ANALYSIS OF A CIRCULAR CODE MODEL

Jérôme Lacan¹ and Christian J. Michel²*

¹ Laboratoire d’Informatique de Franche-Comté
UNIVERSITE DE FRANCHE-COMTE
IUT de Belfort-Montbéliard
4 Place Tharradin - BP 71427
25211 Montbéliard Cedex, France
Tel.: (33) 3 81 99 47 34
Email: jerome.lacan@pu-pm.univ-fcomte.fr

²* Equipe de Biologie Théorique
UNIVERSITE LOUIS PASTEUR STRASBOURG
Pôle API
Boulevard Sébastien Brant
67400 ILLKIRCH - STRASBOURG, FRANCE
Tel.: (33) 3 90 24 44 62
Email: michel@dpt-info.u-strasbg.fr

* Corresponding author
ABSTRACT

A circular code has been identified in the protein (coding) genes of both eukaryotes and prokaryotes by using a statistical method called Trinucleotide Frequency method (TF method) [Arquès & Michel, (1996) J. Theor. Biol. 182, 45-58]. Recently, a probabilistic model based on the nucleotide frequencies with a hypothesis of absence of correlation between successive bases on a DNA strand, has been proposed by Koch & Lehmann [(1997) J. Theor. Biol. 189, 171-174] for constructing some particular circular codes. Their interesting method which we call here Nucleotide Frequency method (NF method), reveals several limits for constructing the circular code observed with protein genes.
1 INTRODUCTION

This section is divided into 2 parts. The first part summarizes the results of the circular code \(X_0\) identified in the protein genes of both eukaryotes and prokaryotes. The second part recalls the probabilistic model of Koch & Lehmann (1997) based on the Nucleotide Frequency method (NF method).

1.1 The circular code \(X_0\)

The concept of code "without commas" introduced by Crick et al. (1957) for the protein (coding) genes, is a code readable in only one out of three frames. Such a theoretical code without commas, called circular code in theory of codes (e.g. Béal, 1993; Berstel & Perrin, 1985), is a particular set \(X\) of trinucleotides so that a concatenation (a series) of trinucleotides of \(X\), leads to sequences which cannot be decomposed in another frame with a concatenation of trinucleotides of \(X\).

For example, suppose that \(X\) is the following set of trinucleotides: \(X=\{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}\). Some trinucleotides of \(X\) are randomly concatenated, for example as follows:

\[
\ldots CAG, GCC, TTC, AAT, ACC, ACC, CAG, GAA, GAG, GTA, ATT, ACC, AAT, GTA, AAC, TAC, TTC, ACC, ATC \ldots
\]

The commas between the trinucleotides show the frame of construction (reading frame in biology). Suppose now that the commas are "lost" leading to the sequence:

\[
\ldots CAGGCCCTTCAATACCACCCAGGAAAGGTATATTACCAATGTAAACTACTTTACCATC \ldots
\]

The problem is to retrieve the original frame of construction. There are 3 obvious possibilities:

\[
\ldots C, AGG, CCT, TCA, ATA, CCA, CCA, AGG, AAG, AAG, TAA, TTA, CCA, ATG, TAA, ACT, ACT, TCA, CCA, TC \ldots
\]

\[
\ldots CA, GGC, CTT, CAA, TAC, CAC, CCA, GGA, AGA, AGA, GGT, AAT, TAC, CAA, TGT, AAA, CTA, CTT, CAC, CAT, C \ldots
\]

\[
\ldots CAG, GCC, TTC, AAT, ACC, ACC, CAG, GAA, GAG, GTA, ATT, ACC, AAT, GTA, AAC, TAC, TTC, ACC, ATC \ldots
\]

If the set \(X\) of trinucleotides is a circular code, then there is an unique solution:

\[
\ldots CAGGCCTTCAATACCACCCAGGAAAGGTATATTACCAATGTAAACTACTTTACCATC \ldots
\]

This unique solution is obtained by choosing a window (sufficiently large) in any position in the sequence and then, to verify the belonging of the trinucleotides of the window to \(X\):

\[
\ldots CAGGCCCTTCAATACCACCCAGGAAAGGTATATTACCAATGTAAACTACTTTACCATC \ldots
\]

\[
\ldots CAGGCCCTTCAATACCACCCAGGAAAGGTATATTACCAATGTAAACTACTTTACCATC \ldots
\]

\[
\ldots CAGGCCCTTCAATACCACCCAGGAAAGGTATATTACCAATGTAAACTACTTTACCATC \ldots
\]

The first decomposition proposed is rejected immediately as the first trinucleotides AGG in the window does not belong to \(X\). The second decomposition proposed is rejected with a window of 13 nucleotides. Indeed, the first nucleotide A in the window may belong to several trinucleotides of \(X\), e.g. GTA. The trinucleotides GGT, AAT and TAC following A belong to \(X\). The next trinucleotide CAA does not belong to \(X\) as the 13th nucleotide A (from the beginning of the window) differs from the unique possibility G of CAG belonging to \(X\). The third decomposition is the original one as all the trinucleotides in the window belong to \(X\). The original decomposition of the sequence is automatically deduced.

