}}$ has 24 elements and any bijective mapping $\pi : \mathcal{B} \to \mathcal{B}$ can be applied component wise to $x \in \mathcal{B}^3$ and thus induces a bijective map $\mathcal{B}^3 \to \mathcal{B}^3$ which we will also denote by π. The complementary structure of the DNA double helix is based on the complementary transformation c from $S_{\mathcal{B}}$

$$c : (A, T, C, G) \to (T, A, G, C)$$

which exchanges A and T, and C and G.

A second symmetric group used in the following sections is the group $(S_3, \circ)$, the symmetric group on the 3 numbers 1, 2 and 3, where

$$S_3 := \{\alpha : \{1, 2, 3\} \to \{1, 2, 3\} \mid \alpha \text{ is bijective}\}$$

and $\circ$ denotes the composition of mappings. Clearly, any such α induces a mapping on the set of trinucleotides $\mathcal{B}^3$ by permuting the order of the bases in the trinucleotides. Let us focus on the subgroup of *cyclical permutations* of $(S_3, \circ)$ denoted by

$$\mathcal{A}_3 := \{\alpha_0, \alpha_1, \alpha_2\} \subset S_3$$

where

$$\alpha_0(N_1N_2N_3) = N_1N_2N_3, \quad \alpha_1(N_1N_2N_3) = N_2N_3N_1 \quad \text{and} \quad \alpha_2(N_1N_2N_3) = N_3N_1N_2$$

for any codon $N_1N_2N_3 \in \mathcal{B}^3$. $(\mathcal{A}_3, \circ)$ is known as the *alternating subgroup* of $(S_3, \circ)$ and contains the 2 shift operations α_1 and α_2. However, it does not contain the second important biological transformation related to the antiparallel structure of the DNA double helix: the reversing permutation of the indices. We will indicate this permutation by $^{\leftarrow}$ so that a given trinucleotide $x = (N_1, N_2, N_3) \in \mathcal{B}^3$ leads to $\overleftarrow{x} := (N_3, N_2, N_1)$. Obviously, for a trinucleotide (codon) $x \in \mathcal{B}^3$ the reversed complementary trinucleotide (anticodon) can now be described as $\overleftarrow{c(x)}$.

Definition 1 Let $X \subseteq \mathcal{B}^3$. The code X is *self-complementary* if for each codon $x \in X$ its anticodon $\overleftarrow{c(x)} \in X$.

Definition 2 Let $X \subseteq \mathcal{B}^n, n = 2, 3$. We say that X is

(1) a *circular code* if any word over the alphabet $\mathcal{B}$ written on a circle has at most one decomposition into words from X;
(2) a *comma-free code* if for any given 2 words $x_1, x_2 \in X$ any sub-word of the concatenation x_1x_2, except x_1, x_2 themselves, does not belong to X;
(3) a *strong comma-free code* if no $v \in \mathcal{B}^+$ (i.e. $v \neq \epsilon$) appears both as a prefix and a suffix in X;
(4) a *maximal circular (comma-free, strong comma-free) n-letter code* if it is not contained in a larger circular (comma-free, strong comma-free) n-letter code;
(5) a *maximum circular (comma-free, strong comma-free) n-letter code* (or equivalently code of maximal size) if $|Y| \leq |X|$ whenever Y is a circular (comma-free, strong comma-free) n-letter code over $\mathcal{B}$.

Definition 3 A trinucleotide code $X \subseteq \mathcal{B}^3$ is a C^3*-code* if X as well as $\alpha_1(X)$ and $\alpha_2(X)$ are circular.

For a codon $x \in X$, we denote the complete cyclic equivalence class $\overline{x}$ by

$$\overline{x} = \{x, \alpha_1(x), \alpha_2(x)\}.$$

All mappings from $S_{\mathcal{B}}$ preserve the property of circularity (Fimmel et al. 2015): Any permutation from $S_{\mathcal{B}}$ preserves comma-freeness and circularity, hence the C^3 property.

Finally, the *reversing permutation* which reverses a codon also preserves self-complementarity and circularity (Fimmel et al. 2015).

3 Probabilistic Approach to Construct Trinucleotide Circular Codes

A simple probabilistic approach allows the construction of some, but not all, trinucleotide C^3-codes. With the aim of describing our generalization of this approach and to have this article self-contained, we explicitly recall and rephrase here the main historical results which were obtained by Koch and Lehman in 1997 (Koch and Lehmann 1997), and Lacan and Michel in 2001 (Lacan and Michel 2001).

3.1 Koch and Lehman Results in 1997

In 1997, Koch and Lehmann proposed a probabilistic approach to construct trinucleotide C^3-codes (Koch and Lehmann 1997). An essential assumption of the KL method (Koch and Lehmann method) is the independence of the distribution of nucleotides in a gene. In a codon of a gene, there are 3 positions noted 1, 2 and 3. We assign a probability p_i, $i = 1, 2, 3$, for each nucleotide in each position that can be represented by a 4×3 probability distribution table P

Base	Pos. 1	Pos. 2	Pos. 3
A	$p_1(A)$	$p_2(A)$	$p_3(A)$
C	$p_1(C)$	$p_2(C)$	$p_3(C)$
G	$p_1(G)$	$p_2(G)$	$p_3(G)$
T	$p_1(T)$	$p_2(T)$	$p_3(T)$
Σ	1	1	1

Note that such probabilities p_i can easily be obtained from a codon usage table.

Furthermore, let us consider a complete (i.e. of 3 elements) cyclic equivalence class

$$\overline{N_1N_2N_3} = \{N_1N_2N_3, N_2N_3N_1, N_3N_1N_2\}, \quad N_i \in \mathcal{B}.$$

Thus, the periodic trinucleotides $\{AAA, CCC, GGG, TTT\}$ are excluded.

Construction 1 Given a probability distribution table P, we will assign a codon $N_1N_2N_3 \in \overline{x}$ from a complete cyclic equivalence class $\overline{x}$ to the constructed trinucleotide code $X = X(P)$ if the following 2 conditions are satisfied:

$$\begin{aligned} p_1(N_1)p_2(N_2)p_3(N_3) > p_1(N_3)p_2(N_1)p_3(N_2) \quad \text{and} \\ p_1(N_1)p_2(N_2)p_3(N_3) > p_1(N_2)p_2(N_3)p_3(N_1). \end{aligned} \tag{1}$$

Inequalities (1) mean that a codon has a greater probability than the 2 other codons of the same equivalence class in the selected (reading) frame. Of course, one or both inequalities in the equivalence class can be not fulfilled for all 3 codons, i.e. some probabilities are identical. In this case, no codon from this equivalence class is assigned to the code X to be constructed.

For the convenience of the reader, we will give an example for Construction 1.

