J. theor. Biol. (2001) **213,** 159–170
doi:10.1006/jtbi.2001.2416, available online at http://www.idealibrary.com on IDEAL®

Analysis of a Circular Code Model

JÉRÔME LACAN* AND CHRISTIAN J. MICHEL†‡

*Laboratoire d'Informatique de Franche-Comté, Université de Franche-Comté, IUT de Belfort-Montbéliard, 4 Place Tharradin - BP 71427, 25211 Montbéliard Cedex, France and †Equipe de Bioinformatique Théorique, LSIIT (UA CNRS 7005), Université Louis Pasteur Strasbourg, Pôle API, Boulevard Sébastien, Brant, 67400 Illkirch, France

(*Received on* 20 *March* 2001, *Accepted in revised form on* 27 *July* 2001)

A circular code has been identified in the protein (coding) genes of both eukaryotes and prokaryotes by using a statistical method called trinucleotide frequency (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 (NF) method, reveals several limits for constructing the circular code observed with protein genes.

1. Introduction

This section is divided into two 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 the 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:

… CAG,GCC,TTC,AAT,ACC,ACC,CAG,GAA, GAG,GTA,ATT,ACC,AAT,GTA,AAC,TAC, TTC,ACC,ATC …

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:

… CAGGCCTTCAATACCACCCAGG-AAGAGGTAATTACCAATGTAAACTACT-TCACCATC …

‡Author to whom correspondence should be addressed. E-mails: michel@dpt-info.u-strasbg.fr; jerome.lacan@pu-pm.univ-fcomte.fr

0022–5193/01/220159 + 12 $35.00/0

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

… C,AGG,CCT,TCA,ATA,CCA,CCC,AGG, AAG,AGG,TAA,TTA,CCA,ATG,TAA,ACT, ACT,TCA,CCA,TC …

… CA,GGC,CTT,CAA,TAC,CAC,CCA,GGA, AGA,GGT,AAT,TAC,CAA,TGT,AAA,CTA, CTT,CAC,CAT,C …

… CAG,GCC,TTC,AAT,ACC,ACC,CAG,GAA, GAG,GTA,ATT,ACC,AAT,GTA,AAC,TAC, TTC,ACC,ATC …

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

… CAG,GCC,TTC,AAT,ACC,ACC,CAG,GAA, GAG,GTA,ATT,ACC,AAT,GTA,AAC,TAC, TTC,ACC,ATC …

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:

… CAGGCCTTCAATACCACCCAGGAAG ~~AGG~~,TAATTACCAATGTAAACTACTTCACCATC …

… CAGGCCTTCAATACCACCCAGGAAG A,GGT,AAT,TAC,~~CAA~~,TGTAAACTACTTCACCATC …

… CAGGCCTTCAATACCACCCAGGAAG AG,GTA,ATT,ACC,AAT,GTA,AAC,TAC,TTC,ACC,ATC, …

The first decomposition proposed is rejected immediately as the first trinucleotide 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 two 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 three possible decompositions: … AAA,AAA,AAA, … , … A, AAA,AAA,AA … 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 two 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 three trinucleotides so that, in each class, the three 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.

In the late 1950s, no set of 20 trinucleotides leading to a circular code has been found. Furthermore, the two 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 giving up of 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 two 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 two methods.

(i) By computing the occurrence frequencies of the 64 trinucleotides AAA,...,TTT in the three 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 12 288 (3×64^2) 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,GTT,TAC,TTC$\}$ of 20 trinucleotides in frame 0 of protein genes (reading frame). Furthermore, the two 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 two methods, are also maximal circular codes [Table 1(a)]. These three circular codes have several important properties.

(i) *Circularity*: X_0 generates X_1 by one circular permutation and X_2 by another circular permutation (one and two circular permutations of each trinucleotide of X_0 lead to the trinucleotides of X_1 and X_2, respectively) [Table 1(b)].

(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 1(c)]. 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 1(c)].

(iii) *Rarity*: the occurrence probability of X_0 is equal to 6×10^{-8}. As there are 20 classes of three trinucleotides (see above), the number of potential circular codes is $3^{20} = 3\,486\,784\,401$. The computed number of complementary circular codes with two 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 three 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:

  ... GAA,GAG,GTA,GTA,ACC,AAT, GTA,CTC,TAC,TTC,ACC,ATC...
  then, the trinucleotides in frame 1:
  ... G,AAG,AGG,TAG,TAA,CCA,ATG, TAC,TCT,ACT,TCA,CCA,TC...
  and the trinucleotides in frame 2:
  ... GA,AGA,GGT,AGT,AAC,CAA,TGT, ACT, CTA, CTT,CAC,CAT,C...

  mainly belong to X_1 and X_2, respectively. A few trinucleotides are misplaced in the shifted frames. With this example, in frame 1, nine trinucleotides belong to X_1, one trinucleotide (TAC) to X_0 and one trinucleotide (TAA) to