Such a code was proposed by Crick et al. (1957) in order to explain how the reading of a series of nucleotides in the protein genes could code for the amino acids constituting the proteins. The 2 problems stressed were: why are there more trinucleotides than amino acids and how to choose the reading frame? Crick et al. (1957) have then proposed that only 20 among 64 trinucleotides code for the 20 amino acids.
However, the determination of a set of 20 trinucleotides forming a circular code \( X \) depends on a great number of constraints:

(i) A trinucleotide with identical nucleotides (AAA, CCC, GGG or TTT) must be excluded from such a code. Indeed, the concatenation of AAA with itself does not allow to retrieve the reading (original) frame as there are 3 possible decompositions: \(...)AAA,AAA,AAA,...\) and \(...)AA,AAA,AAA,A...

(ii) Two trinucleotides related to circular permutation, e.g. ATC and TCA, must be excluded from such a code. Indeed, the concatenation of ATC with itself does not allow the retrieval of the reading (original) frame as there are 2 possible decompositions: \(...)ATC,ATC,ATC,...\) and \(...)A,TCA,TCA,TC...

Therefore, by excluding AAA, CCC, GGG and TTT and by gathering the 60 remaining trinucleotides in 20 classes of 3 trinucleotides so that, in each class, the 3 trinucleotides are deduced from each other by circular permutations, e.g. ATC, TCA and CAT, a circular code has only one trinucleotide per class and therefore contains at most 20 trinucleotides (maximal circular code). This trinucleotide number is identical to the amino acid number leading to a circular code assigning one trinucleotide per amino acid.

No set of 20 trinucleotides leading to a circular code has been found at this time. Furthermore, the 2 discoveries that the trinucleotide TTT, an "excluded" trinucleotide in the concept of circular code, codes for phenylalanine (Nirenberg & Matthaei, 1961) and that the protein genes are placed in the reading frame with a particular trinucleotide, namely the start trinucleotide ATG, have led to give up the concept of circular code on the alphabet \( \{A,C,G,T\} \). For several biological reasons, in particular the interaction between mRNA and tRNA, the concept of circular code is resumed later on the alphabet \( \{R,Y\} \) (R=purine=\( A \) or \( G \), \( Y=pyrimidine=C \) or \( T \)) with 2 trinucleotide models for the primitive protein genes: \( RRY \) (Crick et al., 1976) and \( RNY \) (\( N=R \) or \( Y \)) (Eigen & Schuster, 1978).

Unexpectedly, a maximal circular code has recently been identified in the protein genes of both eukaryotes and prokaryotes on the alphabet \( \{A,C,G,T\} \) (Arquès & Michel, 1996). This circular code has been obtained by 2 methods:

(i) by computing the occurrence frequencies of the 64 trinucleotides AAA,...,TTT in the 3 frames of protein genes and then, by assigning each trinucleotide to the frame associated with its highest frequency (Arquès & Michel, 1996). This Trinucleotide Frequency method is called TF method.

(ii) by computing the 12288 (\( 3\times64^2 \)) autocorrelation functions analysing the probability that a trinucleotide in any frame occurs any \( i \) bases \( N \) after a trinucleotide in a given frame of protein genes and then, by classifying these autocorrelation functions according to their modulo 3 periodicity for deducing a frame for each trinucleotide (Arquès & Michel, 1997a).

The maximal circular code identified is the set \( X_0 = \{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,GAG,GAT,GCC,GGC,GGT,GTA,GTC,GTG,TAC,TCA,GCT \} \) of 20 trinucleotides in frame 0 of protein genes (reading frame). Furthermore, the 2 sets \( X_1 \) and \( X_2 \) of 20 trinucleotides identified in the frames 1 and 2 respectively (frames 1 and 2 being the frame 0 shifted by 1 and 2 nucleotides respectively in the 5'-3' direction) by these 2 methods, are also maximal circular codes (Table 1a). These 3 circular codes have several important properties:

(i) circularity: \( X_0 \) generates \( X_1 \) by one circular permutation and \( X_2 \) by another circular permutation (1 and 2 circular permutations of each trinucleotide of \( X_0 \) lead to the trinucleotides of \( X_1 \) and \( X_2 \) respectively) (Table 1b).
(ii) complementarity: \( X_0 \) is self-complementary (10 trinucleotides of \( X_0 \) are complementary to the 10 other trinucleotides of \( X_0 \)) and, \( X_1 \) and \( X_2 \) are complementary to each other (the 20 trinucleotides of \( X_1 \) are complementary to the 20 trinucleotides of \( X_2 \)) (Table 1c). Note that this property is also verified with \( T_0 = X_0 \cup \{AAA, TTT\} \) and, \( T_1 = X_1 \cup \{CCC\} \) and \( T_2 = X_2 \cup \{GGG\} \) (Table 1c).

(iii) rarity: the occurrence probability of \( X_0 \) is equal to \( 6 \times 10^{-8} \). As there are 20 classes of 3 trinucleotides (see above), the number of potential circular codes is \( 3^{20} = 3486784401 \). The computed number of complementary circular codes with 2 shifted circular codes (called \( C^3 \) codes), such as \( X_0 \), is 216. Therefore, its probability is \( 216/3^{20} = 6 \times 10^{-8} \).