Example 1 Assume the following probability distribution table P

Base	Pos. 1	Pos. 2	Pos. 3
A	0.4	0.1	0.2
C	0.3	0.2	0.3
G	0.2	0.3	0.2
T	0.1	0.4	0.3
$\sum$	1	1	1

Then Construction 1 generates the following (non-maximal) C^3-code

$$X = \{ACA, AGA, ATA, ACC, CGA, AGC, AGG, AGT, ATC, ATT, CTC, CGT, CTT, ATG, CGC, CTG, GGT, GTT\}.$$

For instance, $ACA \in X$ since

$$p_1(A)p_2(C)p_3(A) = 0.4 \cdot 0.2 \cdot 0.2 > p_1(A)p_2(A)p_3(C) = 0.4 \cdot 0.1 \cdot 0.3$$

and

$$p_1(A)p_2(C)p_3(A) = 0.4 \cdot 0.2 \cdot 0.2 > p_1(C)p_2(A)p_3(A) = 0.3 \cdot 0.1 \cdot 0.2.$$

None of the codons *CGG*, *GGC*, *GCG* is in the code, because

$$p_1(C)p_2(G)p_3(G) = 0.3 \cdot 0.3 \cdot 0.2 = p_1(G)p_2(G)p_3(C) > p_1(G)p_2(C)p_3(G) = 0.2 \cdot 0.2 \cdot 0.2.$$

The following assertion was proved in Koch and Lehmann (1997) but was not explicitly highlighted as such. We therefore present it in an explicit and compact form.

Proposition 1 *(Koch and Lehmann (1997)) A trinucleotide code X generated by Construction 1 is a C^3-code.*

Proof Let us consider

$$w = N_1^1 N_2^1 N_3^1 N_1^2 N_2^2 N_3^2 ... N_1^n N_2^n N_3^n$$

where $N_1^i N_2^i N_3^i \in X, i = 1, ..., n$. In correspondence with Construction 1, we have

$$p_1(N_1^i)p_2(N_2^i)p_3(N_3^i) > p_1(N_2^i)p_2(N_3^i)p_3(N_1^i)$$

for all $i \in \{1, ..., n\}$ and, thus,

$$\prod_{i=1}^{n} p_1(N_1^i)p_2(N_2^i)p_3(N_3^i) > \prod_{i=1}^{n} p_1(N_2^i)p_2(N_3^i)p_3(N_1^i). \quad (*)$$

Suppose that w can be decomposed on a circle in the following way:

$$w = N_1^1|N_2^1N_3^1N_1^2|N_2^2N_3^2...N_1^n|N_2^nN_3^n$$

so that $N_2^iN_3^iN_1^{i+1} \in X, i = 1, ..., n$, and $N_1^{n+1} = N_1^1$. That implies that

$$p_1(N_2^i)p_2(N_3^i)p_3(N_1^{i+1}) > p_1(N_1^{i+1})p_2(N_2^i)p_3(N_3^i)$$

for all $i \in \{1, ..., n\}$. Then, by rearranging the factors in the product of probabilities

$$\prod_{i=1}^{n} p_1(N_2^i)p_2(N_3^i)p_3(N_1^{i+1}) = \prod_{i=1}^{n} p_1(N_2^i)p_2(N_3^i)p_3(N_1^i) >$$
$$\prod_{i=1}^{n} p_1(N_1^{i+1})p_2(N_2^i)p_3(N_3^i) = \prod_{i=1}^{n} p_1(N_1^i)p_2(N_2^i)p_3(N_3^i).$$

This is a contradiction to (*). Similarly, it can be proved that the shift by 2 positions in w also leads to a contradiction. The circularity of $\alpha_1(X)$ and $\alpha_2(X)$ are obtained similarly. □

A natural question arising from Proposition 1 is whether any C^3-code can be obtained in this way. This question was answered negatively by Lacan and Michel (2001) four years later in 2001. We reproduce here the proof of Lacan and Michel (2001).

3.2 Lacan and Michel Results in 2001

Proposition 2 *(Lacan and Michel (2001)) There are trinucleotide C^3-codes that cannot be generated by Construction 1.*

Proof Let us consider a C^3-code $X \subseteq \mathcal{B}^3$ that contains the following combination of codons

$$N_1N_2N_3, \quad N_3N_1N_4, \quad N_4N_4N_2 \in X, \quad N_i \in \mathcal{B}, N_i \neq N_j, i \neq j.$$

Such C^3-codes exist. Indeed, the C^3-code identified in genes contains the codons *ATC*, *CAG* and *GGT* (Arquès and Michel 1996).

Since $N_4N_4N_2 \in X$ applies, we have according to Construction 1

$$p_1(N_4)p_2(N_4)p_3(N_2) > p_1(N_4)p_2(N_2)p_3(N_4)$$

and consequently,

$$p_2(N_4)p_3(N_2) > p_2(N_2)p_3(N_4). \quad (*)$$

Since $N_3N_1N_4 \in X$ applies, we have according to Construction 1

$$p_1(N_3)p_2(N_1)p_3(N_4) > p_1(N_1)p_2(N_4)p_3(N_3). \quad (**)$$

As $N_1N_2N_3 \in X$ and $p_3(N_4) > 0$, we also have

$$p_1(N_1)p_2(N_2)p_3(N_3) > p_1(N_3)p_2(N_1)p_3(N_2) = p_1(N_3)p_2(N_1)p_3(N_4)\frac{p_3(N_2)}{p_3(N_4)}.$$

Using (**) we thus obtain

$$p_1(N_1)p_2(N_2)p_3(N_3) > p_1(N_1)p_2(N_4)p_3(N_3)\frac{p_3(N_2)}{p_3(N_4)}$$

which is equivalent to the inequality

$$p_2(N_2)p_3(N_4) > p_2(N_4)p_3(N_2)$$

after truncating the equal probabilities on both sides. This is a contradiction to the inequality (*). □

We have an immediate consequence which divides the set of 216 maximal self-complementary C^3-codes into 2 subsets depending on whether they can be constructed by Construction 1 or not.

Proposition 3 *(Lacan and Michel (2001)) There are 88 maximal self-complementary C^3-codes that can be generated by Construction 1 for some probability distribution table P. Thus, there are 128 (= 216 − 88) maximal self-complementary C^3-codes that cannot be generated by Construction 1 for any probability distribution table.*

Proof By algorithm and computer calculus (Table 3 in Lacan and Michel (2001)). □

4 Probabilistic Approach to Construct Dinucleotide Circular Codes

Now we transfer the above trinucleotide probabilistic approach to the case of dinucleotide circular codes. In this case, the complete equivalence classes consist of 2 elements that are also reverse to each other

$$\overline{N_1N_2} = \{N_1N_2, N_2N_1\}, \quad N_i \in \mathcal{B}, N_1 \neq N_2$$

and the probability distributions of individual nucleotides can be represented by a 4×2 probability distribution table P

Base	Pos. 1	Pos. 2
A	$p_1(A)$	$p_2(A)$
C	$p_1(C)$	$p_2(C)$
G	$p_1(G)$	$p_2(G)$
T	$p_1(T)$	$p_2(T)$
Σ	1	1

Construction 2 Given a probability distribution table P, we will assign a dinucleotide $N_1N_2 \in \overline{x}$ from a complete cyclic equivalence class $\overline{x}$ to the constructed dinucleotide code X if the following condition is satisfied:

$$p_1(N_1)p_2(N_2) > p_1(N_2)p_2(N_1). \tag{2}$$

Inequality (2) means that a dinucleotide has a greater probability than the other dinucleotide of the same equivalence class in the selected (reading) frame. Of course, the inequality in the equivalence class can be not fulfilled for both dinucleotides, i.e. the 2 probabilities are equal. In this case, no dinucleotide from this equivalence class is assigned to the code X to be constructed.