TABLE 1

(a) *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*: $T_0 = X_0 \cup \{\text{AAA,TTT}\}$ *in frame* 0, $T_1 = X_1 \cup \{\text{CCC}\}$ *in frame* 1 *and* $T_2 \cup \{\text{GGG}\}$ *in frame* 2. *The three sets* X_0, X_1 *and* X_2 *of* 20 *trinucleotides are maximal circular codes.* (b) *Circularity property with the three circular codes* X_0, X_1 *and* X_2 *of* 20 *trinucleotides identified in protein coding genes of eukaryotes and prokaryotes* [*Table* 1(*a*)]. (c) *Complementarity property with the three circular codes* X_0, X_1 *and* X_2 *of* 20 *trinucleotides identified in protein coding genes of eukaryotes and prokaryotes* [*Table* 1(*a*)]. *This property is also verified with* T_0 (AAA *and* TTT) *and*, T_1 *and* T_2 (CCC *and* GGG)

(a)

T_0:	AAA	AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC	TTT
T_1:	AAG	ACA	ACG	ACT	AGC	AGG	ATA	ATG	CCA	CCC	CCG	GCG	GTG	TAG	TCA	TCC	TCG	TCT	TGC	TTA	TTG	
T_2:	AGA	AGT	CAA	CAC	CAT	CCT	CGA	CGC	CGG	CGT	CTA	CTT	GCA	GCT	GGA	GGG	TAA	TAT	TGA	TGG	TGT	

(b)

X_0:	AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC
X_1:	ACA	ATA	CCA	TCA	TTA	AGC	TCC	TGC	AAG	ACG	AGG	ATG	CCG	GCG	GTG	TAG	TCG	TTG	ACT	TCT
X_2:	CAA	TAA	CAC	CAT	TAT	GCA	CCT	GCT	AGA	CGA	GGA	TGA	CGC	CGG	TGG	AGT	CGT	TGT	CTA	CTT

(c)

T_0:	AAA	AAC	AAT	ACC	ATC	CAG	CTC	GAA	GAC	GCC	GTA										
T_0:	TTT	GTT	ATT	GGT	GAT	CTG	GAG	TTC	GTC	GGC	TAC										
T_1:	AAG	ACA	ACG	ACT	AGC	AGG	ATA	ATG	CCA	CCC	CCG	GCG	GTG	TAG	TCA	TCC	TCG	TCT	TGC	TTA	TTG
T_2:	CTT	TGT	CGT	AGT	GCT	CCT	TAT	CAT	TGG	GGG	CGG	CGC	CAC	CTA	TGA	GGA	CGA	AGA	GCA	TAA	CAA

X_2. In frame 2, eight trinucleotides belong to X_2, two trinucleotides (GGT, AAC) to X_0 and one 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 the frequency equality of X_0 in frames 1 and 2 and on the other, 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 four types of nucleotides occur in the three trinucleotide sites with the three circular codes X_0, X_1 and X_2 [Table 1(a)].

(v) *Evolutionary*: an evolutionary analytical model at three 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 three trinucleotide sites, respectively, retrieves the frequencies of X_0, X_1 and X_2 observed in the three 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 two-letter genetic alphabets, the genetic code and the amino acid frequencies in proteins, are given in Arquès & Michel (1996, 1997a). 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).

TABLE 2

(a) *Nucleotide frequencies* $p_i(\theta)$ *at position* $i \in \{1,2,3\}$ *of the reading frame for the prokaryotes* (*Koch & Lehmann*, 1997, *Table* 1). (b) *Three self-complementary circular codes generated by the NF method with the frequencies of Table* 2(*a*) *modified according to relation* (3): $p_1(\mathrm{A}) = p_3(\mathrm{T})$, $p_1(\mathrm{C}) = p_3(\mathrm{G})$, $p_1(\mathrm{G}) = p_3(\mathrm{C})$, $p_1(\mathrm{T}) = p_3(\mathrm{A})$, $p_2(\mathrm{A}) = p_2(\mathrm{T})$, *and* $p_2(\mathrm{C}) = p_2(\mathrm{G})$

(a)

Base θ	$p_1(\theta)$	$p_2(\theta)$	$p_3(\theta)$
A	0.276	0.315	0.222
T	0.166	0.285	0.268
C	0.204	0.228	0.268
G	0.354	0.172	0.242

(b)

Circular codes	p_1(A)	p_1(C)	p_1(G)	p_1(T)	p_2(A)	p_2(C)
AACAATACCAGCATCATTCTCGAAGACGAGGATGCCGCTGGCGGTGTAGTCGTTTACTTC	0.276	0.204	0.354	0.166	0.285	0.215
AACAAGAATATCATTCACCAGCTCCTGCTTGACGAGGATGCCGGCGTAGTCGTGGTTTAC	0.268	0.242	0.268	0.222	0.285	0.215
AACAAGAATAGCATCATTCACCTCCTTGACGAGGATGCCGCTGGCGTAGTCGTGGTTTAC	0.272	0.223	0.311	0.194	0.300	0.200

1.2. THE NF METHOD

Koch & Lehmann (1997, p. 171) have recently suggested that the self-complementary circular code X_0 observed in protein genes could be explained by a method for generating circular codes from nucleotide frequencies. This method called here as the NF method, is briefly recalled by retaining the same notations.