(iv) flexibility:
- the lengths of the minimal windows to retrieve automatically the frames 0, 1 and 2 with the 3 circular codes \( X_0 \), \( X_1 \) and \( X_2 \) respectively, are all equal to 13 nucleotides and represent the largest window length among the 216 \( C^3 \) codes.
- the frequency of misplaced trinucleotides in the shifted frames is equal to 24.6%. If the trinucleotides of \( X \) are randomly concatenated, for example as follows:
  
  \[ \ldots GAA, GAG, GTA, ACC, AAT, GTA, CTC, TAC, TTC, ACC, ATC \ldots \]

  then, the trinucleotides in frame 1:
  
  \[ \ldots G, AAC, TAG, TAA, CCA, ATG, TAC, TCT, ACT, TCA, CCA, AAC, ATC \ldots \]

  and the trinucleotides in frame 2:
  
  \[ \ldots GA, AGA, GGG, TGT, ACT, TCT, CTA, CTT, CAT, CAC, CAT, C \ldots \]

  mainly belong to \( X_1 \) and \( X_2 \) respectively. A few trinucleotides are misplaced in the shifted frames. With this example, in frame 1, 9 trinucleotides belong to \( X_1 \), 1 trinucleotide (TAC) to \( X_0 \) and 1 trinucleotide (TAA) to \( X_2 \). In frame 2, 8 trinucleotides belong to \( X_2 \), 2 trinucleotides (GGT, AAC) to \( X_0 \) and 1 trinucleotide (ACT) to \( X_1 \). By computing exactly, the average frequencies of misplaced trinucleotides in frame 1 are 11.9% for \( X_0 \) and 12.7% for \( X_2 \). In frame 2, the average frequencies of misplaced trinucleotides are 11.9% for \( X_0 \) and 12.7% for \( X_1 \). The complementarity property explains on the one hand that the frequency equality of \( X_0 \) in frames 1 and 2 and on the other hand, the frequency equality of \( X_2 \) in frame 1 and \( X_1 \) in frame 2. The sum of percentages of misplaced trinucleotides in frame 1 (\( X_0 \) and \( X_2 \)) is equal to the sum of percentages of misplaced trinucleotides in frame 2 (\( X_0 \) and \( X_1 \)) and is equal to 24.6%. This value is close to the highest frequency (27.9%) of misplaced trinucleotides among the 216 \( C^3 \) codes.
- the 4 types of nucleotides occur in the 3 trinucleotide sites with the 3 circular codes \( X_0 \), \( X_1 \) and \( X_2 \) (Table 1a).

(v) evolutionary: an evolutionary analytical model at 3 parameters \((p, q, t)\) based on an independent mixing of the 20 trinucleotides of \( X_0 \) with equiprobability \((1/20)\) followed by \( t \approx 4 \) substitutions per trinucleotide according to the proportions \( p \approx 0.1, q \approx 0.1 \) and \( r = 1 - p - q \approx 0.8 \) in the 3 trinucleotide sites respectively, retrieves the frequencies of \( X_0 \), \( X_1 \) and \( X_2 \) observed in the 3 frames of protein genes.

The proof that \( X_0 \), \( X_1 \) and \( X_2 \) are circular codes, the detailed explanation of the properties (i-iv) and the different biological consequences, in particular on the 2-letter genetic alphabets, the genetic code and the amino acid frequency frequencies in proteins, are given in Arquès & Michel (1996, 1997a). The property (v) is
described in Arquès et al. (1998, 1999). Note: a non-complementary circular code has recently been identified in the mitochondrial protein genes (Arquès & Michel, 1997b).

1.2 The Nucleotide Frequency method (NF method)

Koch & Lehmann (1997, p. 171) have recently suggested that the self-complementary circular code $X_o$ observed in protein genes could be explained by a method for generating circular codes from nucleotide frequencies. This method called here Nucleotide Frequency method (NF method), is briefly recalled by keeping the same notations.

Let $p_i(\theta)$ be the occurrence probability of a given base $\theta \in \{A,C,G,T\}$ at position $i \in \{1,2,3\}$ in a trinucleotide (triplet) observed in a DNA strand read in frame 0. By supposing that there is no correlation between successive bases on a DNA strand, the probability of finding the triplet $\alpha \beta \gamma$ in the frame 0 is given by the probabilities product $p_i(\alpha)p_i(\beta)p_i(\gamma)$ (independent probabilities). The belonging of the triplet $\alpha \beta \gamma$ to a preferential set $0Y$ of triplets in frame 0 is then equivalent to the 2 following probability inequalities

$$p_i(\alpha)p_i(\beta)p_i(\gamma) > p_i(\gamma)p_i(\alpha)p_i(\beta) \quad (1)$$

and

$$p_i(\alpha)p_i(\beta)p_i(\gamma) > p_i(\beta)p_i(\gamma)p_i(\alpha) \quad (2)$$

Similar probability inequalities imply that the triplet $\beta \gamma \alpha$ (resp. $\gamma \alpha \beta$) belongs to the preferential set $1Y$ (resp. $2Y$) of triplets in frame 1 (resp. 2).

Koch & Lehmann (1997, p. 173) prove that a preferential set generated from any set of probabilities $p_i(\theta)$ with this method, is a circular code.