We would like to remark that it is actually not necessary to work with probabilities. In fact Construction 2 also works for any table P with non-negative entries whose sum does not even have to be one. It can be easily seen that such a table can be converted into one with proper probabilities by dividing the entries by the sum of the corresponding columns. The constructed code however stays the same.

Example 2 Assume the following general distribution table P

Base	Pos. 1	Pos. 2
A	4	1
C	3	2
G	2	3
T	1	4
$\sum$	10	10

as well as the probability distribution table P'

Base	Pos. 1	Pos. 2
A	$0.4 = \frac{4}{10}$	$0.1 = \frac{1}{10}$
C	$0.3 = \frac{3}{10}$	$0.2 = \frac{2}{10}$
G	$0.2 = \frac{2}{10}$	$0.3 = \frac{3}{10}$
T	$0.1 = \frac{1}{10}$	$0.4 = \frac{4}{10}$
$\sum$	1	1

Then Construction 2 based on P and also P' assigns the following code

$$X = \{AC, AG, AT, CG, CT, GT\}.$$

For simplicity and to be more general, we sometimes use probability distribution tables (P) in the following form:

Base	Pos. 1	Pos. 2
N_1	$p_1(N_1) = p_1$	$p_2(N_1) = q_1$
N_2	$p_1(N_2) = p_2$	$p_2(N_2) = q_2$
N_3	$p_1(N_3) = p_3$	$p_2(N_3) = q_3$
N_4	$p_1(N_4) = p_4$	$p_2(N_4) = q_4$
Σ	1	1

In the next section we will describe all dinucleotide circular codes that can be obtained by Construction 2.

4.1 Classes of Dinucleotide Circular Codes

As for trinucleotide codes, we have the following result. For dinucleotide codes, the circularity already implies the C^2-property. In fact, any dinucleotide circular code is automatically a C^2-code, since its shifted code is equal to its reversed code and the reversing transformation preserves the circularity (Michel and Pirillo 2013; Michel et al. 2016; Fimmel et al. 2015, 2017b).

Proposition 4 *A dinucleotide code X generated by Construction 2 is circular, and hence a C^2-code.*

Proof As in the proof of Proposition 1. □

However, in this case, it turns out that any maximum dinucleotide circular code and even any maximal dinucleotide (strong) comma-free code can be obtained by Construction 2. Recall that a maximum dinucleotide circular code is of size 6.

Proposition 5 *Any maximum dinucleotide circular code can be obtained through Construction 2.*

Proof According to Michel and Pirillo (2013); Fimmel et al. (2016), each maximum dinucleotide circular code has the form

$$X = \{N_1N_2, N_1N_3, N_1N_4, N_2N_3, N_2N_4, N_3N_4\}, \quad N_i \in \mathcal{B}, N_i \neq N_j, i \neq j.$$

Assume the following probability distribution table P

Base	Pos. 1	Pos. 2
N_1	p_1	p_4
N_2	p_2	p_3
N_3	p_3	p_2
N_4	p_4	p_1
Σ	1	1

with $p_1 > p_2 > p_3 > p_4$. It is easy to check that using P, Construction 2 generates the desired code. □

We now turn to dinucleotide comma-free codes. Recall that a maximum dinucleotide comma-free code is of size 5.

Proposition 6 *Any maximum dinucleotide comma-free code can be obtained using Construction 2.*

Proof According to Fimmel and Strüngmann (2016), there are 36 maximum dinucleotide comma-free codes. Furthermore, they are all contained in a maximum dinucleotide circular code and have one of the following 3 types (12 of each type)

(1) $\{N_1N_2, N_1N_3, N_1N_4, N_2N_4, N_3N_4\}, \quad N_i \in \mathcal{B}, N_i \neq N_j, i \neq j.$

(2) $\{N_1N_2, N_1N_3, N_1N_4, N_2N_3, N_2N_4\}, \quad N_i \in \mathcal{B}, N_i \neq N_j, i \neq j.$
(3) $\{N_1N_3, N_1N_4, N_2N_3, N_2N_4, N_3N_4\}, \quad N_i \in \mathcal{B}, N_i \neq N_j, i \neq j.$

The codes of one type are connected by bijective transformations from $S_{\mathcal{B}}$. Therefore, it is sufficient to show that all 3 variants can be obtained from a maximum dinucleotide circular code by excluding matching dinucleotides. Without loss of generality, consider the maximum dinucleotide circular code $\{AC, AG, AT, CG, CT, GT\}$. According to the above description of the comma-free codes, the dinucleotides AC, CG and GT, correspondingly, should be excluded from the code. Let us now specify the appropriate probability distributions excluding these 3 dinucleotides. For example, consider the following probability distribution table P excluding:
in the case of GT

Base	Pos. 1	Pos. 2
A	0.4	0.2
C	0.3	0.2
G	0.2	0.4
T	0.1	0.2
$\sum$	1	1

in the case of CG

Base	Pos. 1	Pos. 2
A	0.4	0.2
C	0.3	0.3
G	0.2	0.2
T	0.1	0.3
$\sum$	1	1

and in the case of AC

Base	Pos. 1	Pos. 2
A	0.4	0.2
C	0.2	0.1
G	0.2	0.3
T	0.2	0.4
$\sum$	1	1

In this case, we get the maximum dinucleotide comma-free code$_{\{AG, AT, CG, CT, GT\}}$ since $0.4 \cdot 0.1 = p_1(A) \cdot p_2(C) = p_1(C) \cdot p_2(A) = 0.2 \cdot 0.2$ is fulfilled. □

We finally consider dinucleotide strong comma-free codes. Recall that a maximum dinucleotide strong comma-free code is of size 4.

Proposition 7 *Any maximum dinucleotide strong comma-free code can be obtained using Construction 2.*

Proof According to Fimmel et al. (2017a, 2017b), any maximum dinucleotide strong comma-free code has the form

$$\{N_1N_3, N_1N_4, N_2N_3, N_2N_4\}, \quad N_i \neq N_j, i \neq j.$$

Let $0.25 < p < 0.5$ and thus, $p > 0.5 - p > 0$. Consider the following probability distribution table P

Base	Pos. 1	Pos. 2
N_1	p	$0.5-p$
N_2	p	$0.5-p$
N_3	$0.5-p$	p
N_4	$0.5-p$	p
Σ	1	1

In this case, we get the code$\{N_1N_3, N_1N_4, N_2N_3, N_2N_4\}$, since $p \cdot (0.5-p) = p_1(N_1) \cdot p_2(N_2) = p_1(N_2) \cdot p_2(N_1) = (0.5-p) \cdot p$, $and\, p \cdot (0.5-p) = \quad p_1(N_3) \cdot p_2(N_4) = \quad p_1(N_4) \cdot p_2(N_3) = (0.5-p) \cdot p$ and are fulfilled. □

In the next section, we will show that for the cardinalities 4, 5 and 6, any dinucleotide code constructed by Construction 2 has to be maximum strong comma-free, comma-free and circular, respectively.

4.2 Transitive Dinucleotide Codes

We first prove an essential property of dinucleotide codes generated by some probability distribution table P. Let us start with the following definition.

Definition 4 Let $X \subseteq \mathcal{B}^2$ be a dinucleotide code. Then X is called *transitive* if $N_1N_2 \in X$ and $N_2N_3 \in X$ implies $N_1N_3 \in X$ for any bases $N_1, N_2, N_3 \in \mathcal{B}$.