Let $p_i(\theta)$ be the occurrence probability of a given base $\theta \in \{\mathrm{A,C,G,T}\}$ 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 $p_1(\alpha)p_2(\beta)p_3(\gamma)$ (independent probabilities). The belonging of the triplet $\alpha\beta\gamma$ to a preferential set Y_0 of triplets in frame 0 is then equivalent to the following two probability inequalities:

$$p_1(\alpha)p_2(\beta)p_3(\gamma) > p_1(\gamma)p_2(\alpha)p_3(\beta) \qquad (1)$$

and

$$p_1(\alpha)p_2(\beta)p_3(\gamma) > p_1(\beta)p_2(\gamma)p_3(\alpha). \qquad (2)$$

Similar probability inequalities imply that the triplet $\beta\gamma\alpha$ (resp. $\gamma\alpha\beta$) belongs to the preferential set Y_1 (resp. Y_2) 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

$$p_1(\theta) = p_3(\mathrm{C}(\theta)) \text{ and } p_2(\theta) = p_2(\mathrm{C}(\theta)), \qquad (3)$$

where $\mathrm{C}(\theta)$ denote the complementary base of θ, then the circular code Y_0 is necessarily self-complementary and the two permutated circular codes Y_1 and Y_2 are complementary (called C^3 codes in Arquès & Michel, 1996).

Table 1 in Koch & Lehmann (1997) gives the 12 nucleotide observed frequencies $p_i(\theta)$ of a base $\theta \in \{\mathrm{A,C,G,T}\}$ at position $i \in \{1,2,3\}$ of the reading frame for the prokaryotes. These data have been obtained from the 44-th release of the prokaryotic EMBL database. This Table 1 is recalled in this paper with the Table 2(a). These 12 probabilities with the NF method lead to a new circular code $Y_0 = \{$AAT, AAC, ATT, ATC, ACT, CAC, CTT, CTC, GAA, GAT, GAC, GAG, GTA, GTT, GTC, GTG, GCA, GCT, GCC, GCG$\}$. This code Y_0 contains 13 trinucleotides of the code X_0 [Table 1(a)].

2. Method and Results

2.1. THE NF METHOD CANNOT GENERATE THE CIRCULAR CODE X_0

2.1.1. *The NF Method Does not Generate a 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 relation (3), then the circular code computed by the NF method is a self-complementary code. However, relation (3) which contains six 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 2(a) do not precisely verify 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 2(a) are slightly modified for verifying relation (3). Three examples of such self-complementary circular codes are presented in Table 2(b). The first circular code is obtained with observed frequencies from the first and second columns of Table 2(a): $p_1(A) = p_3(T) = 0.276$, $p_1(C) = p_3(G) = 0.204$, $p_1(G) = p_3(C) = 0.354$, $p_1(T) = p_3(A) = 0.166$, $p_2(A) = p_2(T) = 0.285$, 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 2(a): $p_1(A) = p_3(T) = 0.268$, $p_1(C) = p_3(G) = 0.242$, $p_1(G) = p_3(C) = 0.268$, $p_1(T) = p_3(A) = 0.222$, $p_2(A) = p_2(T) = 0.285$, and $p_2(C) = p_2(G) = (1-2 \times 0.285)/2$. The third circular code is obtained with average frequencies from Table 2(a): $p_1(A) = p_3(T) = (0.276 + 0.268)/2 = 0.272$, $p_1(C) = p_3(G) = (0.204 + 0.242)/2 = 0.223$, $p_1(G) = p_3(C) = (0.354 + 0.268)/2 = 0.311$, $p_1(T) = p_3(A) = (0.166 + 0.222)/2 = 0.194$, $p_2(A) = p_2(T) = (0.315 + 0.285)/2 = 0.3$, 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 a unique self-complementary circular code. Furthermore, in the next section we shall prove 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 three 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 three triplets considered. As the circular code X_0 contains such three triplets (ATC, GGT, CAG), then X_0 cannot be generated by the NF method.

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

$$p_1(\alpha)p_2(\beta)p_3(\gamma) > p_1(\gamma)p_2(\alpha)p_3(\beta). \tag{4}$$

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

$$p_1(\delta)p_2(\delta)p_3(\beta) > p_1(\delta)p_2(\beta)p_3(\delta). \tag{5}$$

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

$$p_2(\delta)p_3(\beta) > p_2(\beta)p_3(\delta). \tag{6}$$

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

$$p_1(\gamma)p_2(\alpha)p_3(\delta) > p_1(\alpha)p_2(\delta)p_3(\gamma). \tag{7}$$

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

$$p_1(\alpha)p_2(\beta)p_3(\gamma) > p_1(\gamma)p_2(\alpha)p_3(\delta) \times p_3(\beta)/p_3(\delta). \tag{8}$$

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

$$p_1(\alpha)p_2(\beta)p_3(\gamma) > p_1(\alpha)p_2(\delta)p_3(\gamma) \times p_3(\beta)/p_3(\delta). \tag{9}$$

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

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

i.e.

$$p_2(\beta)p_3(\delta) > p_2(\delta)p_3(\beta). \tag{10}$$

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

This proof can be applied to the circular code X_0 containing the three 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 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 two 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 four bases at the three positions in the range [0,1] according to 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 two circular permuted patterns $P_1 = \{\beta\gamma\alpha, \delta\beta\delta, \alpha\delta\gamma\}$ 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 (Section 2.1.2). Similarly the proof obtained 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 three 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 NF 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 relation (3), as the proof obtained does not make any hypothesis on the probabilities.