Koch & Lehmann (1997, p. 172) also show that, if the probabilities $p_i(\theta)$ verify the relation

$$p_i(\theta) = p_3(C(\theta)) \text{ and } p_2(\theta) = p_2(C(\theta)) \quad (3)$$

where $C(\theta)$ denote the complementary base of $\theta$, then the circular code $0Y$ is necessarily self-complementary and the 2 permutated circular codes $1Y$ and $2Y$ are complementary (called C1 codes in Arquès & Michel, 1996).

The Table 1 in Koch & Lehmann (1997) gives the 12 nucleotide observed frequencies $p_i(\theta)$ of a base $\theta \in \{A,C,G,T\}$ at position $i \in \{1,2,3\}$ of the reading frame for the prokaryotes. These data have been obtained from the 44th release of the prokaryotic EMBL database. This Table 1 is recalled in this paper with the Table 2a. These 12 probabilities with the NF method lead to a new circular code $0Y$={$AAT$, $AAC$, $ATT$, $ATC$, $ACT$, $CAC$, $CTT$, $CCT$, $GAA$, $GAT$, $GAC$, $GAG$, $GTA$, $GTC$, $GTG$, $GCA$, $GCT$, $GCC$, $GCG$}. This code $0Y$ contains 13 trinucleotides of the code $X_o$ (Table 1a).
2 METHOD AND RESULTS

2.1 The Nucleotide Frequency method (NF method) cannot generate the circular code $X_0$

2.1.1 The NF method does not generate an unique self-complementary circular code from the observed probabilities

The approach of Koch & Lehmann (1997) tries to link the self-complementary code $X_0$ and the NF method. However, the code $Y_0$ obtained by the NF method from the observed probabilities $p_i(\theta)$ of a base $\theta \in \{A,C,G,T\}$ at position $i \in \{1,2,3\}$ of the reading frame for the prokaryotes is not self-complementary as, for example, $ACT \in Y_0$ but $C(ACT) = AGT \notin Y_0$. So, this section is devoted to obtain a self-complementary circular code with the NF method from probabilities closed to the observed ones.

If the 12 probabilities $p_i(\theta)$ verify the relation (3), then the circular code computed by the NF method is a self-complementary code. However, the relation (3) which contains 6 probability equalities, cannot be easily used with observed probabilities.

Koch & Lehmann (1997, p. 172) have mentioned that the 12 probabilities $p_i(\theta)$ in Table 2a do not precisely verify the relation (3) and then, no self-complementary circular code has been proposed.

Furthermore, the NF method generates several self-complementary circular codes if the probabilities of Table 2a are slightly modified for verifying the relation (3). Three examples of such self-complementary circular codes are presented in Table 2b. The first circular code is obtained with observed frequencies from the first and second columns of Table 2a: $p_1(A) = p_3(T) = 0.276$, $p_1(C) = p_3(G) = 0.204$, $p_1(G) = p_3(C) = 0.354$, $p_1(T) = p_3(A) = 0.166$, $p_2(A) = p_2(T) = 0.285$ and $p_2(C) = p_2(G) = (1 - 2 \times 0.285)/2$.

The second circular code is obtained with observed frequencies from the second and third columns of Table 2a: $p_1(A) = p_3(T) = 0.268$, $p_1(C) = p_3(G) = 0.242$, $p_1(G) = p_3(C) = 0.268$, $p_1(T) = p_3(A) = 0.222$, $p_2(A) = p_2(T) = 0.285$ and $p_2(C) = p_2(G) = (1 - 2 \times 0.285)/2$. The third circular code is obtained with average frequencies from Table 2a: $p_1(A) = p_3(T) = (0.276 + 0.268)/2 = 0.272$, $p_1(C) = p_3(G) = (0.204 + 0.242)/2 = 0.223$, $p_1(G) = p_3(C) = (0.354 + 0.268)/2 = 0.311$, $p_1(T) = p_3(A) = (0.166 + 0.222)/2 = 0.194$, $p_2(A) = p_2(T) = (0.315 + 0.285)/2 = 0.3$ and $p_2(C) = p_2(G) = (0.228 + 0.172)/2 = 0.2$.

In summary, the NF method is not well adapted to reveal an unique self-complementary circular code. Furthermore, we shall prove in the next section that the NF method cannot generate the self-complementary circular code $X_0$ which has been identified in the protein genes of both eukaryotes and prokaryotes (Arquès & Michel, 1996).
2.1.2 Proof that the NF method cannot generate the circular code $X_0$

This section presents a mathematical proof that the NF method cannot generate the circular code $X_0$. The idea of this proof is the following one. We take the hypothesis that a circular code $X$ containing the 3 triplets $\alpha\beta\gamma$, $\delta\delta\beta$ and $\gamma\alpha\delta$ where $\alpha,\beta,\gamma,\delta \in \{A,C,G,T\}$ is generated by the NF method from the occurrence probabilities $p_i(\theta)$ of a base $\theta \in \{A,C,G,T\}$ at the position $i \in \{1,2,3\}$. Then, this hypothesis is refuted by considering several probability inequalities associated with the 3 triplets considered. As the circular code $X_0$ contains such 3 triplets (ATC, GGT, CAG), then $X_0$ cannot be generated by the NF method.