We illustrate this property with the following example.

Example 3 The dinucleotide code $X = \{AC, AG, CG, TG\}$ is transitive.

It turns out that dinucleotide codes generated by Construction 2 are automatically transitive.

Proposition 8 *Let $X \subseteq \mathcal{B}^2$ be a dinucleotide code generated by Construction 2. Then X is transitive.*

Proof Assume that X is generated by the following probability distribution table P according to Construction 2:

Base	Pos. 1	Pos. 2
N_1	p_1	q_1
N_2	p_2	q_2
N_3	p_3	q_3
N_4	p_4	q_4
Σ	1	1

Furthermore, we assume that $N_1N_2, N_2N_3 \in X$ and $N_1N_3 \notin X$. At the level of probabilities, this means

$$(I) \quad p_1q_2 > p_2q_1$$

and

$$(II) \quad p_2q_3 > p_3q_2$$

but

$$(III) \quad p_1q_3 \leq p_3q_1.$$

Multiplying (*II*) by p_1 ($p_1 \neq 0$ by (*I*)), we obtain $p_1p_2q_3 > p_1p_3q_2$ and hence using (*I*) that

$$(*) \quad p_1p_2q_3 > p_1p_3q_2 > p_3p_2q_1.$$

By (*II*), it follows that $p_2 \neq 0$, so we may divide $(*)$ by p_2 to obtain $p_1q_3 > p_3q_1$ which contradicts assumption (*III*). □

We now prove properties of transitive codes.

Theorem 1 *Let $X \subseteq \mathcal{B}^2 \backslash \{AA, CC, GG, TT\}$ be a transitive dinucleotide code. Then the following statements hold:*

(1) *X is circular.*
(2) *If X is of size at most 5 then X is comma-free.*

Proof (1): The first part of the claim was proved for dinucleotide codes using the theory of tournaments in Fimmel et al. (2016). We give here a direct proof of the assertion without using the graph-theoretical approach. Assume that X is not circular and let $d_1 \cdots d_n \in X^n$ be a sequence of dinucleotides $d_i = N_1^i N_2^i$ from X that has a second decomposition over X, i.e.

$$N_2^i N_1^{i+1} N_2^{i+1} \cdots N_2^n N_1^1 \in X^n.$$

It implies that $N_2^i N_1^{i+1} \in X$ and also $N_2^n N_1^1 \in X$. Using that $N_1^1 N_2^1 \in X$ and $N_2^1 N_1^2 \in X$, transitivity implies $N_1^1 N_1^2 \in X$. Since also $N_1^2 N_2^2 \in X$, transitivity again implies that also $N_1^1 N_2^2 \in X$. Inductively, we get $N_1^1 N_2^i \in X$ for all $i \leq n$ and finally $N_1^1 N_1^1 \in X$ since also $N_2^n N_1^1 \in X$. This contradicts the fact that $X \subseteq \mathcal{B}^2 \backslash \{AA, CC, GG, TT\}$.

(2): Now assume that X is of size at most 5 and that X is not comma-free. Thus, the code must contain the 3 dinucleotides of the form $N_1N_2, N_2N_3, N_3N_4 \in X$. By circularity, these 3 dinucleotides cannot form a circle, i.e. $N_1 \neq N_4$. However, transitivity implies that also $N_1N_3, N_2N_4, N_1N_4 \in X$, meaning that X is of size 6 - a contradiction. □

Let us make the following remark.

Remark 1 There are dinucleotide circular codes that are not transitive.

Example 4 The dinucleotide code $X = \{AC, CG\}$ is circular but not transitive.

Now we show that all dinucleotide codes of size 4 which are generated by Construction 2 (a probability distribution table) are strong comma-free.

Theorem 2 *Let $X \subseteq \mathcal{B}^2 \backslash \{AA, CC, GG, TT\}$ be a dinucleotide code generated by Construction 2. If X is of size 4 then X is strong comma-free.*

Proof Assume that X is generated by a probability distribution table P with probabilities $p_1, \cdots, p_4$ and $q_1, \cdots, q_4$. By Theorem 1, we know that X is comma-free. Let us assume that X is not strong comma-free. This means that there are 2 dinucleotides in the code, so that the suffix of one is the prefix of the other $N_1N_2, N_2N_3 \in X$. By transitivity (Proposition 8), we know that also $N_1N_3 \in X$. Since X has size 4 there is one more dinucleotide in X. The reversed dinucleotides $N_2N_1, N_3N_1, N_3N_2 \notin X$ are excluded, since X is comma-free. This leaves 6 possible dinucleotides, namely

$$N_3N_4,\ N_2N_4,\ N_4N_2,\ N_4N_1,\ N_1N_4 \quad \text{and} \quad N_4N_3.$$

The first 4 of them, namely $N_3N_4, N_2N_4, N_4N_2, N_4N_1$, are not possible, because a 5th dinucleotide must also be in the code due to the transitivity of the code - N_1N_4 in the first two cases and N_4N_3 or N_4N_2, correspondingly, in the two last ones. Assume $N_1N_4 \in X$. Since neither N_4N_2 nor N_3N_4 or their reversed dinucleotides cannot be in the code due to its size, we have

$$(*) \quad p_2q_4 = p_4q_2 \quad \text{and} \quad p_4q_3 = p_3q_4$$

which results in $p_3q_2 = p_2q_3$ and therefore $N_2N_3 \notin X$ - a contradiction. Similarly, it can be shown that the choice of N_4N_3 leads to the conclusion $N_1N_2 \notin X$ - a contradiction - and so X is strong comma-free. □

As an immediate corollary we have the following theorem.

Theorem 3 *Let $X \subseteq \mathcal{B}^2 \backslash \{AA, CC, GG, TT\}$ be a dinucleotide code of size n. Then the following statements are true:*

(1) *If $n = 6$ then X is a maximum circular code if and only if X is generated by Construction 2.*
(2) *If $n = 5$ then X is a maximum comma-free code if and only if X is generated by Construction 2.*
(3) *If $n = 4$ then X is a maximum strong comma-free code if and only if X is generated by Construction 2.*

Proof Follows immediately from Theorem 1, Proposition 8 and Theorem 2. □

Remark 2 Not all transitive dinucleotide codes can be generated by Construction 2, i.e. by some probability distribution table.

Example 5 The dinucleotide code $X = \{AC, CG, AG, TG\}$ is transitive but it is not generated by any probability distribution table because X has the size 4 and is not strong comma-free because C is the suffix of AC and prefix of CG.

Finally, the structure of dinucleotide codes of size ≤ 3 generated by a probability distribution table is almost clear since their form is determined by transitivity. We start with dinucleotide codes of size 2.

Theorem 4 *Let $X \subseteq \mathcal{B}^2 \backslash \{AA, CC, GG, TT\}$ be a dinucleotide code of size 2 generated by some probability distribution table. Then X is of the form*

$$X = \{N_1N_2, N_1N_3\} \quad \text{or} \quad \overleftarrow{X} = \{N_2N_1, N_3N_1\}.$$

Proof Assume that X is generated by the probability distribution table P. Since X is of size 2 its 2 dinucleotides can either have a vertex in common or not. Thus we have 4 possibilities:

(1) $X = \{N_1N_2, N_1N_3\}$.
(2) $X = \{N_2N_1, N_3N_1\}$.
(3) $X = \{N_1N_2, N_3N_1\}$.
(4) $X = \{N_1N_2, N_3N_4\}$.

