Journal of Theoretical Biology 275 (2011) 21–28

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Circular codes revisited: A statistical approach

D.L. Gonzalez [a,b], S. Giannerini [b,*], R. Rosa [b,a]

[a] CNR-IMM, Sezione di Bologna, Via Gobetti 101, I-40129 Bologna, Italy
[b] Dipartimento di Scienze Statistiche, Università di Bologna, via delle Belle Arti 41, I-40126 Bologna, Italy

ARTICLE INFO

Article history:
Received 1 September 2010
Received in revised form
18 January 2011
Accepted 19 January 2011
Available online 26 January 2011

Keywords:
Comma-free codes
Circular codes
Reading frame synchronization
Statistical analysis
Protein synthesis accuracy

ABSTRACT

In 1996 Arquès and Michel [1996. A complementary circular code in the protein coding genes. J. Theor. Biol. 182, 45–58] discovered the existence of a common circular code in eukaryote and prokaryote genomes. Since then, circular code theory has provoked great interest and underwent a rapid development. In this paper we discuss some theoretical issues related to the synchronization properties of coding sequences and circular codes with particular emphasis on the problem of retrieval and maintenance of the reading frame. Motivated by the theoretical discussion, we adopt a rigorous statistical approach in order to try to answer different questions. First, we investigate the covering capability of the whole class of 216 self-complementary, C^3 maximal codes with respect to a large set of coding sequences. The results indicate that, on average, the code proposed by Arquès and Michel has the best covering capability but, still, there exists a great variability among sequences. Second, we focus on such code and explore the role played by the proportion of the bases by means of a hierarchy of permutation tests. The results show the existence of a sort of optimization mechanism such that coding sequences are tailored as to maximize or minimize the coverage of circular codes on specific reading frames. Such optimization clearly relates the function of circular codes with reading frame synchronization.

1. Introduction

All the steps of genetic information processing that characterizes normal functioning of living organisms share an astonishing level of accuracy that pervades all the biochemical processes involved. DNA replication, RNA transcription and mRNA editing, protein translation, represent outstanding examples of such capabilities. In particular, protein translation accuracy depends mainly on the synthesis process performed at the level of the active site A of the ribosome and the correct charging of tRNAs with their cognate amino acid. But if the ribosome complex is able to maintain high levels of translation accuracy, as is experimentally observed, the mRNA template has to carry sufficient information for ensuring a correct decoding behaviour. Thus, the mRNA base sequence needs to include (explicitly or implicitly) appropriate information regarding: punctuation signs, direction of translation, synchronism with the normal reading frame and faithful assignation of codons to cognate amino acids. In fact, different organizational structures of mRNA sequences have been associated to these functions and the ability of the genetic code of carrying different levels of information along gene sequences has been recently remarked (Itzkovitz and Alon, 2007).

In some cases, codon translation has been shown to be context dependent; this indicates that linear reading of codons does not suffice for ensuring faithful translation of amino acids and punctuation signs. It is well known that this is particularly true for start and stop signals for which the presence of specific sequences of bases are needed in order to confirm the function of the punctuation codons. This is the case, for example, of the Shine–Dalgarno sequence situated some bases up-stream to the start codon site in prokaryotes. The need for sophisticated biochemical methods of mRNA information reading for translation accuracy has been put in terms of error correction of synthesis start signals (May, 2002). Also, in order to ensure a correct amino acid translation different reading hypotheses have been proposed; these include the use of intron information (Fordsyke, 1981) and non-linear decoding of di-nucleotides as a proof-reading mechanism (Gonzalez, 2008a, b).

Another important factor related to the accuracy of protein translation is the maintenance of the correct reading frame. This capability of the translation machinery has been associated to base periodicity and to the existence of circular codes in coding sequences. In particular, base periodicity has been identified on the purine/pyrimidine alphabet at the gene and gene population levels by Shepherd (1981), Fickett (1982), and Michel (1986). After these pioneering works, several researchers, studied these results on the other 2-letter alphabets (strong/weak, etc.) and the 4-letter alphabet (Arquès and Michel, 1996). Normal reading

* Corresponding author. Tel.: +39 0512098262.
E-mail addresses: gonzalez@bo.imm.cnr.it (D.L. Gonzalez), simone.giannerini@unibo.it (S. Giannerini), rodolfo.rosa@unibo.it (R. Rosa).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.01.028

frame synchronism has been related to the presence of circular codes in the structure of protein coding sequences (Michel, 2008; Trifonov, 1987). Such codes also explain the purine/pyrimidine periodicity (Arquès and Michel, 1996).

In Farabough and Björk (1999) it is shown experimentally that the accuracy in amino acid translation is related to the accuracy of frame maintenance. This fact is remarkable because it indicates a complex interaction between the two informational levels similarly to what happens in error detection/correction codes developed for transmitting digital information in technological applications.

In this paper we study the organizational structure of mRNA information in relation to the property of frame detection and maintenance by using circular code properties. In Section 2 we present a theoretical discussion on code theory in theoretical biology with particular emphasis on comma-free codes and on circular codes. The theoretical issues arisen from the discussion are studied in Section 3 by adopting a rigorous statistical approach. In particular, we aim at (i) quantifying the fitness of an arbitrary circular code with respect to a single sequence or an ensemble of coding sequences; (ii) investigating the existence of some sort of optimization mechanism that links circular codes to coding sequences. These task are accomplished by setting up a series of tests based on appropriate resampling techniques. In Section 4 we discuss the results and outline the conclusions.

The main results point to the existence of an optimization mechanism such that coding sequences maximize or minimize the coverage of circular codes on specific reading frames. Also, the two levels of information regarding amino acid recognition and frame maintenance are indeed independent. This means that a coding sequence can carry the information ensuring frame synchronization independently from the information regarding accurate amino acid translation. Of course, the synthesis mechanism, actuated by the ribosome, can take profit simultaneously of both kinds of information in a synergic process for ensuring overall protein translation accuracy. The results also shows that the roles of bases according to their relative position along codons are differentiated: frame maintenance seems to be mainly associated with the first two bases, while amino acid translation accuracy is mainly related to the third base.