The existence of 3 probabilities $p_i(\theta)$ generating $X$ by the NF method is taking as hypothesis. According to the inequality (1) of the NF method, the triplet $\alpha\beta\gamma$ belonging to $X$ leads to the following probability inequality

$$p_1(\alpha)p_2(\beta)p_3(\gamma) > p_1(\gamma)p_2(\alpha)p_3(\beta) \quad (4)$$

According to the inequality (2) of the NF method, the triplet $\delta\delta\beta$ belonging to $X$ leads to the following probability inequality

$$p_1(\delta)p_2(\delta)p_3(\beta) > p_1(\delta)p_2(\beta)p_3(\delta) \quad (5)$$

Clearly, $p_1(\delta) > 0$ otherwise the inequality (5) cannot be verified. Therefore, by simplifying (5)

$$p_2(\delta)p_3(\beta) > p_2(\beta)p_3(\delta) \quad (6)$$

According to the inequality (2) of the NF method, the triplet $\gamma\alpha\delta$ belonging to $X$ leads to the following probability inequality

$$p_1(\gamma)p_2(\alpha)p_3(\delta) > p_1(\alpha)p_2(\delta)p_3(\gamma) \quad (7)$$

Clearly, $p_3(\delta) > 0$ otherwise the inequality (7) cannot be verified. By rewriting (4) as follows

$$p_1(\alpha)p_2(\beta)p_3(\gamma) > p_1(\gamma)p_2(\alpha)p_3(\delta) \times p_3(\beta)/p_3(\delta) \quad (8)$$

By using (7) with the second member of (8), we obtain

$$p_1(\alpha)p_2(\beta)p_3(\gamma) > p_1(\alpha)p_2(\delta)p_3(\gamma) \times p_3(\beta)/p_3(\delta) \quad (9)$$

As $p_1(\alpha) > 0$ and $p_3(\gamma) > 0$, the inequality (9) can be simplified as follows

$$p_2(\beta) > p_2(\delta) \times p_3(\beta)/p_3(\delta)$$

i.e.

$$p_2(\beta)p_3(\delta) > p_2(\delta)p_3(\beta) \quad (10)$$

The inequality (10) is in contradiction with the inequality (6). Therefore, the hypothesis of existence of 3 probabilities $p_i(\theta)$ generating $X$ is refuted.

This proof can be applied to the circular code $X_0$ containing the 3 triplets ATC, GGT and CAG which follow the pattern $\alpha\beta\gamma$, $\delta\delta\beta$ and $\gamma\alpha\delta$. Therefore, the circular code $X_0$ cannot be generated by the NF method.
2.1.3 Development of two algorithms in complement of the proof

The previous section (ii) has proved that the self-complementary circular code $X_0$ cannot be generated by the NF method. This section consists in determining all the self-complementary circular codes which can be generated by this NF method.

The first algorithm A1 developed allows the determination of a set $S$ of self-complementary circular codes $Y$ based on the NF method. The NF method implies the following property with each code $Y$ of $S_Y$. The 2 sets of 20 words obtained by circular permutations of a code $Y$, are complementary circular codes (Koch & Lehmann, 1997, p. 173). Such codes $Y$ are called $C^3$ codes (Arquès & Michel, 1996).

The principle of the algorithm A1 consists in varying the probabilities $p_i(\theta)$ of the 4 bases at the 3 positions in the range $[0,1]$ according to the relation (3). For each probability variation step, the algorithm A1 computes a $C^3$ code by using the NF method and tests whether this $C^3$ code has been previously generated. Indeed, several sets of probabilities $p_i(\theta)$ can lead to the same $C^3$ codes. By varying the probabilities $p_i(\theta)$ with steps becoming smaller and smaller, the number of $C^3$ codes $Y$ in $S_Y$ remains constant and equal to 88. These 88 codes $Y$ are listed in Table 3.

The algorithm A1 generates 88 $C^3$ codes $Y$. However, the flower automaton method identifies 216 $C^3$ codes (Arquès & Michel, 1996). In order to explain the 216-88=128 remaining $C^3$ codes, we extend the proof (ii) based on the pattern $P_0=$\{$\alpha\beta\gamma$, $\delta\delta\beta$, $\gamma\alpha\delta$\} to its 2 circular permuted patterns $P_1=$\{$\beta\gamma\alpha$, $\delta\delta\beta$, $\alpha\delta\gamma$\}$ and $P_2=$\{$\gamma\alpha\beta$, $\beta\delta\delta$, $\delta\gamma\alpha$\}. Any circular code containing the pattern $P_0$ cannot be generated by the NF method (proof (ii)). Similarly, the proof (ii) also shows that any circular code containing a circular permuted pattern $P_1$ or $P_2$, cannot be generated by the NF method. The algorithm A2 developed determines the $C^3$ codes among the 216 ones which contains at least one of the 3 previous patterns. There are exactly 128 such $C^3$ codes. Therefore, the algorithm A2 confirms the number 88 of $C^3$ codes $Y$ determined by the algorithm A1 whatever the probability variation step used.

In summary, the number of $C^3$ codes which can be generated by the Nucleotide Frequency method is exactly 88. It is important to stress that the 128 other $C^3$ codes cannot be generated from any sets of probabilities, even probabilities which do not verify the relation (3), as the proof (ii) does not make any hypothesis on the probabilities.
2.2 Remarks on the hypothesis of no correlation between successive bases used in the Nucleotide Frequency method (NF method)

The hypothesis of no correlation between successive bases has been justified by the entropy approach (Koch & Lehmann, 1997, p. 173). We briefly recall the elementary principles of the entropy.