The first two cases are transitive and easily seen to be generated by a probability distribution table (see below Theorem 6).

The 3rd case is not possible because of the transitivity of the code X, since, together with the dinucleotides N_3N_1 and N_1N_2, the dinucleotide N_3N_2 must also be in the code and thus $|X| > 2$ must apply – a contradiction.

In the 4th case, due to $N_1N_3, N_1N_4 \notin X$, we have

$$(*) \quad p_1q_3 = p_3q_1 \quad \text{and} \quad p_1q_4 = p_4q_1$$

which results in $p_4q_3 = p_3q_4$ and therefore $N_3N_4 \notin X$ - a contradiction. □

We now analyse the dinucleotide codes of size 3.

Theorem 5 *Let $X \subseteq \mathcal{B}^2 \backslash \{AA, CC, GG, TT\}$ be a dinucleotide code of size 3 generated by some probability distribution table. Then X is of the form*

$$X = \{N_1N_2, N_1N_3, N_1N_4\} \quad \text{or} \quad X = \{N_1N_2, N_2N_3, N_1N_3\}$$
$$\text{or} \quad X = \{N_2N_1, N_3N_1, N_4N_1\}.$$

Proof All 3 cases are transitive and it is easily seen that they are generated by a probability distribution table (see below Theorem 6).

In all 3 cases, each pair of dinucleotides from the code has a common nucleotide. Furthermore, these 3 possibilities are the only ones with this property, taking into account the transitivity of the code. Now let us assume that 2 dinucleotides from the code do not share a common nucleotide. Without loss of generality, let it be $N_1N_2, N_3N_4 \in X, N_i \neq N_j$ (and consequently $N_2N_1, N_4N_3 \notin X$ due to the circularity of the code). Let us see which 3rd dinucleotide comes into question. Due to the transitivity of the code, the following applies

$$N_1N_3,\ N_3N_1,\ N_2N_4,\ N_4N_2,\ N_2N_3,\ N_4N_1 \notin X,$$

since in each of these cases a 4th dinucleotide must also be in X due to its transitivity. Assuming $N_3N_2 \in X$ then, as in the proof of Theorem 4, we get

$$(*) \quad p_1q_3 = p_3q_1 \quad \text{and} \quad p_1q_4 = p_4q_1$$

which results in $p_4q_3 = p_3q_4$ and therefore $N_3N_4 \notin X$ - a contradiction. Similarly, N_1N_4 is also excluded from the code. So in this case $|X| \leq 2$, which is a contradiction to the size of the code. □

Remark 3 The dinucleotide code $X = \{N_1N_2\}$, $N_1 \neq N_2$, of size 1 is obviously generated by a probability distribution table.

Theorems 3, 4 and 5, and Remark 3 allow a complete characterization of such codes.

4.3 Construction of a Dinucleotide Code From Nucleotide Probabilities of the Code

Any dinucleotide code generated by some probability distribution table is also generated by a special distribution table, namely the *nucleotide distribution table* which is defined as follows.

Construction 3 Let $X \subseteq \mathcal{B}^2$ and $|X| = n$. For $N \in \mathcal{B}$ let

$$p(N) = \mid \{N' \in \mathcal{B} \mid NN' \in X\} \mid$$

and

$$q(N) = \mid \{N' \in \mathcal{B} \mid N'N \in X\} \mid .$$

Then the *nucleotide distribution table Q of X* is defined as

Base	Pos. 1	Pos. 2
N_1	$\frac{p(N_1)}{n}$	$\frac{q(N_1)}{n}$
N_2	$\frac{p(N_2)}{n}$	$\frac{q(N_2)}{n}$
N_3	$\frac{p(N_3)}{n}$	$\frac{q(N_3)}{n}$
N_4	$\frac{p(N_4)}{n}$	$\frac{q(N_4)}{n}$
Σ	1	1

As a consequence of the the above results, we have the following theorem.

Theorem 6 *A dinucleotide code $X \subseteq \mathcal{B}^2$ is generated by some probability distribution table P if and only if it is generated by its nucleotide distribution table Q.*

Proof One implication is trivial since any code generated by its nucleotide distribution table is also generated by a distribution table (in fact, the same table). Assume conversely, that $X \subseteq \mathcal{B}^2$ is generated by an arbitrary distribution table. By Theorems 3, 4 and 5 we have that X is of a completely determined form, namely (all N_i are different)

(1) $|X|= 6 \Longrightarrow X = \{N_1N_2, N_1N_3, N_1N_4, N_2N_3, N_2N_4, N_3N_4\}$
(2) $|X|= 5 \Longrightarrow$
  (a) $\{N_1N_2, N_1N_3, N_1N_4, N_2N_4, N_3N_4\}$ or
  (b) $\{N_1N_2, N_1N_3, N_1N_4, N_2N_3, N_2N_4\}$ or
  (c) $\{N_1N_3, N_1N_4, N_2N_3, N_2N_4, N_3N_4\}$.
(3) $|X|= 4 \Longrightarrow \{N_1N_3, N_1N_4, N_2N_3, N_2N_4\}$.
(4) $|X|= 3 \Longrightarrow$
  (a) $X = \{N_1N_2, N_1N_3, N_1N_4\}$ or
  (b) $X = \{N_1N_2, N_2N_3, N_1N_3\}$ or
  (c) $X = \{N_2N_1, N_3N_1, N_4N_1\}$.
(5) $|X|= 2 \Longrightarrow$
  (a) $X = \{N_1N_2, N_1N_3\}$ or
  (b) $X = \{N_2N_1, N_3N_1\}$.
(6) $|X|= 1 \Longrightarrow X = \{N_1N_2\}$.

It is now straight forward to check that in each case, the codes are also generated by their nucleotide distribution table. For example, in case (1), the code is generated by

Base	Pos. 1	Pos. 2
N_1	$\frac{3}{6}$	0
N_2	$\frac{2}{6}$	$\frac{1}{6}$
N_3	$\frac{1}{6}$	$\frac{2}{6}$
N_4	0	$\frac{3}{6}$
Σ	1	1

in case (2)(a), the code is generated by

Base	Pos. 1	Pos. 2
N_1	$\frac{3}{5}$	0
N_2	$\frac{1}{5}$	$\frac{1}{5}$
N_3	$\frac{1}{5}$	$\frac{1}{5}$
N_4	0	$\frac{3}{5}$
Σ	1	1

in case (3), the code is generated by

Base	Pos. 1	Pos. 2
N_1	$\frac{2}{4}$	0
N_2	$\frac{2}{4}$	0
N_3	0	$\frac{2}{4}$
N_4	0	$\frac{2}{4}$
Σ	1	1

All other cases are similar. □

5 Biological Application: Dinucleotide Circular Codes Observed in Gene Codons and Constructed Theoretically

5.1 Codon Usage of Bacterial, Archaeal and Eukaryotic Genes

From the codon statistics database (CSD) (Appendix A), the codon usage is computed in 34,020,997 genes of 8345 bacterial genomes Bac, 1,280,890 genes of 432 archaeal genomes Arc and 20,206,058 genes of 1150 eukaryotic genomes Euk (Table 3 of Appendix B). The massive statistics of genes in the three kingdoms of bacteria, archaea and eukaryotes given in Table 3 of Appendix B, also enable the reader to carry out other research studies.