2. Synchronizable codes

2.1. Comma-free codes

A typical problem encountered in the transmission of a message composed by words made up of strings of symbols taken from an alphabet is that of synchronization. In order to ensure a faithful decoding of words we need not only a faithful decoding of symbols but also an accurate signalling and decoding of start and end points of words. This allows to group the alphabet signs as to form admissible words. In common written language this last task is implemented by using a separation sign (the blank character) between adjacent words. For example the following message:

SOME ANTS ARE MIGRANTS

is easily interpreted by identifying any word of the message as the string between successive blanks. Now, if we eliminate the blanks the message reads,

SOMEANTSAREMIGRANTS

Now, the semantic interpretation of the message is ambiguous because it can be decoded in at least another different way, i.e.,

SO MEAN TSAR EMIGRANTS

All the words decoded in this last way are valid words of English language and all the alphabet symbols have been correctly decoded; still, this alternative decoding is erroneous. The bug resides in the identification of start and end points of the words, indeed, a frame decoding error. The blank symbol is used for separating adjacent words but any other symbol not used in word composition, that is, other than letters of the alphabet, can do the job as well; since commas are used for separating parts of a sentence, comma is a symbol that is naturally used for the former scope; for this reason, a code that can be decoded without the need of separation symbols is said a *comma-free* code. Moreover, if in a digital communication system all the transmitted words have the same length (as is the case, for example, of bytes of fixed length in binary computer systems), the problem of word separation becomes a problem of synchronization. Since the decoding operation in binary messages is implemented at a constant symbol rate, say t symbols/second, the time for decoding words of constant length n is constant and equal to the time needed to decode n symbols, i.e., nt seconds. Thus, a correct decoding of words is achieved by starting the decoding process exactly in synchrony with the start of a word and resetting for decoding a new word periodically after nt seconds. As the inverse of a time should be interpreted as a frequency, the problem is equivalent to keeping at the same time the word decoding frequency at $1/(nt)$ and the symbol decoding frequency at $1/t$. Now, maintaining two frequencies in a fixed rational ratio, i.e., $(1/nt)/(1/t)=1/n$, represents a problem of frequency synchronization; in this spirit, we call the general problem of word frame decoding, a synchronization problem. Frequency synchronization allows to achieve accurately the word length; of course, we also need phase accuracy for localizing correctly the starting points and reading the words in their correct frame. Consider the following example:

THE OWL WAS HOT (correct reading frame)

T HEO WLW ASH OT (frame $-$ shift $+1$ reading)

In this example we see a sequence of words of three letters. Thus, the word decoding frequency needs to be 1/3 of the symbol decoding frequency. By ensuring a frequency synchronization at this value (1/3) we achieve a correct length reading of words. But if we use a wrong starting point (phase shift) the words are decoded incorrectly. This is shown in the second line of the example where a frame-shift of 1 symbol (+1 frame-shift) has been introduced. Observe that now most of the words have no sense so that we can argue that an error has been introduced. This is the main idea behind comma-free codes: a frame-shift reading produces non-allowed words.

If the synchronization is ensured only at the beginning of one specific message, any perturbation along the decoding process can break the synchronization and, thus, the correct decoding of the reading frame; for this reasons, man-made digital communication systems implement means for monitoring (continuously or intermittently) the quality of frequency and phase synchronization and possibly restoring them in case of errors. Coding regions of DNA are very similar to digital computer messages where the alphabet symbols are the four bases (A, T, C, G) and individual words are formed by strings of three symbols forming an mRNA codon. The meaning of the words in coding sequences of DNA correspond to the coded amino acids which are sequentially assembled as to form the proteins in the ribosome complex; this is the so-called translation or protein synthesis step in the central dogma of molecular biology. Remarkably, the decoding frequency of this process can attain more than 20 amino acids per second in *Escherichia coli* at 37 °C (Bremer and Dennis, 2008). As in the transmission of any digital message, a faithful protein synthesis needs appropriate means for ensuring synchronization

of the decoding process with the correct reading frame of codons. Such ability is called *reading frame maintenance*; in this context frequency synchronization is not important because a correct word length reading is ensured by the stereo-chemical properties of the ribosome; instead, the correctness of the starting point is fundamental because once a frame-shift decoding is started it continues producing wrong amino acids until a wrong (out-of-frame) stop codon is encountered and the synthesis is stopped. A loss of frame maintenance (or frame-shift) produces a completely erroneous translation because the genetic code is compact (any sequence of three bases represents one amino-acid) and frame-shift reading errors usually lead to non-equivalent codons, that is, codons that represent different amino acids (as it can be deduced by inspection of the genetic code).

The problem of synchronization (or reading frame location and maintenance) in mRNA translation has been identified from the beginning of molecular genetics. Crick et al. (1957) proposed an ingenious solution for the problems of both amino acid coding through codons and frame maintenance. They proposed an ancestral genetic code with the property of being comma-free and non-degenerate. In doing so, Crick et al. (1957) introduced the comma-free codes for the first time, an interesting example of communication theory problems generated by a biological insight. Indeed, these codes have provoked a great interest from the mathematical point of view of coding theory (see e.g. Golomb et al., 1958). The main idea of comma-free codes is to use a subset of available words in such a way that a concatenation of valid words in the correct reading frame produces inadmissible words when it is read out of frame. In this sense, comma-free codes are error-correcting codes. A classic example is represented by the comma-free code implemented in a two symbol alphabet, i.e., the R, Y alphabet (R=purine=A, G; Y=pyrimidine=T, C). This alphabet has been proposed as a primeval alphabet related to the origin of the genetic code. If we consider the set formed by the two codons {RRY, RYY} they can be concatenated in the following ways:

RRYRRY
RRYRYY
RYYRYY
RYYRRY

If we allow any possible reading frame, the newly generated codons are: RYR, YRR, YRY, and YYR. Notice that all of them are not allowed codons, hence, the original set, {RRY, RYY} represents a comma-free code for words of three-letters over an alphabet of the two symbols, R, Y. If this code is used for coding amino acids, there are two valid words out of 8 (2^3) that have the property of automatic frame retrieval (there are no valid words in the out of frame reading). However, the price to be paid is the high redundancy which amounts to 3/4 of the total quantity of information (2/8=1/4 is the proportion of valid words). Furthermore, despite the great interest arisen, these kind of codes have not been found in current mRNA protein coding sequences but could have existed in primitive genes on a purine/pyrimidine alphabet.