2.2.1 Method

Let X be a discrete random variable taking the value \( a_i \in \{A,C,G,T\} \) with the probability \( P(a_i) = Pr(X = a_i) \). The entropy \( H(X) \) of the discrete random variable \( X \) can be defined, in a simple approach, by the measure of the average information quantity associated with this variable \( X \), i.e.

\[
H(X) = -\sum_{i=1}^{4} P(a_i) \log_2 P(a_i)
\]

The entropy \( H(X) \) defined for the words of length 1 (nucleotides) is extended for words \( w_i = a_{i_1}...a_{i_n} \), \( i \in \{1,...,4^n\} \), of a given length \( n \) as follows

\[
H_n = -\sum_{i=1}^{4^n} P(w_i) \log_2 P(w_i)
\]

where \( P(w_i) \) is the occurrence probability of the word \( w_i \). Note \( H_n = H(X) \).

As the protein genes are read in the reading frame, the entropy \( H_n \) defined for the words of length \( n \) is extended to the entropy \( H_{n,f} \), \( f \in \{0,1,2\} \), computed from the occurrence probabilities \( P_f(w_i) \) of the word \( w_i = a_{i_1}...a_{i_n} \) in the frame \( f \), as follows

\[
H_{n,f} = -\sum_{i=1}^{4^n} P_f(w_i) \log_2 P_f(w_i)
\]

Notes:
(i) For the word \( w_i \) of length 1 (\( n=1 \)), there is the obvious relation

\[
P_f(w_i) = p_{f+1}(w_i) \quad (11)
\]

where \( p_{f+1}(w_i) \) is the probability of a base at the position \( f+1 \in \{1,2,3\} \) in the NF method.

(ii) \( H_{3,0} \) can be considered as a classical entropy \( H(Y) \) for the discrete random variable \( Y \) taking the 64 values in \{AAA,...,TTT\} in reading frame.

When the probabilities follow a random discrete uniform law, i.e. all the probabilities are equal, then the maxima of the entropy functions \( H_n \) and \( H_{n,f} \) are reached and are equal to

\[
\sum_{i=1}^{4^n} \frac{1}{4^n} \log_2 4^n = \log_2 4^n = 2n
\]

bits (Cover & Thomas, 1991).

Classically, an entropy function is expressed in bits per nucleotide with a maximal value equal to 2 corresponding to an uniform random distribution (Loewenstern & Yianilos,1999). Then, the introduced functions are normalised as follows:
The two statistical methods presented in Section 1, the Trinucleotide Frequency (TF) method (Arquès & Michel, 1996) and the Nucleotide Frequency (NF) method (Koch & Lehmann, 1997), allow to construct circular codes from data observed in the coding genes. The circular codes constructed by both methods, are sets of trinucleotides in frame 0. The construction of these different codes are based on the occurrence probabilities of the triplets in frame 0.

The TF method directly uses these probabilities.

In contrast, the NF method assumes the independence between successive bases for using the occurrence probabilities of the bases at the different positions in a trinucleotide (triplet) observed in frame 0. The computation of the entropies associated with the 2 models of probabilities will measure the real influence of the hypothesis of non-correlation between successive bases.

The NF method is based on the occurrence probability \( p_i(\theta) \) of a given base \( \theta \in \{A,C,G,T\} \) at position \( i \in \{1,2,3\} \) in a trinucleotide (triplet) observed in frame 0. By assuming the non-correlation between successive bases, the occurrence probability \( P_0(\alpha\beta\gamma) \) of the trinucleotide \( \alpha\beta\gamma \) in frame 0, is then deduced by the product of individual probabilities which is equal by using the relation (11) to

\[
P_0(\alpha\beta\gamma) = P_0(\alpha)P_2(\beta)P_3(\gamma) = p_i(\alpha)p_j(\beta)p_k(\gamma)
\]

Then, the entropy \( H_{nf} \) associated to these probabilities is

\[
H_{nf} = - \sum_{\alpha,\beta,\gamma=\{A,C,G,T\}} p_i(\alpha)p_j(\beta)p_k(\gamma) \log_2 (p_i(\alpha)p_j(\beta)p_k(\gamma))
\]

By assuming the non-correlation between successive bases and by using the relation (11), basic results lead to the entropy \( H_{nf} \) equal to (Cover & Thomas, 1991)

\[
H_{nf} = - \sum_{\alpha,\beta,\gamma=\{A,C,G,T\}} p_i(\alpha)p_j(\beta)p_k(\gamma) \log_2 (p_i(\alpha)p_j(\beta)p_k(\gamma))
= - \sum_{i=0}^2 \sum_{\alpha=\{A,C,G,T\}} p_i(\alpha) \log_2 p_i(\alpha)
= - \sum_{i=0}^2 \sum_{\alpha=\{A,C,G,T\}} P_i(w_i) \log_2 P_i(w_i)
= \sum_{i=0}^2 H_{nf}
\]

The TF method is based on the observed occurrence probabilities of the trinucleotides in the frame 0. Therefore, its entropy \( H_{tf} \) is equal to

\[
H_{tf} = - \sum_{\alpha,\beta,\gamma=\{A,C,G,T\}} P_0(\alpha\beta\gamma) \log_2 P_0(\alpha\beta\gamma) = H_{3,0}
\]