5.2 Dinucleotide Circular Codes Observed in Codons of Bacterial, Archaeal and Eukaryotic Genes

From the codon usage Table 3 of Appendix B, the frequency f of the 16 dinucleotides in the 3 codon sites of 34,020,997 genes of 8345 bacterial genomes Bac, 1,280,890 genes of 432 archaeal genomes Arc and 20,206,058 genes of 1150 eukaryotic genomes Euk, are computed: $f_{1,2}$ in the codon site $1-2$ and $f_{2,3}$ in the codon site $2-3$ (Table 4 of Appendix C).

We briefly recall the approach and the results recently obtained in Fimmel et al. (2024). As the codon site $2-3$ is a shift of 1 nucleotide from the codon site $1-2$, the dinucleotides of highest frequency are assigned in the codon sites $1-2$ or $2-3$ by comparing their frequencies $f_{1,2}$ and $f_{2,3}$ (in bold in Table 4 of Appendix C). For example in bacteria, the dinucleotide AC has $f_{1,2} = 5.76\%$ and $f_{2,3} = 8.05\%$, and thus AC is assigned to the codon site $2-3$, etc. The periodic dinucleotides $\{AA, CC, GG, TT\}$ are not considered for a search of dinucleotide circular codes. Very surprisingly, two maximal dinucleotide circular codes are identified

$$D_{1,2} = \{AT, CA, CT, GA, GC, GT\}$$

in the codon site $1-2$ and

$$D_{2,3} = \{AC, AG, CG, TA, TC, TG\}$$

in the codon site $2-3$, with the additional property

$$D_{2,3} = \alpha_1(D_{1,2}) = \overleftarrow{D_{1,2}}.$$

Importantly, the same maximal circular codes $D_{1,2}$ and $D_{2,3}$ are observed in genes of bacteria, archaea as well as eukaryotes, without any exception in the three kingdoms (Table 4 of Appendix C).

5.3 Dinucleotide Circular Codes Constructed Theoretically

5.3.1 Nucleotide Frequencies in the 3 Codon Sites of Bacterial, Archaeal and Eukaryotic Genes

In Table 1, the frequency p (%) of the 4 nucleotides in the 3 codon sites of 34,020,997 genes of 8345 bacterial genomes Bac, 1,280,890 genes of 432 archaeal genomes Arc and 20,206,058 genes of 1150 eukaryotic genomes Euk, are computed from the codon usage Table 3 of Appendix B: p_1 in the codon site 1, p_2 in the codon site 2 and p_3 in the codon site 3.

5.3.2 Dinucleotide Circular Code Constructed By the Nucleotide Probabilities

By assuming independence, the probability $p_{1,2}(N_1N_2)$ of a dinucleotide N_1N_2 in the codon site $1-2$ is obviously equal to the nucleotide probability product

$$p_{1,2}(N_1N_2) = p_1(N_1) \cdot p_2(N_2).$$

From Table 1, the frequency $p_{1,2}$ (%) of the 16 dinucleotides in the codon site $1-2$ of 34,020,997 genes of 8345 bacterial genomes Bac, 1,280,890 genes of 432 archaeal

Table 1 Frequency p (%) of the 4 nucleotides in the 3 codon sites of 34,020,997 genes of 8345 bacterial genomes Bac (11,087,876,805 codons), 1,280,890 genes of 432 archaeal genomes Arc (367,937,932 codons) and 20,206,058 genes of 1150 eukaryotic genomes Euk (10,374,305,634 codons) computed from the codon usage in Table 3 of Appendix B: p_1 in the codon site 1, p_2 in the codon site 2 and p_3 in the codon site 3. The numbers are rounded to the 2nd decimal place

	Bacteria			Archaea			Eukaryota		
	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3
A	24.09	28.11	16.28	23.55	29.57	15.68	27.78	31.44	20.92
C	22.92	24.66	35.04	20.87	23.77	39.38	22.71	23.81	27.73
G	37.64	18.29	30.20	41.25	18.06	29.95	31.57	18.30	26.96
T	15.35	28.94	18.48	14.32	28.60	14.99	17.94	26.45	24.39

genomes Arc and 20,206,058 genes of 1150 eukaryotic genomes Euk are computed in Table 2.

By Construction 2, the dinucleotide code $E_{1,2}$ in the codon site $1-2$ can be identified

$$E_{1,2} = \{AT, CA, CT, GA, GC, GT\}.$$

Importantly, the constructed dinucleotide code $E_{1,2}$ is a maximal circular code and is equal to the observed dinucleotide code $D_{1,2}$

$$E_{1,2} = D_{1,2}.$$

$E_{1,2}$ is observed in genes of bacteria, archaea as well as eukaryotes, without any exception in the three kingdoms (Table 2).

6 A Theoretical Model of a Possible Precursor to the Genetic Code

We propose a theoretical model of a "primitive" genetic code based on the 2 dinucleotide codes identified in Sect. 5.2

$$D_{1,2} = \{AT, CA, CT, GA, GC, GT\}$$

and its deduced reversed code

$$\overleftarrow{D_{1,2}} = \{AC, AG, CG, TA, TC, TG\} = D_{2,3}.$$

Table 2 Frequency $p_{1,2}$ (%) in the codon site $1-2$ of the 16 dinucleotides computed from the nucleotide frequency product $p_1 \cdot p_2$ of Table 1 for 34,020,997 genes of 8345 bacterial genomes Bac (11,087,876,805 codons), 1,280,890 genes of 432 archaeal genomes Arc (367,937,932 codons) and 20,206,058 genes of 1150 eukaryotic genomes Euk (10,374,305,634 codons). The highest dinucleotide frequencies $p_{1,2}$ between a dinucleotide and its permuted dinucleotide are in bold, the periodic dinucleotides being not considered

	Bacteria	Archaea	Eukaryota
AA	6.77	6.96	8.73
AC	5.94	5.60	6.61
AG	4.41	4.25	5.08
AT	**6.97**	**6.74**	**7.35**
CA	**6.44**	**6.17**	**7.14**
CC	5.65	4.96	5.41
CG	4.19	3.77	4.16
CT	**6.63**	**5.97**	**6.01**
GA	**10.58**	**12.20**	**9.93**
GC	**9.28**	**9.81**	**7.52**
GG	6.88	7.45	5.78
GT	**10.89**	**11.80**	**8.35**
TA	4.31	4.23	5.64
TC	3.79	3.40	4.27
TG	2.81	2.59	3.28
TT	4.44	4.10	4.75

We now build the concatenation of $D_{1,2}$ and $\overleftarrow{D_{1,2}}$ by merging dinucleotides of the form N_1N and NN_2 to N_1NN_2. This operation is denoted by $\diamond$.

Definition 5 Let $D_1, D_2 \subseteq \mathcal{B}^2$, then $D_1 \diamond D_2 = \{N_1N_2N_3 \mid N_1N_2 \in D_1, N_2N_3 \in D_2\}$.