2.2. Circular codes

Another class of codes that allows frame retrieval with less redundancy, is represented by the so called circular codes. These codes have the property of *synchronizability*, that is, they allow to retrieve the correct reading frame by using an appropriate window of mRNA bases. Circular codes obey less restrictive rules than comma-free ones. As shown before, comma-free codes are based on a kind of zip coding; only some codons have a sense and these are always placed in the correct reading frame. In fact, when such codes are built on three letter words over an alphabet of four they should have at most 20 meaningful codons out of the 64 possible. Circular codes, instead, are characterized by less redundancy. They possess the circular property, i.e., any word written on a circle (the last letter becoming the first in a torus like fashion) can be decomposed in at most one way in words of the circular code.

Suppose that a given codon belongs to a specific circular code; due to the circular property, the same code cannot contain any circular permutation of this codon. This is because the concatenation of a given codon with itself generates all its circular permutations if read in an out-of-frame situation. For instance, if the codon ATC belongs to a given code then its circular permutations—TCA, and CAT, which can be generated by reading out-of-frame the concatenation of the codon with itself ATCATCATC—need to be excluded from such code. Now, since the identical codons, AAA, TTT, CCC, GGG, coincide with their circular permutations, these codons are immediately excluded from any circular code. If we eliminate these codons from the 64 possible ones, we are left with 60 codons that can be grouped in 20 sets of three codons each. These sets are built by a codon together with its two circular permutations. In this way we can form an arbitrary circular code by choosing a number of codons from the 20 sets.

A circular code that contains exactly 20 codons is called a *maximal* circular code. The total quantity of maximal circular codes is 3^{20}. Thus, the probability of generating by chance a particular maximal circular code is $3^{-20} \approx 2.9 \times 10^{-10}$. For this reason, it is remarkable that specific maximal circular codes (and also more restrictive versions of them) have been actually found in genes. In fact, by studying the probability of occurrence of codons in the reading frames Arquès and Michel (1996) have identified a particular maximal circular code. The codons that form such code, called the X0 code, have a preferential frequency of occurrence in the normal reading frame; moreover, the authors assert that X0 is a common code for both eukaryote and prokaryote organisms. The code X0 exhibits many interesting symmetry properties; for example, the first circular permutation of X0, called X1, is found preferentially in the +1 out-of-frame condition; also, the second circular permutation, X2, is found in the +2 (or -1) out-of-frame condition. X1 and X2 are also maximal circular codes. This property of X0 is called the C^3 property. Moreover, the X0 code has also a peculiar characteristic regarding its symmetry under the complementary transformation, i.e., when the bases of a codon are changed by their complementary ones (A↔T, and C↔G), and the codon is read in the reverse direction. The X0 code is invariant under this transformation, or, in other words, the X0 code is self-complementary.

Other maximal circular codes have been found in different life domains (bacteria, archea, and mitochondria) (Arquès and Michel, 1997; Frey and Michel, 2003, 2006), but these codes do not have the self-complementary property. However, no biochemical mechanism linked to the functional use of circular codes has been reported so far. For this reason, tentative explanations for the existence of circular codes have been proposed. For example, Koch and Lehman (1997) noted that the distribution of bases occurrence in the different codon positions can generate in a natural way some circular codes. Assuming that these frequencies are related by a self-complementary symmetry, the generated codes share also such property. From these assumptions, as shown by Lacan and Michel (2001), it is possible to generate 88 maximal, C^3 and self-complementary circular codes. However, in the same article the authors proved that the common eukaryote–prokaryote X0 code cannot be generated in this way.

The authors implement an algorithm (the flower automaton) for generating all the 216 maximal self-complementary circular codes. Since the proportion of bases in actual genomic sequences

do not satisfy exactly the self-complementary condition the issue on which codes can be obtained by relaxing the conditions established in Koch and Lehman (1997) remains open. Moreover, recent works (Arquès and Michel, 1997; Frey and Michel, 2003, 2006) reveal the presence of many other circular codes in different organisms; also, the existence of evolutionary mechanisms responsible of codon variations from the archetypical common X0 circular code are hypothesized (Arquès et al., 1999; Ahmed et al., 2010). Moreover, Michel et al. (2008) identify a relation between the comma-free codes and the circular codes by constructing a hierarchy of these two classes of codes. All these results prompt two fundamental questions: (i) quantifying the fitness of an arbitrary circular code with respect to a single sequence or an ensemble of coding sequences; (ii) investigating the existence of some sort of optimization mechanism that links circular codes to coding sequences. In the next section, we explore these issues by means of a rigorous statistical approach.

3. Statistical analysis

In their approach Arquès and Michel (1996) (see also Michel, 2008) start from a set of coding sequences and derive circular codes from the frequency distribution of the codons in the three reading frames of such set. In particular, they find a common circular code, (AM code in the following) from a large set of eukaryote and prokaryote coding sequences. Our approach is complementary to theirs in that we look for the best codes among the whole class of 216 C^3 codes. To this aim, we define a measure (called *covering capability*) for describing the ability of a given circular code in covering a mRNA sequence. Then, we study the covering capability of the whole class of maximal, C^3, and self-complementary codes (denoted by $\mathfrak{C}$, in the following) in a sample of coding sequences. Moreover, we explore the covering capability of the 216 codes as a function of their distance from the AM code. We focus on the best code in terms of mean covering capability and assess the significance of the result by means of a bootstrap test. In addition, we analyse how the covering capability of a code is linked to (i) the organizational structure of a sequence; (ii) the proportion of bases in the three codon positions; and (iii) the reading frames. To this aim we implement a series of permutation tests. Finally, we analyse in some detail the distribution of the covering capability of some of the best codes. The 216 C^3 codes used have been taken from the lists in Michel et al. (2008).

We define the *covering capability* $\mathcal{C}_{x,s}$ of a circular code $x \in \mathfrak{C}$ for a sequence s, a parameter which describes the number of bases covered simultaneously by the code $x0$ in frame, $x1$ on the frame-shift +1 and $x2$ on the frame-shift +2 conditions. Formally we have the following:

Definition 1. Let $s=(c_1,\ldots,c_n)$ be a sequence of codons where ($s0$, $s1$, $s2$) denote the sequence s read in frame, with frame shift +1, with frame shift +2, respectively. In terms of bases, the i-th codon of the r-th frame shift is denoted by $\{c_i^r\}=(b_{3i+r-2},b_{3i+r-1}, b_{3i+r})$. Define the coverage of the circular code $x=\{x0,x1,x2\}$ over s as

$$\mathcal{C}_{x,s}=\frac{card[\bigcup_{r=0}^{2}\bigcup_{i=1}^{n} I(c_i^r,xr)]}{3n} \tag{1}$$

here, and in the following, $card[A]$ is the cardinality of the set A. $I(a,A)$ is the function defined as

$$I(a,A)=\begin{cases} a & \text{if } a\in A \\ \emptyset & \text{else} \end{cases} \tag{2}$$

Notice also that xr, with $r=0,1,2$, means $x0$, $x1$, $x2$.