2.2. REMARKS ON THE HYPOTHESIS OF NO CORRELATION BETWEEN SUCCESSIVE BASES USED IN THE 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 $a_i \in \{\text{A,C,G,T}\}$ 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.

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

TABLE 3

List of the 88 *self-complementary circular codes generated by the NF method according to the six probabilities* $p_1(A) = p_3(T)$, $p_1(C) = p_3(G)$, $p_1(G) = p_3(C)$, $p_1(T) = p_3(A)$, $p_2(A) = p_2(T)$ *and* $p_2(C) = p_2(G)$

Circular codes	$p_1(A)$	$p_1(C)$	$p_1(G)$	$p_1(T)$	$p_2(A)$	$p_2(C)$
ACA AGA CCA CGA GCA GCC GGA GGC GTA TAA TAC TCA TCC TCG TCT TGA TGC TGG TGT TTA	0.06	0.06	0.12	0.76	0.06	0.44
ACA CCA CGA GAA GCA GCC GGA GGC GTA TAA TAC TCA TCC TCG TGA TGC TGG TGT TTA TTC	0.06	0.06	0.12	0.76	0.18	0.32
CAA CCA CGA GAA GCA GCC GGA GGC GTA TAA TAC TCA TCC TCG TGA TGC TGG TTA TTC TTG	0.06	0.06	0.12	0.76	0.30	0.20
CAA CCA GAA GAC GCA GCC GGA GGC GTA GTC TAA TAC TCA TCC TGA TGC TGG TTA TTC TTG	0.06	0.06	0.12	0.76	0.42	0.08
CAA CAC CTC GAA GAC GAG GCA GCC GGC GTA GTC GTG TAA TAC TCA TGA TGC TTA TTC TTG	0.06	0.06	0.12	0.76	0.48	0.02
CAA CAC GAA GAC GCA GCC GGA GGC GTA GTC GTG TAA TAC TCA TCC TGA TGC TTA TTC TTG	0.06	0.06	0.18	0.70	0.42	0.08
ACA CCA GAA GAC GCA GCC GGA GGC GTA GTC TAA TAC TCA TCC TGA TGC TGG TGT TTA TTC	0.06	0.06	0.24	0.64	0.24	0.26
ATC CAA CAC CTC GAA GAC GAG GAT GCA GCC GGC GTA GTC GTG TAA TAC TGC TTA TTC TTG	0.06	0.06	0.30	0.58	0.48	0.02
ACA ACC GAA GAC GCA GCC GGA GGC GGT GTA GTC TAA TAC TCA TCC TGA TGC TGT TTA TTC	0.06	0.06	0.48	0.40	0.06	0.44
AAC ACC GAA GAC GCA GCC GGA GGC GGT GTA GTC GTT TAA TAC TCA TCC TGA TGC TTA TTC	0.06	0.06	0.48	0.40	0.24	0.26
AAC CAC GAA GAC GCA GCC GGA GGC GTA GTC GTG GTT TAA TAC TCA TCC TGA TGC TTA TTC	0.06	0.06	0.48	0.40	0.30	0.20
AAC ATC CAC CTC GAA GAC GAG GAT GCA GCC GGC GTA GTC GTG GTT TAA TAC TGC TTA TTC	0.06	0.06	0.48	0.40	0.48	0.02
AAC ATC CAC GAA GAC GAT GCA GCC GGA GGC GTA GTC GTG GTT TAA TAC TCC TGC TTA TTC	0.06	0.06	0.54	0.34	0.42	0.08
AAC ACC ATC GAA GAC GAT GCA GCC GGA GGC GGT GTA GTC GTT TAA TAC TCC TGC TTA TTC	0.06	0.06	0.72	0.16	0.24	0.26
AAC ATC CAC CAG CTC CTG GAA GAC GAG GAT GCC GGC GTA GTC GTG GTT TAA TAC TTA TTC	0.06	0.06	0.78	0.10	0.48	0.02
AAC AAT ACC AGC ATC ATT GAA GAC GAT GCC GCT GGA GGC GGT GTA GTC GTT TAC TCC TTC	0.06	0.06	0.84	0.04	0.06	0.44
AAC AAT ACC AGC ATC ATT CTC GAA GAC GAG GAT GCC GCT GGC GGT GTA GTC GTT TAC TTC	0.06	0.06	0.84	0.04	0.24	0.26
AAC AAT AGC ATC ATT CAC CTC GAA GAC GAG GAT GCC GCT GGC GTA GTC GTG GTT TAC TTC	0.06	0.06	0.84	0.04	0.30	0.20
AAC AAT ATC ATT CAC CAG CTC CTG GAA GAC GAG GAT GCC GGC GTA GTC GTG GTT TAC TTC	0.06	0.06	0.84	0.04	0.48	0.02
ACA AGA CCA CCG CGA CGG CTA GCA GGA TAA TAG TCA TCC TCG TCT TGA TGC TGG TGT TTA	0.06	0.12	0.06	0.76	0.06	0.44
AGA CAA CCA CCG CGA CGG CTA GCA GGA TAA TAG TCA TCC TCG TCT TGA TGC TGG TTA TTG	0.06	0.12	0.06	0.76	0.18	0.32
CAA CCA CCG CGA CGG CTA GAA GCA GGA TAA TAG TCA TCC TCG TGA TGC TGG TTA TTC TTG	0.06	0.12	0.06	0.76	0.30	0.20
CAA CAG CCA CCG CGA CGG CTA CTG GAA GGA TAA TAG TCA TCC TCG TGA TGG TTA TTC TTG	0.06	0.12	0.06	0.76	0.42	0.08
CAA CAC CAG CCG CGA CGG CTA CTC CTG GAA GAG GTG TAA TAG TCA TCG TGA TTA TTC TTG	0.06	0.12	0.06	0.76	0.48	0.02
CAA CAC CAG CTC CTG GAA GAC GAG GCC GGC GTA GTC GTG TAA TAC TCA TGA TTA TTC TTG	0.06	0.12	0.18	0.64	0.48	0.02
ATC CAA CAC CAG CTC CTG GAA GAC GAG GAT GCC GGC GTA GTC GTG TAA TAC TTA TTC TTG	0.06	0.12	0.24	0.58	0.48	0.02