Thus the trinucleotide code $\mathcal{T}$ is equal to

$$\mathcal{T} = D_{1,2} \diamond \overleftarrow{D_{1,2}} =$$
$$\{ATA, ATC, ATG, CAC, CAG, CTA, CTC, CTG, GAC, GAG, GCG, GTA, GTC, GTG\}$$

and satisfies interesting properties:

Proposition 9 *The trinucleotide code* $\mathcal{T} = D_{1,2} \diamond \overleftarrow{D_{1,2}}$ *satisfies the following properties:*

- $|\mathcal{T}| = 14$.
- $\mathcal{T}$ is comma-free.
- $\mathcal{T}$ is C^3: the 2 shifted trinucleotide codes are circular (but not comma-free).
- $\mathcal{T} = \overleftarrow{\mathcal{T}}$.
- $\pi_{1,2}(\mathcal{T}) = D_{1,2}$.
- $\pi_{2,3}(\mathcal{T}) = \overleftarrow{D_{1,2}} = D_{2,3}$.

The trinucleotide code $\mathcal{T}$ has strong error detecting properties: comma-freeness and C^3. However, $\mathcal{T}$ is not self-complementary. In order to have a closure under self-complementarity, we add the complementary codons of $\mathcal{T}$ - not reversed - to $\mathcal{T}$ leading to a trinucleotide code $\mathcal{T}'$

$$\mathcal{T}' = \{ATA, ATC, ATG, CAC, CAG, CAT, CGC, CTA, CTC, CTG,$$
$$GAC, GAG, GAT, GCG, GTA, GTC, GTG, TAC, TAG, TAT\}.$$

The trinucleotide code $\mathcal{T}'$ looses its error detecting properties, e.g. the codons *ATC* and *CAT* are permuted, but it retrieves additional properties of the genetic code.

Proposition 10 *The code* $\mathcal{T}'$ *has the following properties:*

- $|\mathcal{T}'| = 20$.
- $\mathcal{T}' = \overleftarrow{\mathcal{T}'}$.
- $\mathcal{T}'$ is self-complementary.
- $\mathcal{T}'$ contains the start codon *ATG* and the stop codon *TAG*.
- $\pi_{1,2}(\mathcal{T}') = D_{1,2} \cup \{TA, CG\}$.
- $\pi_{2,3}(\mathcal{T}') = \overleftarrow{D_{1,2}} \cup \{AT, GC\} = D_{2,3} \cup \{AT, GC\}$.

This trinucleotide code $\mathcal{T}'$ with its property could form a double helix structure and even could start and terminate a translational process.

7 Discussion and Conclusion

A maximal C^3 self-complementary trinucleotide circular code X is identified in genes of bacteria, archaea, eukaryotes, plasmids and viruses (Arquès and Michel 1996; Michel 2017). Circular codes are defined by their ability to recover the correct reading frame of genes at any position. From an evolutionary perspective, particularly in the context of the prebiotic soup, it can be hypothesized that trinucleotide circular codes, composed of 3-letter words, emerged from simpler circular codes such as dinucleotide circular codes, composed of 2-letter words. Another possible alternative, which is beyond the scope of this article, would be circular codes defined over a reduced alphabet with 2- or 3-letter words, for example a 2-letter purine–pyrimidine alphabet.

In this work, we demonstrated that several classes of dinucleotide circular codes can be constructed from nucleotide probability products, following Construction 2. We introduced the new concept of transitive dinucleotide codes and established a set of theorems linking transitivity to circularity and comma-free properties. By analysing codon usage across thousands of bacterial, archaeal and eukaryotic genomes, 2 "universal" maximal dinucleotide circular codes are consistently observed in the codon positions $1-2$ and $2-3$, without exception in the three kingdoms: $D_{1,2} = \{AT, CA, CT, GA, GC, GT\}$ in the codon site $1-2$ and $D_{2,3}$ in the codon site $2-3$. Notably, $D_{2,3}$ can be derived from $D_{1,2}$ by 1-letter cyclical permutation α_1 or identically by reversing permutation $D_{2,3} = \alpha_1(D_{1,2}) = \overleftarrow{D_{1,2}}$.

Remarkably, assuming independence of nucleotide sites, we further show that the dinucleotide code $E_{1,2}$ through Construction 2 from products of nucleotide frequencies in the codon sites 1 and 2, is a maximal dinucleotide circular code and exactly matches the observed dinucleotide code: $E_{1,2} = D_{1,2}$.

From a biological perspective, these findings suggest that dinucleotide circular codes may naturally arise from nucleotide distributions, as could have occurred in prebiotic nucleotide mixtures. Notably, the maximal dinucleotide circular code $D_{1,2}$ observed consistently across the three kingdoms of life which is equal to theoretical code $E_{1,2}$, is located at the codon position $1-2$. All these result align with the biological hypothesis proposed by Copley et al. (2005), who suggested that dinucleotides may have played a catalytic role in prebiotic amino acid synthesis, thus prefiguring the modern codon–amino acid relationship. According to their model, the 3rd nucleotide in present-day codons was likely a later evolutionary addition, with the initial dinucleotide N_1N_2 functioning as a proto-codon, and the 3rd nucleotide N_3 later introducing redundancy into the genetic code. A dinucleotide may have contributed to the synthesis of the same amino acids as their corresponding modern codons, e.g. *GG* for glycine (Gly G), and *CG* for ornithine, a simpler analogue of arginine (Arg R). In some cases, a dinucleotide may have coded ancestral amino acids that were chemically simpler precursors of modern ones, e.g. *AG* for 2,4-diaminobutyric

acid instead of serine (Ser S), and AC for homoserine, a precursor to threonine (Thr T). Further support for this hypothesis comes from quantum mechanical calculations using density functional theory (DFT), which suggest that such dinucleotide-based pathways for amino acid synthesis are chemically plausible (Yaman and Harvey 2021). The mathematical framework developed in this study, grounded in dinucleotide circular codes, provides an additional theoretical contribution supporting this prebiotic model for the origin of the genetic code.

Appendix A: Data

A codon statistics database (CSD) has recently been developed by the Alvarez-Ponce group (Subramanian et al. 2022). 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 (July 2022) the codon usage of genomes of three kingdoms: bacteria, archaea and eukaryota. The few exceptional genomes in which the codon usage of the stop codons is not given, are not considered.

The Archaea (Id 2157) can be directly extracted leading to 432 archaeal genomes Arc with 1,280,890 genes representing 367,937,932 codons.

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). This data construction leads to 8345 bacterial genomes Bac with 34,020,997 genes representing 11,087,876,805 codons.

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). This data construction leads to 1150 eukaryotic genomes Euk with 20,206,058 genes representing 10,374,305,634 codons.

Appendix B: Codon usage of bacterial, archaeal and eukaryotic genes

See Table 3.