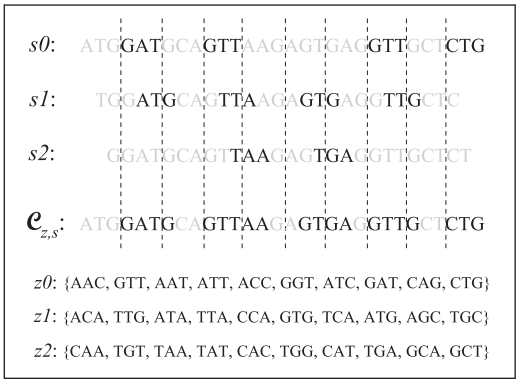

Fig. 1. Scheme explaining the covering capability $\mathcal{C}_{z,s}$ of a circular code z formed by 10 codons, with respect to a sequence of codons s. The circular code is represented as $z0$, while $z1$ is the first circular permutation of $z0$, and $z2$ is the second circular permutation of $z0$. $s0$, $s1$ and $s2$ denote the sequence s read in frame, with frame shift +1, with frame shift +2. The black letters in $\mathcal{C}_{z,s}$ are the bases of s covered simultaneously by $z0$, $z1$, $z2$ circular codes. The dashed lines guide the eye between codons.

As an example, let

$$s=(c_1,\ldots,c_{10})=\text{ATGGATGCAGTTAAGAGTGAGGTTGCTCTG}$$

be a sequence composed by 30 bases, see Fig. 1. Denote the sequence read in frame as $s0$. Now take a (non-maximal) circular code z of 10 codons. In the figure we report $z0$, i.e., the code z itself, $z1$, i.e., the first circular permutation of $z0$, and $z2$, i.e., the second circular permutation of $z0$. We see that out of the 10 codons of $s0$, there are four codons, GAT, CTG, GTT (twice), which belong also to $z0$ (in black in the figure); hence, we say that $z0$ "covers" 12 bases of $s0$. The same procedure is repeated with the code $z1$ as to obtain the covered bases of $s1$. In this case, the number of the covered bases of $s1$ is still 12. Lastly, we consider $z2$ and $s2$ and get six covered bases. Taking into account the "union" of the coverages by $z0$, $z1$ and $z2$, we obtain the total covered bases of the sequence s, represented in black in the fourth line of Fig. 1, namely 20 bases. Finally, the *covering capability* (in percentage) results $\mathcal{C}_{z,s}=20/30\times 100=66.7\%$.

In order to study the covering capability of the whole $\mathfrak{C}$ class, we use a data set consisting of 3408 nucleotide sequences obtained from genbank[1] by means of the R package *seqinr*[2] (Charif and Lobry, 2007). In detail, we have extracted all the coding sequences from the 13 classes of proteins reported in Table 1, where the number of sequences analysed for each class is listed. Such ensemble has been reduced by eliminating duplicate and short (<120 bases) sequences. The final data set consists of 3248 sequences. For each code $x\in\mathfrak{C}$ we have obtained a distribution of covering capability whose mean we call the *average covering capability* denoted by $\overline{\mathcal{C}}_x$ and defined as

$$\overline{\mathcal{C}}_x=\frac{1}{3248}\sum_{s=1}^{3248}\mathcal{C}_{x,s}$$

Fig. 2 displays the histogram of the average covering capability $\overline{\mathcal{C}}_x$ of the 216 circular codes. The dashed line represents a kernel density estimate. The distribution appears bimodal, with one peak around 53.8% and a second one around 63.8%. Such bimodality is representative of a cluster of 12 codes whose average covering capability is around 66%. Moreover, the average covering capability ranges from $\approx 35.53\%$ up to $\approx 66.88\%$ that corresponds to the AM code. This result shows that also from this

[1] http://www.ncbi.nlm.nih.gov/genbank/
[2] The query script is available upon request.

Table 1
Classes of proteins whose sequences have been analysed. The third and fourth columns report the number of sequences and the number of kilobases (kb) of each class, respectively.

	Protein	No. of seqs	kb
1	Albumin	142	101.5
2	Alpha-globin	57	15.2
3	Beta-globin	141	50.3
4	Carboxypeptidase A	31	35.2
5	Globulin	6	2.9
6	Glycogen synthase	346	458.3
7	Heat shock protein 70	1022	1148.1
8	Insulin	45	9.9
9	Lactate dehydrogenase	164	113.7
10	Lysozyme	327	161.5
11	Phosphoglycerate kinase	1077	1221.7
12	Phosphorylase kinase	10	18.2
13	Troponin C	40	17.6

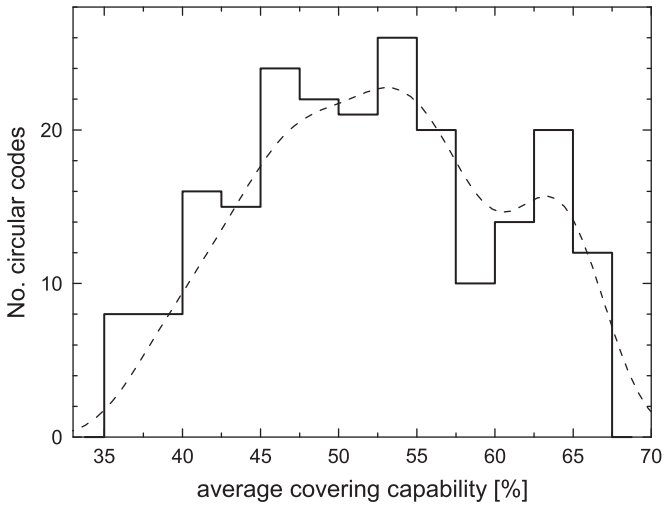

Fig. 2. Histogram of the average covering capability $\overline{\mathcal{C}}_x$ of the 216 circular codes. The dashed line shows a Gaussian kernel density estimate.

perspective, the AM code is a special one, in that, on average, it performs better than the other codes.