CAA CAG CCA CCG CGA CGG CTA CTC CTG GAA GAG TAA TAG TCA TCG TGA TGG TTA TTC TTG	0.06	0.18	0.06	0.70	0.42	0.08
CAA CAC CAG CCG CGG CTA CTC CTG GAA GAC GAG GTC GTG TAA TAG TCA TGA TTA TTC TTG	0.06	0.18	0.12	0.64	0.48	0.02
AGA CAA CAG CCA CCG CGA CGG CTA CTG GGA TAA TAG TCA TCC TCG TCT TGA TGG TTA TTG	0.06	0.24	0.06	0.64	0.24	0.26
ATG CAA CAC CAG CAT CCG CGG CTA CTC CTG GAA GAC GAG GTC GTG TAA TAG TTA TTC TTG	0.06	0.24	0.12	0.58	0.48	0.02
AAC AAT ACC ACT AGC AGT ATC ATT GAA GAC GAT GCC GCT GGA GGC GGT GTC GTT TCC TTC	0.06	0.24	0.66	0.04	0.06	0.44
ATG CAA CAC CAG CAT CCG CGA CGG CTA CTC CTG GAA GAG GTG TAA TAG TCG TTA TTC TTG	0.06	0.30	0.06	0.58	0.48	0.02
AAC AAT ACC ACT AGC AGT ATC ATT CTC GAA GAC GAG GAT GCC GCT GGC GGT GTC GTT TTC	0.06	0.30	0.60	0.04	0.06	0.44
AAC AAG AAT ACC ACT AGC AGT ATC ATT CTC CTT GAC GAG GAT GCC GCT GGC GGT GTC GTT	0.06	0.36	0.54	0.04	0.06	0.44
AAC AAG AAT ATC ATT CAC CAG CTC CTG CTT GAC GAG GAT GCC GGC GTA GTC GTG GTT TAC	0.06	0.36	0.54	0.04	0.12	0.38
AGA AGG CAA CAG CCA CCG CCT CGA CGG CTA CTG TAA TAG TCA TCG TCT TGA TGG TTA TTG	0.06	0.48	0.06	0.40	0.06	0.44
AAG AGG CAA CAG CCA CCG CCT CGA CGG CTA CTG CTT TAA TAG TCA TCG TGA TGG TTA TTG	0.06	0.48	0.06	0.40	0.24	0.26
AAG CAA CAG CCA CCG CGA CGG CTA CTC CTG CTT GAG TAA TAG TCA TCG TGA TGG TTA TTG	0.06	0.48	0.06	0.40	0.30	0.20
AAG ATG CAA CAC CAG CAT CCG CGA CGG CTA CTC CTG CTT GAG GTG TAA TAG TCG TTA TTG	0.06	0.48	0.06	0.40	0.48	0.02
AAC AAG AAT ACG ACT AGG AGT ATG ATT CAC CAG CAT CCG CCT CGG CGT CTG CTT GTG GTT	0.06	0.48	0.42	0.04	0.06	0.44
AAC AAG AAT ATG ATT CAC CAG CAT CCG CGG CTA CTC CTG CTT GAC GAG GTC GTG GTT TAG	0.06	0.48	0.42	0.04	0.12	0.38
AAG ATG CAA CAG CAT CCA CCG CGA CGG CTA CTC CTG CTT GAG TAA TAG TCG TGG TTA TTG	0.06	0.54	0.06	0.34	0.42	0.08
AAG ATG CAA CAC CAG CAT CCG CGG CTA CTC CTG CTT GAC GAG GTC GTG TAA TAG TTA TTG	0.06	0.54	0.30	0.10	0.24	0.26
AAG AAT ACG ACT AGG AGT ATG ATT CAA CAC CAG CAT CCG CCT CGG CGT CTG CTT GTG TTG	0.06	0.60	0.30	0.04	0.06	0.44
AAG AAT ACG ATG ATT CAA CAC CAG CAT CCG CGG CGT CTA CTC CTG CTT GAG GTG TAG TTG	0.06	0.60	0.30	0.04	0.12	0.38
AAG AAT ATG ATT CAA CAC CAG CAT CCG CGG CTA CTC CTG CTT GAC GAG GTC GTG TAG TTG	0.06	0.60	0.30	0.04	0.18	0.32
AAG AGG ATG CAA CAG CAT CCA CCG CCT CGA CGG CTA CTG CTT TAA TAG TCG TGG TTA TTG	0.06	0.66	0.18	0.10	0.12	0.38
AAG AAT ACG ACT AGG AGT ATG ATT CAA CAG CAT CCA CCG CCT CGG CGT CTG CTT TGG TTG	0.06	0.66	0.24	0.04	0.06	0.44
AAG AAT ACG AGG ATG ATT CAA CAG CAT CCA CCG CCT CGG CGT CTA CTG CTT TAG TGG TTG	0.06	0.72	0.18	0.04	0.06	0.44
AAG AAT ACG AGG ATG ATT CAA CAC CAG CAT CCG CCT CGG CGT CTA CTG CTT GTG TAG TTG	0.06	0.72	0.18	0.04	0.12	0.38
ACA ACT AGA AGT CCA CGA GCA GCC GGA GGC TAA TCA TCC TCG TCT TGA TGC TGG TGT TTA	0.12	0.06	0.12	0.70	0.06	0.44
ACA ACC AGA CGA GCA GCC GGA GGC GGT GTA TAA TAC TCA TCC TCG TCT TGA TGC TGT TTA	0.12	0.06	0.30	0.52	0.06	0.44
ACA ACC AGA GAC GCA GCC GGA GGC GGT GTA GTC TAA TAC TCA TCC TCT TGA TGC TGT TTA	0.12	0.06	0.36	0.46	0.12	0.38
ACA ACC ACT AGA AGT GAC GCA GCC GGA GGC GGT GTC TAA TCA TCC TCT TGA TGC TGT TTA	0.12	0.06	0.66	0.16	0.06	0.44
AAT ACA ACC ACT AGA AGC AGT ATC ATT GAC GAT GCC GCT GGA GGC GGT GTC TCC TCT TGT	0.12	0.06	0.72	0.10	0.06	0.44
AAC AAT ACC ACT AGA AGC AGT ATC ATT GAC GAT GCC GCT GGA GGC GGT GTC GTT TCC TCT	0.12	0.06	0.78	0.04	0.06	0.44
ACA ACT AGA AGT CCA CCG CGA CGG GCA GGA TAA TCA TCC TCG TCT TGA TGC TGG TGT TTA	0.12	0.12	0.06	0.70	0.06	0.44