In order to express the entropies \( H_3 \), \( H_{nf} \) and \( H_{tf} \) in bits per nucleotide, the functions are normalized according to (12) and (13)

\[
\tilde{H}_3 = H_3/3
\]

\[
\tilde{H}_{nf} = H_{nf}/n\quad (12)
\]

\[
\tilde{H}_{tf} = H_{tf}/n\quad (13)
\]
Remark:
With gene populations containing several millions of nucleotides (e.g. Arquès & Michel, 1996; Koch & Lehmann, 1997), the computed probabilities are stable (law of large numbers). Therefore, the values obtained here from such probabilities lead to a precise approximation of the entropy functions.

\[ H_{\text{NF}} = \frac{H_{\text{NF}}}{3} \]
\[ H_{\text{TF}} = \frac{H_{\text{TF}}}{3} \]

2.2.2 Results

The values of these entropies in the prokaryotic protein genes are presented in the Table 4. The values of \( H_1 \) (resp. \( H_3 \)) are associated with the nucleotides (resp. the trinucleotides) without considering the existence of the reading frame in the prokaryotic protein genes. As expected, these values are closed to 2 representing the random situation. The value \( H_3 \) (1.984 bit per nucleotide) is slightly less than the value of \( H_1 \) (1.998 bit per nucleotide), showing that the basic element of information in the protein genes, is the trinucleotide and not the nucleotide.

The value of \( H_{\text{TF}} \) (1.918 bit per nucleotide) associated with the TF method, is significantly lower than the value of \( H_{\text{NF}} \) (1.965 bit per nucleotide) associated with the NF method. The \( H_{\text{TF}} \) value can be compared with the classical estimate of entropy of coding genes which is about 1.92 (Loewenstern & Yianilos, 1999). This value of 1.92 can be improved by considering particular sequences or by using specific algorithms as shown in Table 4 of Loewenstern & Yianilos (1999) for a non-redundant collection of 490 human genes. The improvement of the estimate of the entropy is not the aim of this paper. But, the fact that the value of \( H_{\text{TF}} \) corresponds to the classical estimate, implies that the probability model used in the TF method can be considered as a correct representation of the structure of the coding genes.

In contrast, the value of \( H_{\text{NF}} \) differs significantly from the classical estimate. The hypothesis of independence between successive bases has then a strong effect on the values of the entropies. Therefore, the probability model used in the NF method does not reveal neither the internal structure of the coding genes nor the occurrence probabilities of the triplets in frame 0.

3 DISCUSSION

Koch & Lehmann (1997) have proposed a probabilistic model for constructing the circular code observed in the protein genes. Their method (called here Nucleotide Frequency (NF) method) is based on the nucleotide frequencies with a hypothesis of absence of correlation between successive bases on a DNA strand for deducing a circular code from the product of the 3 occurrence probabilities of nucleotides in the positions of trinucleotide read in frame 0. It allows a simple construction of some particular circular codes but reveals several limits for constructing the circular code associated with protein genes:

(i) Several self-complementary circular codes, but not an unique one, are generated by the NF method from the observed probabilities (Section 2.1.1).

(ii) The self-complementary circular code \( X_0 \) observed in the in the protein genes of both eukaryotes and prokaryotes cannot be generated by the NF method (Section 2.1.2).
(iii) 88 among 216 self-complementary circular codes can be generated by the NF method (Section 2.1.3). They are listed in Table 3.

(iv) The hypothesis used in the NF method of no correlation between successive bases in the protein genes, is not verified (Section 2.2.2). Indeed, this hypothesis has been justified by computing the entropy with occurrence probabilities of words of length 1 to 6 (Koch & Lehmann, 1997). However, any probability model can produced a value of entropy. The choice of the function for revealing the genetic information in the sense of the information theory defined by Shannon (1949), is very important as the value of the entropy strongly varies among the functions used. Several examples of different functions estimating the value of the entropy are presented in Chatzidimitriou-Dreismamm et al (1996), Lio et al (1996), Loewenstern & Yianilos (1999), etc. In order to evaluate the hypothesis of non-correlation between successive bases, 2 estimates of the entropy are computed here. The first estimate associated with the TF method, is based on the 64 occurrence probabilities of triplets in frame 0. The entropy value \( H_{TF} \) associated with these probabilities, is equal to 1.918 bit per nucleotide and is similar to the classical estimate (1.92) of the entropy of coding genes (Loewenstern & Yianilos,1999). The second estimate associated with the NF method, is based on the 12 occurrence probabilities of nucleotides in the 3 triplet sites. These nucleotide probabilities with the hypothesis of non-correlation between successive bases, allow to deduce the occurrence probabilities of triplets in frame 0 more simply (with 12 values compared to 64 ones, but with a probability hypothesis). However, its entropy value \( H_{NF} \) is equal to 1.965 bit per nucleotide and significantly differs from \( H_{TF} \). Therefore, the hypothesis of non-correlation between successive bases is not verified.

4 CONCLUSION

The method introduced by Koch & Lehmann (1997) is a new approach for constructing circular codes. This NF method constructs in a simple way a sub-set of circular codes which is included in the set of circular codes generated by the flower automaton method. The NF method has an obvious interest in the field of the theory of codes. In this paper, some new results are presented in this respect, in particular the number of codes generated by this NF method and some patterns of code words excluded by the NF method.