We can go further into the matter by computing $\overline{\mathcal{C}}_x$ as a function of the set distance between each $\mathfrak{C}$ code and the AM code. If x and y belong to the class $\mathfrak{C}$, then the distance $\delta_{x,y}$ between them is defined as $\delta_{x,y}=20-card\{x\cap y\}$ (see also Ahmed et al., 2010). The result is reported in Fig. 3.

The figure shows that codes having average covering capability similar to the AM code are also close to it from the point of view of the distance $\delta_{\mathrm{AM},y}$. Note that the set distance $\delta_{\mathrm{AM},y}$ is always a multiple of 2 as the codes are self-complementary. In particular, there are 6 codes at distance 2 from the AM code, that is, these codes differ from the AM code by 2 codons. Moreover, the greatest number of codes (42) are at distance 14 and at the greatest possible distance 20 there are 8 codes. As mentioned before, $\overline{\mathcal{C}}_{\mathrm{AM}}=66.88\%$. Notice that there are 11 codes with average covering capabilities greater than 65% (see dashed line in Fig. 3). Also, in the figure we have framed the points belonging to the subset of Koch and Lehmann codes (KL codes, in the following), formed by 88 codes taken from Table 3 of Lacan and Michel, 2001. We see that 2 out of such 88 codes are at distance 2 from the AM code, and there are 3 codes with average covering capabilities greater than 65%, i.e., very close to the AM code. Notice that, for any given distance, the best code is of the non-KL type; on the contrary, the worst one is always of KL type. We can make a coarse assessment of the probability of the joint occurrence that

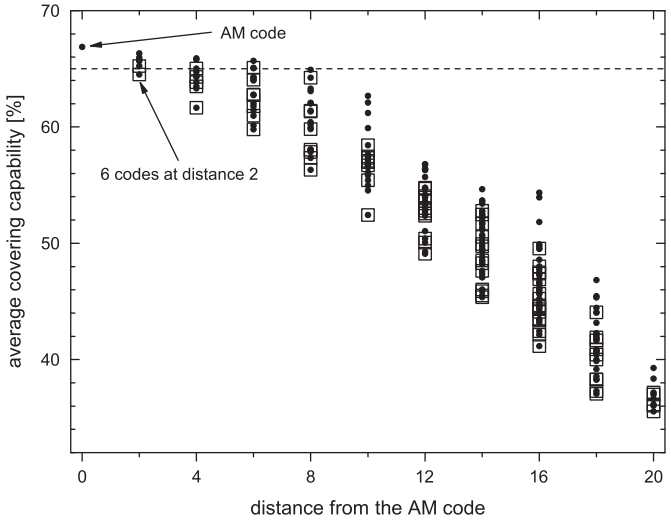

Fig. 3. Distance from the Arquès and Michel code *versus* average covering capability of each circular code. The 88 points framed with a square belong to the subset of Koch and Lehmann codes.

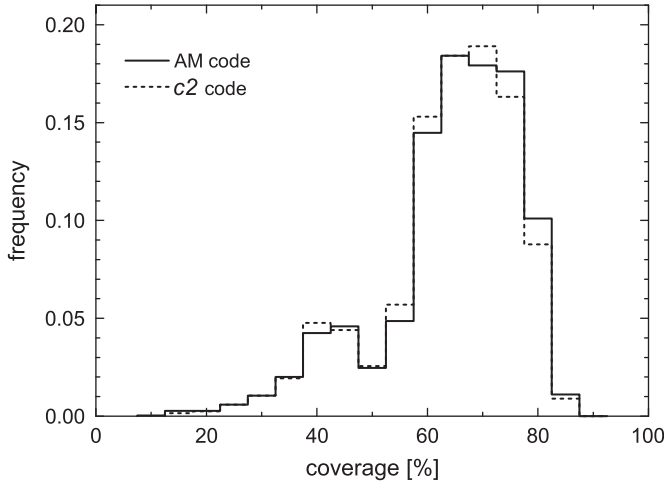

Fig. 4. Distributions of the covering capability of the Arquès and Michel code (continuous line) and of a code $c2\in\mathfrak{C}$ with the second highest average covering capability (dashed line).

by chance the best code is of non-KL type for all the distances. At a given distance $2k$ with $k=1,\ldots,10$ we have that the event *the best code is of KL type* can be modelled as a Bernoulli variable with parameter $p_{2k}=n^*_{2k}/n_{2k}$ where n_{2k} is the overall number of codes at distance $2k$ and n^*_{2k} is the number of codes of KL type at distance $2k$. Hence, the joint probability is given by $\prod_{k=1}^{10}(1-p_{2k})=0.0035$. This result assumes that the variables are independent and that the codes are equally likely; still, we can be quite confident of the unlikeliness that such an occurrence happened by chance.

Consider now the circular code having the average covering capability equal to 66.33%, which is the second highest after the AM code (call it *c2*). Fig. 4 shows the distributions of the covering capability of such two codes.

From the figure, we see that, for instance, the AM code, around its maximum, covers 598 sequences out of 3248 (i.e., 18.4%) with a covering capability between 62.5% and 67.5%. It appears that both distributions are quite similar and bimodal in shape. The histograms are not clearly separated, so at first sight, we are induced to think that the difference between the average covering capabilities of the AM code and the one of *c2*,

i.e., $d = 66.88 - 66.33 = 0.55\%$, is *not significant*. Nevertheless, by means of a rigorous statistical analysis, it turned out to be the opposite.

Formally, the problem reduces to testing the null hypothesis H_0, that the populations from which the observed samples have been drawn have the same mean or, equivalently, whether the difference of the mean values is zero. In this context, there are two aspects which deserve some care. The first one is that we are not legitimate to introduce the hypothesis of either normality or homoscedasticity, on which "traditional" statistical methods rest. The second aspect is that the samples under study are not independent; in other words we are in presence of *paired data*. A well-known example of such a situation is that of measuring on the same subject the effect (if any) of two different pharmacological treatments. If both measurements are made on the same subject, the variability between subjects is eliminated from the comparison, so that small treatment effects can be detected, even though the response of different subjects may be quite different. In our case, the "subjects" are the coding sequences and the "treatments" are the circular codes.