ACA ACC ACT AGA AGT CGA GCA GCC GGA GGC GGT TAA TCA TCC TCG TCT TGA TGC TGT TTA	0.12	0.18	0.48	0.22	0.06	0.44
AAT ACA ACC ACT AGA AGC AGT ATT GAC GCC GCT GGA GGC GGT GTC TCA TCC TCT TGA TGT	0.12	0.18	0.60	0.10	0.06	0.44
AAC AAT ACC ACT AGA AGC AGG AGT ATC ATT CCT GAC GAT GCC GCT GGC GGT GTC GTT TCT	0.12	0.24	0.60	0.04	0.06	0.44
AAC AAG AAT ACT AGC AGT ATC ATT CAC CTC CTT GAC GAG GAT GCC GCT GGC GTC GTG GTT	0.12	0.24	0.60	0.04	0.18	0.32
AAC AAG AAT AGC ATC ATT CAC CTC CTT GAC GAG GAT GCC GCT GGC GTA GTC GTG GTT TAC	0.12	0.24	0.60	0.04	0.24	0.26
ACA AGA AGG CCA CCG CCT CGA CGG CTA GCA TAA TAG TCA TCG TCT TGA TGC TGG TGT TTA	0.12	0.30	0.06	0.52	0.06	0.44
AAT ACA ACC ACG ACT AGA AGC AGT ATT CGT GCC GCT GGA GGC GGT TCA TCC TCT TGA TGT	0.12	0.30	0.48	0.10	0.06	0.44
AAC AAG AAT ACC ACT AGC AGG AGT ATC ATT CCT CTT GAC GAT GCC GCT GGC GGT GTC GTT	0.12	0.30	0.54	0.04	0.06	0.44
ACA AGA AGG CAG CCA CCG CCT CGA CGG CTA CTG TAA TAG TCA TCG TCT TGA TGG TGT TTA	0.12	0.36	0.06	0.46	0.12	0.38
AAT ACA ACC ACG ACT AGA AGC AGG AGT ATT CCT CGT GCC GCT GGC GGT TCA TCT TGA TGT	0.12	0.36	0.42	0.10	0.06	0.44
AAC AAG AAT ACC ACG ACT AGC AGG AGT ATC ATT CCT CGT CTT GAT GCC GCT GGC GGT GTT	0.12	0.36	0.48	0.04	0.06	0.44
AAC AAG AAT ACT AGT ATC ATT CAC CAG CTC CTG CTT GAC GAG GAT GCC GGC GTC GTG GTT	0.12	0.36	0.48	0.04	0.18	0.32
ACA ACT AGA AGG AGT CCA CCG CCT CGA CGG GCA TAA TCA TCG TCT TGA TGC TGG TGT TTA	0.12	0.42	0.30	0.16	0.06	0.44
AAT ACA ACC ACG ACT AGA AGC AGG AGT ATT CCG CCT CGG CGT GCT GGT TCA TCT TGA TGT	0.12	0.42	0.36	0.10	0.06	0.44
AAT ACA ACG ACT AGA AGC AGG AGT ATT CCA CCG CCT CGG CGT GCT TCA TCT TGA TGG TGT	0.12	0.48	0.30	0.10	0.06	0.44
AAC AAG AAT ACC ACG ACT AGC AGG AGT ATG ATT CAT CCG CCT CGG CGT CTT GCT GGT GTT	0.12	0.48	0.36	0.04	0.06	0.44
AAC AAG AAT ACT AGT ATG ATT CAC CAG CAT CCG CGG CTC CTG CTT GAC GAG GTC GTG GTT	0.12	0.48	0.36	0.04	0.18	0.32
ACA ACT AGA AGG AGT CAG CCA CCG CCT CGA CGG CTG TAA TCA TCG TCT TGA TGG TGT TTA	0.12	0.54	0.18	0.16	0.06	0.44
AAT ACA ACG ACT AGA AGG AGT ATT CAG CCA CCG CCT CGG CGT CTG TCA TCT TGA TGG TGT	0.12	0.54	0.24	0.10	0.06	0.44
AAC AAG AAT ACC ACG ACT AGG AGT ATG ATT CAG CAT CCG CCT CGG CGT CTG CTT GGT GTT	0.12	0.54	0.30	0.04	0.06	0.44
AAC AAG AAT ACG ACT AGT ATG ATT CAC CAG CAT CCG CGG CGT CTC CTG CTT GAG GTG GTT	0.12	0.54	0.30	0.04	0.18	0.32
AAG AAT ACA ACC ACG ACT AGG AGT ATG ATT CAG CAT CCG CCT CGG CGT CTG CTT GGT TGT	0.12	0.60	0.24	0.04	0.06	0.44
AAC AAG AAT ACG ATG ATT CAC CAG CAT CCG CGG CGT CTA CTC CTG CTT GAG GTG GTT TAG	0.12	0.60	0.24	0.04	0.24	0.26
AAT ACA ACG ACT AGA AGG AGT ATG ATT CAG CAT CCA CCG CCT CGG CGT CTG TCT TGG TGT	0.12	0.66	0.12	0.10	0.06	0.44
AAG AAT ACA ACG ACT AGG AGT ATG ATT CAG CAT CCA CCG CCT CGG CGT CTG CTT TGG TGT	0.12	0.66	0.18	0.04	0.06	0.44
AAT ACA ACC ACG ACT AGA AGC AGT ATC ATT CGT GAT GCC GCT GGA GGC GGT TCC TCT TGT	0.18	0.24	0.48	0.10	0.06	0.44
AAC AAT ACC ACG ACT AGA AGC AGG AGT ATC ATT CCT CGT GAT GCC GCT GGC GGT GTT TCT	0.18	0.24	0.54	0.04	0.06	0.44
AAT ACA ACC ACG ACT AGA AGC AGG AGT ATC ATT CCT CGT GAT GCC GCT GGC GGT TCT TGT	0.18	0.30	0.42	0.10	0.06	0.44
AAT ACA ACC ACG ACT AGA AGC AGG AGT ATG ATT CAT CCG CCT CGG CGT GCT GGT TCT TGT	0.18	0.42	0.30	0.10	0.06	0.44
AAT ACA ACG ACT AGA AGC AGG AGT ATG ATT CAT CCA CCG CCT CGG CGT GCT TCT TGG TGT	0.18	0.48	0.24	0.10	0.06	0.44
AAG AAT ACA ACC ACG ACT AGC AGG AGT ATG ATT CAT CCG CCT CGG CGT CTT GCT GGT TGT	0.18	0.54	0.24	0.04	0.06	0.44