However, the main purpose of the NF method was to explain the circular code \( X_0 \) identified in the protein genes of both eukaryotes and prokaryotes (Arquès & Michel, 1996). Several results were presented here concerning the relations between the NF method and the code \( X_0 \). The NF method does not generate a unique self-complementary circular code. Furthermore, it cannot generate the code \( X_0 \). Finally, the hypothesis of non-correlation between successive bases at the basis of the NF method, is rejected as the different computations of the entropy clearly show that the probabilities used by the NF method does not respect the internal structure of the coding genes. In conclusion, the NF method is not an appropriate model for explaining the circular code \( X_0 \).
REFERENCES


Table 1a List per frame and in lexicographical order of the trinucleotides of the complementary circular code identified in protein coding genes of eukaryotes and prokaryotes (Arquès & Michel, 1996). Three subsets of trinucleotides can be identified: $T_0 = X_0 \cup \{\text{AAA}, \text{TTT}\}$ in frame 0, $T_1 = X_1 \cup \{\text{CCC}\}$ in frame 1 and $T_2 = X_2 \cup \{\text{GGG}\}$ in frame 2. The 3 sets $X_0$, $X_1$ and $X_2$ of 20 trinucleotides are maximal circular codes.

Table 1b Circularity property with the 3 circular codes $X_0$, $X_1$ and $X_2$ of 20 trinucleotides identified in protein coding genes of eukaryotes and prokaryotes (Table 1a).

Table 1c Complementarity property with the 3 circular codes $X_0$, $X_1$ and $X_2$ of 20 trinucleotides identified in protein coding genes of eukaryotes and prokaryotes (Table 1a). This property is also verified with $T_0$ (AAA and TTT) and, $T_1$ and $T_2$ (CCC and GGG).
### Table 2a
Nucleotide frequencies $p_i(0)$ at position $i \in \{1, 2, 3\}$ of the reading frame for the prokaryotes (Koch & Lehmann, 1997, Table 1).

<table>
<thead>
<tr>
<th>Base</th>
<th>$p_1(0)$</th>
<th>$p_2(0)$</th>
<th>$p_3(0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.276</td>
<td>0.315</td>
<td>0.222</td>
</tr>
<tr>
<td>T</td>
<td>0.166</td>
<td>0.285</td>
<td>0.268</td>
</tr>
<tr>
<td>C</td>
<td>0.204</td>
<td>0.228</td>
<td>0.268</td>
</tr>
<tr>
<td>G</td>
<td>0.354</td>
<td>0.172</td>
<td>0.242</td>
</tr>
</tbody>
</table>

### Table 2b
Three self-complementary circular codes generated by the Nucleotide Frequency method (NF method) with the frequencies of Table 2a modified according to the relation (3): $p_1(A) = p_3(T)$, $p_1(C) = p_3(G)$, $p_1(G) = p_3(C)$, $p_1(T) = p_3(A)$, $p_2(A) = p_2(T)$ and $p_2(C) = p_2(G)$.

<table>
<thead>
<tr>
<th>Circular codes</th>
<th>$p_1(A)$</th>
<th>$p_1(C)$</th>
<th>$p_1(G)$</th>
<th>$p_1(T)$</th>
<th>$p_2(A)$</th>
<th>$p_2(C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AACAATACCAGCATTTCATCCGAAGACGAGGATGCCGCTGCCGCTGTAGTCTGTTACTTC</td>
<td>0.276</td>
<td>0.204</td>
<td>0.354</td>
<td>0.166</td>
<td>0.285</td>
<td>0.215</td>
</tr>
<tr>
<td>AACAAGATATCATTCACCAGCTCTGCTTAGCAGGATGCCCCGCTGCTGCTGTTTAC</td>
<td>0.268</td>
<td>0.242</td>
<td>0.268</td>
<td>0.222</td>
<td>0.285</td>
<td>0.215</td>
</tr>
<tr>
<td>AACAAGATATCATCATTACCCTCTTGGAGGATGCTGGGGCTCCTGCTGTGTGTATTAC</td>
<td>0.272</td>
<td>0.223</td>
<td>0.311</td>
<td>0.194</td>
<td>0.300</td>
<td>0.200</td>
</tr>
<tr>
<td>Nucleotide (n=1)</td>
<td>Entropy in the frame 0 modulo 3</td>
<td>Classical entropy $H_n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------------------</td>
<td>------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\tilde{H}_{NF} = 1.965$</td>
<td>$H_1 = 1.998$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trinucleotide (n=3)</td>
<td>$\tilde{H}_{TF} = 1.918$</td>
<td>$\tilde{H}_3 = 1.984$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3** List of the 88 self-complementary circular codes generated by the Nucleotide Frequency method (NF method) according to the 6 probabilities $p_x(A) = p_x(T)$, $p_x(G) = p_x(C)$, $p_x(T) = p_x(A)$, $p_x(C) = p_x(G)$.

**Table 4** Computation of different types of entropies (bit per nucleotide) from the occurrence frequencies of the 64 trinucleotides in the frame 0 modulo 3 and in the 3 frames (average frame) of prokaryotic protein coding genes (13686 sequences, 4708758 trinucleotides; data from Arquès & Michel, 1996, p. 49).