Table 3 Codon usage (frequency in %) of 34,020,997 genes of 8345 bacterial genomes Bac (11,087,876,805 codons), 1,280,890 genes of 432 archaeal genomes Arc (367,937,932 codons) and 20,206,058 genes of 1150 eukaryotic genomes Euk (10,374,305,634 codons) obtained from the codon statistics database (CSD) (see Appendix A)

	Bac	Arc	Euk		Bac	Arc	Euk		Bac	Arc	Euk		Bac	Arc	Euk
AAA	2.37	1.74	2.60	*CAA*	1.20	0.59	1.61	*GAA*	2.91	3.04	2.96	*TAA*	0.12	0.11	0.06
AAC	1.80	2.01	2.12	*CAC*	1.27	1.40	1.37	*GAC*	3.42	5.25	2.57	*TAC*	1.56	2.13	1.55
AAG	2.00	1.88	3.12	*CAG*	2.16	1.78	2.75	*GAG*	3.22	5.07	3.69	*TAG*	0.06	0.08	0.05
AAT	1.64	1.05	1.95	*CAT*	0.78	0.46	1.17	*GAT*	2.25	2.02	2.64	*TAT*	1.35	0.96	1.24
ACA	0.93	0.85	1.57	*CCA*	0.63	0.67	1.61	*GCA*	1.43	1.48	1.76	*TCA*	0.64	0.68	1.38
ACC	2.56	2.16	1.63	*CCC*	1.39	1.35	1.49	*GCC*	4.32	3.49	2.23	*TCC*	1.25	1.08	1.60
ACG	1.51	2.20	0.94	*CCG*	2.14	1.91	0.97	*GCG*	3.22	3.42	1.12	*TCG*	1.23	1.62	0.90
ACT	0.76	0.73	1.40	*CCT*	0.67	0.52	1.58	*GCT*	1.30	1.05	2.02	*TCT*	0.67	0.56	1.60
AGA	0.54	0.69	1.27	*CGA*	0.43	0.72	0.71	*GGA*	1.39	1.52	1.76	*TGA*	0.15	0.19	0.08
AGC	1.28	1.27	1.72	*CGC*	2.26	1.80	1.03	*GGC*	3.63	3.40	1.92	*TGC*	0.53	0.41	1.05
AGG	0.38	0.81	1.13	*CGG*	1.64	1.51	0.82	*GGG*	1.43	1.78	1.26	*TGG*	1.31	1.08	1.22
AGT	0.69	0.63	1.27	*CGT*	0.80	0.50	0.68	*GGT*	1.54	1.34	1.43	*TGT*	0.29	0.42	0.95
ATA	0.82	1.33	1.02	*CTA*	0.45	0.47	0.77	*GTA*	1.05	0.86	0.84	*TTA*	1.25	0.75	0.92
ATC	2.76	2.90	2.04	*CTC*	2.23	3.86	1.83	*GTC*	2.66	4.30	1.58	*TTC*	2.11	2.58	2.01
ATG	2.15	2.00	2.26	*CTG*	3.86	2.25	2.87	*GTG*	2.66	1.88	2.28	*TTG*	1.21	0.68	1.58
ATT	1.91	1.31	1.75	*CTT*	1.01	1.09	1.46	*GTT*	1.20	1.36	1.51	*TTT*	1.61	1.00	1.73

Appendix C: Dinucleotide frequency in the codon sites of bacterial, archaeal and eukaryotic genes

See Table 4.

Table 4 Frequency f (%) of the 16 dinucleotides in the 3 codon sites of 34,020,997 genes of 8345 bacterial genomes Bac (11,087,876,805 codons), 1,280,890 genes of 432 archaeal genomes Arc (367,937,932 codons) and 20,206,058 genes of 1150 eukaryotic genomes Euk (10,374,305,634 codons), computed from the codon usage in Table 3 of Appendix B: $f_{1,2}$ in the codon site $1-2$ and $f_{2,3}$ in the codon site $2-3$. The highest dinucleotide frequencies between $f_{1,2}$ and $f_{2,3}$ are in bold. One highest dinucleotide frequency among 96 that does not comply with the rule is in bold and italics: $f_{2,3}=6.45\%$ with *TT* in eukaryota, a periodic dinucleotide not involved in dinucleotide circular codes

	Bacteria		Archaea		Eukaryota	
	$f_{1,2}$	$f_{2,3}$	$f_{1,2}$	$f_{2,3}$	$f_{1,2}$	$f_{2,3}$
AA	**7.80**	6.59	**6.68**	5.48	**9.78**	7.23
AC	5.76	**8.05**	5.94	**10.79**	5.54	**7.60**
AG	2.89	**7.45**	3.39	**8.81**	5.39	**9.60**
AT	**7.63**	6.02	**7.54**	4.49	**7.07**	7.00
CA	**5.42**	3.63	**4.23**	3.68	**6.89**	6.32
CC	4.83	**9.52**	4.45	**8.08**	5.66	**6.95**
CG	5.13	**8.11**	4.53	**9.15**	3.24	**3.94**
CT	**7.55**	3.40	**7.67**	2.86	**6.93**	6.60
GA	**11.80**	2.51	**15.38**	3.11	**11.86**	3.82
GC	**10.28**	7.70	**9.44**	6.88	**7.13**	5.71
GG	**7.98**	4.76	**8.03**	5.18	**6.37**	4.43
GT	**7.57**	3.32	**8.40**	2.89	**6.21**	4.34
TA	3.09	**3.56**	3.27	**3.40**	2.91	**3.55**
TC	3.79	**9.77**	3.95	**13.64**	5.48	**7.46**
TG	2.29	**9.88**	2.10	**6.80**	3.31	**8.99**
TT	**6.18**	5.73	**5.00**	4.76	6.24	***6.45***

Acknowledgements The first and last authors would like to acknowledge the contribution of the COST Action CA21169, supported by COST (European Cooperation in Science and Technology).

Author Contributions All authors wrote the main manuscript text.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflicts of Interest The authors report no conflict of interest.

References

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

Copley SD, Smith E, Morowitz HJ (2005) A mechanism for the association of amino acids with their codons and the origin of the genetic code. Proc Natl Acad Sci U S A 102:4442–4447

Fimmel E, Giannerini S, Gonzalez D, Strüngmann L (2015) Dinucleotide circular codes and bijective transformations. J Theor Biol 386:159–165

Fimmel E, Michel CJ, Strüngmann L (2016) n-Nucleotide circular codes in graph theory. Philos Trans Royal Soc A Math Phys Eng Sci A 374(20150058):1–19

Fimmel E, Michel CJ, Strüngmann L (2017a) Strong comma-free codes in genetic information. Bull Math Biol 79(8):1796–1819

Fimmel E, Michel CJ, Strüngmann L (2017b) Diletter circular codes over finite alphabets. Math Biosci 294:120–129

Fimmel E, Michel CJ, Strüngmann L (2024) Circular cut codes in genetic information. Biosystems 243(105263):1–10

Fimmel E, Strüngmann L (2016) Maximal dinucleotide comma-free codes. J Theor Biol 389:206–213

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

Koch AJ, Lehmann J (1997) About a symmetry of the genetic code. J Theor Biol 189:171–174

Lacan J, Michel CJ (2001) Analysis of a circular code model. J Theor Biol 213:159–170

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(20):1–16

Michel CJ, Pirillo G (2013) Dinucleotide circular codes. ISRN Biomathematics 2013, ID 538631, 1–8

Michel CJ, Pellegrini M, Pirillo G (2016) Maximal dinucleotide and trinucleotide circular codes. J Theor Biol 389:40–46

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

Yaman T, Harvey JN (2021) Computational analysis of a prebiotic amino acid synthesis with reference to extant codon–amino acid relationships. Life 11(1343):1–16

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Elena Fimmel[1] · Christian J. Michel[2] · Lutz Strüngmann[1]

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

Elena Fimmel
e.fimmel@hs-mannheim.de

Lutz Strüngmann
l.struengmann@hs-mannheim.de

[1] Institute of Mathematical Biology, Faculty for Computer Sciences, Mannheim University of Applied Sciences, 68163 Mannheim, Germany

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