Because of the above mentioned issues, we exploited a non-parametric bootstrap method (Efron and Tibshirani, 1993) for paired data. The procedure consists of the following steps. We test the null hypothesis $H_0 : \tilde{C}_{AM} = \tilde{C}_{c2}$ against $H_1 : \tilde{C}_{AM} > \tilde{C}_{c2}$, where $\tilde{C}_{AM}$ and $\tilde{C}_{c2}$ denote the population means. Let $\{d_s\} = \{C_{AM,s} - C_{c2,s}\}$, $s = 1, \ldots, 3248$, be the observed sample of the differences with mean $\overline{d} = \overline{C}_{AM} - \overline{C}_{c2} = 66.88 - 66.33 = 0.55\%$. A bootstrap sample is obtained by randomly sampling 3248 times, with replacement, from $\{d_s\}$, this means to bootstrap the values $C_{AM,s}$ and $C_{c2,s}$ ($s = 1, \ldots, 3248$) in pairs. As a result, any d_s can be drawn more than once, or not at all. Let $\{d_s^*\}_1$ the first *bootstrap sample*, that is the first sample formed by the resampled values from $\{d_s\}$. Compute the first *bootstrap replication* $\overline{d}_1^*$, as the mean of $\{d_s^*\}_1$. Repeat the above computation a large number of times B (e.g. $B = 1000$) to obtain B bootstrap replications $\overline{d}_b^*$, ($b = 1, \ldots, B$). As usual, we choose the level α such as 0.05 or 0.01 and reject H_0 if the proportion of the bootstrap replications less than 0 (i.e., the bootstrap p-value $card\{\overline{d}_b^* < 0\}/B$) is lower than α. In the present case, the result of the bootstrap test is unambiguous: the bootstrap p-value is 0 so that the null hypothesis is rejected, that is, the average covering capabilities of the AM code is greater than those of other codes. Notice that, both the AM code and the *c2* code do not belong to the KL subset. We have repeated the above analysis for the best code of such subset and have found similar results. In other words, also this latter code has (i) a bimodal coverage distribution similar to the two codes studied above; (ii) an average covering capability significantly lower than that of the AM code.

In the following, we investigate in detail how the covering capability of the AM code is related to the structural information of a sequence under different randomization schemes (corresponding to different biological hypotheses) and for different frames. Contrarily to the coverage analysis performed above, this implies the study of the coverage of the AM code in terms of codons in the three reading frames, separately. In practice, we implement a series of permutation tests. The first scheme is produced by permuting without constraints the bases of the sequence (we denote this scheme by *no constr.*). In this scheme, for example, a letter from the first base position can be swapped with letters from the second or third base position; thus, only the global proportions of bases are preserved. Moreover, we explore a set of constrained permutation schemes that are associated to a hierarchy of hypotheses. In particular, we introduce schemes that involve the permutation of the bases in their position. In such a way, permuted sequences will have the same proportion of bases as the original sequence in the corresponding position. For instance, in the scheme denoted by $(b_1, \cdot, b_3)$ the first and the third bases of each codon are permuted (preserving their position) while the second bases do not vary. The most structured situation is created by permuting only the third bases of the sequence $(\cdot, \cdot, b_3)$. In this scheme, besides the proportion of bases, also the amino acid sequence is mostly preserved. Notice that the schemes (b_1, b_2, b_3) and *no constr.* are different in that in the former, letters are permuted by keeping their codon position. The permutation framework is similar to the bootstrap framework as the results of the two tests are asymptotically equivalent. In this context, however, permutation tests are preferred as they allow to preserve the proportion of bases in permuted sequences so that the hypotheses tested are biologically meaningful.

The main motivation of the inquiry is the following: *there is a mechanism that allows the synchronization of the frame by means of circular codes; such mechanism is related to the covering capability of the codes*. Now, the hypothesis we are testing is *Does the mechanism depend on the proportion of bases in the three positions of the codons?* Operationally, this translates into the following system of hypotheses:

As far as the covering is concerned:

$$\begin{cases} H_0 : \text{the original sequence does not differ from permuted sequences} \\ H_1 : \text{the original sequence does differ from permuted sequences} \end{cases}$$

In other words, if H_0 is true if we permute the sequence and preserve the proportion of bases then the coverage obtained from the permuted series will be the same of that derived from the original series. The procedure is the following: for each sequence and for the AM code X0 together with its circular permutations X1 and X2 (i) choose a permutation scheme and generate B randomly permuted sequences (e.g. $B = 1000$); (ii) compute the coverage of the AM code on the B permutations for the three frames as to obtain the distribution of the coverage on the permutations; (iii) perform a right tail test: reject H_0 if the coverage of the AM code on the original sequence is greater than the 95-th percentile of the permutations distribution. The results are reported in Table 2 where we show the percentage of rejections of the tests over the 3248 sequences for each permutation scheme (rows), for the three frames and for X0, X1, X2. For instance, consider the scheme $(b_1, \cdot, \cdot)$ (only the first bases of the codons in the sequence are permuted); we see that for the sequences in frame and for X0, we reject H_0 in the 44.2% of the 3248 sequences; in other words, for 1436 sequences the coverage of the X0 code of the sequences read in frame is significantly greater than those of the permutations.

The most striking results of Table 2 are the following. If we permute without constraints on the base position (*no constr.* scheme) we reject H_0 in about 80% of the sequences when the coverage of X0 (X1, X2, respectively) is computed upon in-frame (frame shift +1 and +2, respectively) sequences. We refer to the combinations: code X0—sequence in frame, code X1—sequence

Table 2
Percentages of rejections of the permutation tests over the 3248 sequences at level 95%, *right tail*, for the three frames and for X0, X1, X2. Each row corresponds to a different permutation scheme.

Bases	In frame			Frame shift+1			Frame shift+2		
	X0	X1	X2	X0	X1	X2	X0	X1	X2
$(b_1, \cdot, \cdot)$	44.2	11.5	3.4	13.2	12.1	21.1	15.3	15.3	11.0
$(\cdot, b_2, \cdot)$	39.4	7.9	6.6	17.8	18.0	17.6	9.0	4.5	29.2
$(\cdot, \cdot, b_3)$	34.8	11.8	10.2	9.8	15.8	10.5	8.7	7.4	17.8
$(b_1, b_2, \cdot)$	42.2	9.5	4.7	14.0	17.1	19.2	9.6	6.7	22.6
$(b_1, \cdot, b_3)$	41.8	9.6	4.6	13.8	17.1	19.0	9.7	6.6	22.4
$(\cdot, b_2, b_3)$	42.0	9.5	4.4	14.0	17.0	19.2	9.8	6.8	22.4
(b_1, b_2, b_3)	42.2	9.5	4.5	13.7	17.0	19.2	9.5	6.7	22.6
no constr.	81.2	8.4	8.2	8.7	79.6	10.4	7.5	8.1	80.8