The entropy $H(X)$ defined for the words of length 1 (nucleotides) is extended for words $w_i = a_1 \ldots a_n$, $i \in \{1, \ldots, 4^n\}$, 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 $H_1 = H(X)$.

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 $w_i = a_1 \ldots a_n$ 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

$$P_f(w_i) = p_{f+1}(w_i), \tag{11}$$

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 $\{\text{AAA}, \ldots, \text{TTT}\}$ 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 attained and are equal to $\sum_1^{4^n} (1/4^n) \log_2 4^n = \log_2 4^n = 2n$ bits (Cover & Thomas, 1991).

Classically, an entropy function is expressed in bits per nucleotide with a maximal value equal to 2 corresponding to a uniform random distribution (Loewenstern & Yianilos, 1999). Then, the introduced functions are normalized as follows:

$$\tilde{H}_n = H_n/n, \tag{12}$$

$$\tilde{H}_{n,f} = H_{n,f}/n. \tag{13}$$

The two statistical methods presented in Section 1, the TF method (Arquès & Michel, 1996) and the 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 two 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 $\theta \in \{\text{A,C,G,T}\}$ 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 relation (11) to

$$P_0(\alpha\beta\gamma) = P_0(\alpha)P_1(\beta)P_2(\gamma) = p_1(\alpha)p_2(\beta)p_3(\gamma).$$

Then, the entropy H_{NF} associated with these probabilities is

$$H_{NF} = -\sum_{\alpha,\beta,\gamma \in \{\text{A,C,G,T}\}} p_1(\alpha)p_2(\beta)p_3(\gamma) \times \log_2 (p_1(\alpha)p_2(\beta)p_3(\gamma)).$$