frame shift +1, code X2—sequence frame shift +2, as the *natural combinations*. This means that the coverage of the AM code in the natural combinations drops significantly when the sequences are permuted. This finding suggests the presence of some sort of informational structure of the sequences that the circular codes are able to capture. The next step is to test the hypothesis that such structure is related to the proportion of bases in their codon positions. This hypothesis is tested by means of the scheme (b_1, b_2, b_3). Clearly, the rejection percentage passes from 81.2 to 42.2 for the coverage of the X0 code on the sequences read in frame. Interestingly, the percentages drop from 80% to about 20% for the other two cases. Also, the permutation of just one base is sufficient to cause the drop. Apart from the natural combinations, the general trend indicates that, in practice, the coverage of the codes on the original sequences is never greater than those of permuted sequences in almost every scheme.

These results suggest that coding sequences might be optimized in a way as to maximize the coverage in the natural positions. On the other hand, with the previous analysis, nothing can be said on a possible minimization process acting out of the natural positions. In other words, consider a coding sequence read in frame; if such sequence is built as to maximize the number of codons that belong to the code X0, then, is it also built in order to minimize the number of codons that belong to X1 and X2? In order to answer this question we need to implement the tests on the left tail of the permutation distributions. The procedure is the same of that described above with the exception of point (iii) that becomes (iii′) perform a left tail test: reject H_0 if the coverage of the AM code on the original sequence is smaller than the 5-th percentile of the permutations distribution. The results are reported in Table 3. For instance, consider the scheme $(b_1, \cdot, \cdot)$ we see that for the sequences in frame and for X0, in 4.6% of the 3248 sequences the coverage of the X0 code on the sequences read in frame is significantly smaller than those of the permutations.

The results of Table 3 are somehow complementary to those of Table 2. In fact, from the *no constr.* row we see that the percentages of rejection for non-natural combinations are high. This means that coding sequences in frame are made as to minimize the proportion of codons that belong to the codes X1 and X2, with a higher percentage for X2. This result is correlated with the unexpected code asymmetry between X1 and X2 observed in reading frames (see Fig. 2 in Arquès et al., 1997). Also, frame shift +1 and +2 sequences seem to minimize the proportion of codons that belong to the code X0 and X1 respectively. In conclusion, the results suggest that the proportion of bases *in their codon positions* is strictly connected to the organizational structure of a sequence and can play a role in the frame synchronization process. In the next section we review and discuss all the results in view of this perspective.

Table 3
Percentages of rejections of the permutation tests over the 3248 sequences at level 95%, *left tail*, for the three frames and for X0, X1, X2. Each row corresponds to a different permutation scheme.

Bases	In frame			Frame shift+1			Frame shift+2		
	X0	X1	X2	X0	X1	X2	X0	X1	X2
$(b_1, \cdot, \cdot)$	4.6	33.5	32.3	12.1	10.5	9.1	5.9	5.4	11.2
$(\cdot, b_2, \cdot)$	4.3	22.1	30.3	15.6	11.8	9.5	21.0	26.4	6.0
$(\cdot, \cdot, b_3)$	6.9	17.2	18.9	7.8	7.4	6.4	8.6	20.8	4.5
$(b_1, b_2, \cdot)$	6.2	28.2	35.0	14.3	11.3	8.7	12.5	22.6	7.5
$(b_1, \cdot, b_3)$	6.1	28.0	35.0	14.5	11.4	8.8	12.2	22.8	7.4
$(\cdot, b_2, b_3)$	6.1	28.1	34.8	14.4	11.7	8.7	12.0	22.7	7.6
(b_1, b_2, b_3)	6.2	27.6	35.0	14.3	11.5	8.9	11.9	22.9	7.5
no constr.	6.9	60.3	73.4	65.1	7.9	44.4	51.2	73.5	7.6

4. Discussion and conclusions

In this paper we have analysed the capability of C^3 (AM-like) codes in describing circular properties of mRNA and DNA protein coding sequences. The work of AM identifies an unique code for prokaryote and eukaryote coding sequences possessing particular symmetry properties. In particular the AM code is self-complementary, that is, if a codon belongs to X0 then also its complementary reversed version belongs to X0. Moreover, the circular permutations of X0, that is, X1 and X2, are also maximal circular codes (though not self-complementary). In relation with symmetry properties, Koch and Lehman (1997) proposed that the AM code should be a consequence of a self-complementary relation regarding the frequency of bases in the different codon positions (along the normal reading frame). The KL model is based on the hypothesis of absence of correlation between successive bases in (protein) genes. Entropy methods, in particular, showed that this hypothesis is not verified in current genes (e.g. Lacan and Michel, 2001). Therefore, the observation of a KL code in genes might be incompatible with a frequency dependence of bases in codon positions. In the framework of coding theory Lacan and Michel (2001) showed that (i) all the 216 codes can be generated with a flower automaton algorithm and (ii) the AM code, as expected, does not belong to the KL class. However, it is not clear if this difference in code building is significant at the functional biological level. The matter is important because, if circular codes can be generated simply from base proportions, then no particular organizational structure that relates a sequence to the circular codes is needed. We have investigated the issue from two different perspectives. In the first approach we have analysed the covering capability of the entire class of 216 C^3 codes. Also we have compared systematically the covering capability of KL versus non-KL codes. In the second approach we have studied how the covering capability of a circular code is related to the dependence structure and to the proportion of bases of a coding sequence. This is achieved by destroying the dependence of a sequence in a controlled fashion by means of different permutation schemes. Such schemes allow to preserve or destroy local and global proportions of bases as to build proper statistical tests.

Our first result confirm the primacy of the AM code over the entire class of 216 C^3 codes. This has been substantiated by means of a bootstrap test. However, by using a set distance we found codes of the KL type that are very close to the AM code. Fig. 3 shows the average covering capability of the 216 C^3 codes as a function of the distance from the AM code. Here, KL codes have been identified with a square. A simple visual inspection shows that the KL codes are interspersed over the entire set. However, for any given distance, the best code is of the non-KL type; on the contrary, the worst one is always of KL type. The probability that by chance the best code if of non-KL type for all the distances is about 0.0035.