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

$$H_{NF} = -\sum_{\alpha,\beta,\gamma \in \{\text{A,C,G,T}\}} p_1(\alpha)p_2(\beta)p_3(\gamma) \times \log_2 (p_1(\alpha)p_2(\beta)p_3(\gamma))$$

$$= -\sum_{i=1}^{3} \sum_{\theta \in \{\text{A,C,G,T}\}} p_i(\theta) \log_2 p_i(\theta)$$

$$= -\sum_{f=0}^{2}\sum_{j=1}^{4} P_f(w_j)\log_2 P_f(w_j)$$

$$= \sum_{f=0}^{2} H_{1,f}.$$

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

$$H_{TF} = -\sum_{\alpha,\beta,\gamma \in \{A,C,G,T\}} P_0(\alpha\beta\gamma)\log_2 P_0(\alpha\beta\gamma) = H_{3,0}.$$

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

$$\tilde{H}_3 = H_3/3, \quad \tilde{H}_{NF} = H_{NF}/3, \quad \tilde{H}_{TF} = H_{TF}/3.$$

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.

2.2.2. *Results*

The values of these entropies in the prokaryotic protein genes are presented in Table 4.

The values of H_1 (resp. $\tilde{H}_3$) 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 close to 2 representing the random situation. The value $\tilde{H}_3$ (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 $\tilde{H}_{TF}$ (1.918 bit per nucleotide) associated with the TF method, is significantly lower than the value of $\tilde{H}_{NF}$ (1.965 bit per nucleotide) associated with the NF method. The $\tilde{H}_{TF}$ 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.

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* (13 686 *sequences*, 4 708 758 *trinucleotides*; *data from Arquès & Michel*, 1996, *p.* 49)

	Entropy in the frame 0 modulo 3	Classical entropy H_n
Nucleotide ($n = 1$)	$\tilde{H}_{NF} = 1.965$	$H_1 = 1.998$
Trinucleotide ($n = 3$)	$\tilde{H}_{TF} = 1.918$	$\tilde{H}_3 = 1.984$

The improvement of the estimate of the entropy is not the aim of this paper. However, the fact that the value of $\tilde{H}_{TF}$ 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 $\tilde{H}_{NF}$ differs significantly from the classical estimate. The hypothesis of independence between successive bases then has a strong effect on the values of the entropies. Therefore, the probability model used in the NF method reveals 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 as 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 three 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 a 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 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 1–6 (Koch & Lehmann, 1997). However, any probability model can produce 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, two 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 $\tilde{H}_{TF}$ 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 three 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 $\tilde{H}_{NF}$ is equal to 1.965 bit per nucleotide and significantly differs from $\tilde{H}_{TF}$. 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 do 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

Arquès, D. G., Fallot, J.-P. & Michel, C. J. (1998). An evolutionary analytical model of a complementary circular code simulating the protein coding genes, the 5′ and 3′ regions. *Bull. Math. Biol.* **60,** 163–194.

Arquès, D. G., Fallot, J.-P., Marsan, L. & Michel, C. J. (1999). An evolutionary analytical model of a complementary circular code. *J. Biosystems* **49,** 83–103.

Arquès, D. G. & Michel, C. J. (1996). A complementary circular code in the protein coding genes. *J. theor. Biol.* **182,** 45–58.

Arquès, D. G. & Michel, C. J. (1997a). A code in the protein coding genes. *J. Biosystems* **44,** 107–134.

Arquès, D. G. & Michel, C. J. (1997b). A circular code in the protein coding genes of mitochondria. *J. theor. Biol.* **189,** 273–290.

Béal, M.-P. (1993). *Codage Symbolique*. Paris: Masson.

Berstel, J. & Perrin, D. (1985). *Theory of Codes*. New York: Academic Press.

Chatzidimitriou-Dreismamm, C. A., Steriffer, R. M. F. & Larhammar, D. (1996). Lack of biological significance in the linguistic features of non-coding DNA. A quantitative analysis. *Nucl. Acids Res.* **24,** 1676–1681.

Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Crick, F. H. C., Griffith, J. S. & Orgel, L. E. (1957). Codes without commas. *Proc. Natl Acad. Sci.* U.S.A. **43,** 416–421.

CRICK, F. H. C., BRENNER, S., KLUG, A. & PIECZENIK, G. (1976). A speculation on the origin of protein synthesis. *Origins of Life* **7,** 389–397.

EIGEN, M. & SCHUSTER, P. (1978). The hypercycle. A principle of natural self-organization. Part C: the realistic hypercycle. *Naturwissenschaften* **65,** 341–369.

KOCH, A. J. & LEHMANN, J. (1997). About a symmetry of the genetic code. *J. theor. Biol.* **189,** 171–174.

LIO, P., POLITI, A., BUIATTI, M. & RUFFO, S. (1996). High statistics block entropy measures of DNA sequences. *J. theor. Biol.* **180,** 151–160.

LOEWENSTERN, D. & YIANILOS, P. N. (1999). Significantly lower entropy estimates for natural DNA sequences. *J. Comp. Biol.* **6,** 125–142.

NIRENBERG, M. W. & MATTHAEI, J. H. (1961). The dependance of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl Acad. Sci. U.S.A.* **47,** 1588–1602.

SHANNON, C. E. (1949). *The Mathematical Theory of Communication.* Champaign, IL: University of Illinois Press.