Even if the average covering capability of the AM code is significantly higher than those of other codes a look at the distribution of the covering over the set of sequences shows an interesting scenario. In fact, the AM code, the second one in average covering capability, and the best code of KL type, have similar distributions (see also Fig. 4). All of them are bimodal with maxima placed around the same values, and all of them show similar dispersion features. The dispersion of the distribution becomes an important aspect of the discussion because it is intimately related to the validity of the original hypothesis of the existence of a single common code. In fact, the high variability of the covering capability, which can be as low as 35.5% for the AM code, shows that some codes (including KL ones) perform better than the AM code for specific sequences and/or classes of proteins.

In the second part of the analysis we have studied the connections between the organizational structure of a sequence, the covering capability of the AM code, the proportion of bases and the reading frames. To this aim, we have implemented a permutation test framework. The results show that there is always a preferred frame for the AM code and its circular permutations. A permutation that preserves the global proportions destroys the covering in nearly 80% of the sequences. But what happens with the remaining 20%? Probably, due to the high dispersion discussed above, the left tail sequences are those that the AM code cannot cover well. Thus, the coverage is not destroyed because there is not coverage at all. If we implement a similar permutation on the base positions but keep the base frequencies (scheme (b_1,b_2,b_3)) we obtain a drastic fall in the percentages (from 80% to approximately 45%). This means that nearly half of the sequences lose their circular properties also when the frequency distribution of the bases in their position is maintained. This fact represents a clear evidence of the existence of an organizational level beyond the frequency distribution; still, it seems that this property is not universal as it is present in only one half of the sequences. In fact, in approximately 35% of the sequences, preserving the local proportions of bases is sufficient for maintaining the circular properties. Of course, this analysis is put forward for the AM code and does not ensure that other codes do not give different results. Again, this fact points to the need of defining in biological terms the appropriateness of a given code for describing a specific sequence.

Another interesting result is obtained by permuting only one letter at a time. It can be observed that the most stable situation is represented by permuting the third letters. This means that the circular coding is most insensible to permutations in the third letter. Since the third letter represents the true degree of freedom of the genetic code, it seems that synonymous codons can be chosen relatively freely without destroying the circular coding features of the entire sequence. In this way, the third letter might be used for implementing error correction mechanisms that do not interfere with the frame synchronization process.

In conclusion, the analysis presented contributes to shed light upon the role played by circular codes in the processing of genetic information. At the same time, new questions arise and more theoretical work is required. A direction of future investigation is that of exploring the connections between circular codes and the theory associated to the mathematical model for the genetic code presented in Gonzalez (2004, 2008b) and Gonzalez et al. (2006, 2008, 2009). Preliminary studies seem to indicate a clear relation between circular codes, global transformations and dichotomic classes.

References

Ahmed, A., Frey, G., Michel, C.J., 2010. Essential molecular functions associated with the circular code evolution. J. Theor. Biol. 264, 613–622.

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

Arquès, D.G., Fallot, J.-P., Michel, C.J., 1997. An evolutionary model of a complementary circular code. J. Theor. Biol. 185, 241–253.

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., 1997. A circular code in the protein coding genes of mitochondria. J. Theor. Biol. 189, 273–290.

Bremer, H., Dennis, P., 2008. Feedback control of ribosome function in *Escherichia coli*. Biochimie 90 (3), 493–499.

Charif, D., Lobry, J., 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla, U., Porto, M., Vendruscolo, M. (Eds.), Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Springer Verlag, New York, pp. 207–232.

Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proc. Math. Acad. Sci. USA 43, 202–209.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Farabough, P.J., Björk, G.R., 1999. How translational accuracy influences reading frame maintenance. EMBO J. 18, 1427–1434.

Fickett, J.W., 1982. Recognition of protein coding regions in DNA sequences. Nucl. Acids Res. 10, 5303–5318.

Fordsyke, D., 1981. Are introns in-series error-detecting sequences? J. Theor. Biol. 93, 861–866.

Frey, G., Michel, C.J., 2003. Circular codes in archaeal genomes. J. Theor. Biol. 223, 413–431.

Frey, G., Michel, C.J., 2006. Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. Comput. Biol. Chem. 30, 87–101.

Golomb, S.W., Gordon, B., Welch, L.R., 1958. Comma-free codes. Canad. J. Math. 10, 202–209.

Gonzalez, D.L., 2004. Can the genetic code be mathematically described? Med. Sci. Monit. 10 (4), 11–17.

Gonzalez, D.L., 2008a. Error detection and correction codes. In: Barbieri, M., Hoffmeyer, J. (Eds.), The Codes of Life: The Rules of Macroevolution, Biosemiotics, vol. 1. Springer, Netherlands, pp. 379–394 (Chapter 17).

Gonzalez, D.L., 2008b. The mathematical structure of the genetic code. In: Barbieri, M., Hoffmeyer, J. (Eds.), The Codes of Life: The Rules of Macroevolution, vol. 1. Springer, Netherlands, pp. 111–152 (Chapter 8).

Gonzalez, D.L., Giannerini, S., Rosa, R., 2006. Detecting structure in parity binary sequences: error correction and detection in DNA. IEEE Eng. Med. Biol. Mag. 25, 69–81.

Gonzalez, D.L., Giannerini, S., Rosa, R., 2008. Strong short-range correlations and dichotomic codon classes in coding DNA sequences. Phys. Rev. E 78 (5), 051918.

Gonzalez, D.L., Giannerini, S., Rosa, R., 2009. The mathematical structure of the genetic code: a tool for inquiring on the origin of life. Statistica LXIX (3–4) 143–157.

Itzkovitz, S., Alon, U., 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res. 17, 405–412.

Koch, A.J., Lehman, J., 1997. About a symmetry of the genetic code. J. Theor. Biol. 189, 171–174.

Lacan, J., Michel, C.J., 2001. Analysis of a circular code model. J. Theor. Biol. 213, 159–170.

May, E.E., 2002. Analysis of coding theory based models for initiating protein translation in prokaryotic organisms. Ph.D. Thesis, North Carolina State University, Raleigh, NC.

Michel, C., Pirillo, G., Pirillo, M., 2008. A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theor. Comput. Sci. 401 (1–3), 17–26.

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

Michel, C.J., 2008. A 2006 review of circular codes in genes. Comput. Math. Appl. 55, 984–988.

Shepherd, J.C.W., 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl Acad. Sci. USA 78, 1596–1600.

Trifonov, E.N., 1987. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. J. Mol. Biol. 194 (4), 643–652.