Strong short-range correlations and dichotomic codon classes in coding DNA sequences

Diego Luis Gonzalez,[1,2] Simone Giannerini,[3,*] and Rodolfo Rosa[3,2]
[1]*CNR-Fondazione scuola di S. Giorgio, I-30124, Venezia, Italy*
[2]*CNR-IMM, Sezione di Bologna, Via Gobetti 101, I-40129, Bologna, Italy*
[3]*Dipartimento di Scienze Statistiche, Università di Bologna, Via delle Belle Arti 41, I-40126, Bologna, Italy*
(Received 9 February 2008; revised manuscript received 7 October 2008; published 19 November 2008)

The study of correlation structures in DNA sequences is of great interest because it allows us to obtain structural and functional information about underlying genetic mechanisms. In this paper we present a study of the correlation structure of protein coding sequences of DNA based on a recently developed mathematical representation of the genetic code. A fundamental consequence of such representation is that codons can be assigned a parity class (odd-even). Such parity can be obtained by means of a nonlinear algorithm acting on the chemical character of the codon bases. In the same setting the Rumer's class can be naturally described and a new dichotomic class, the hidden class, can be defined. Moreover, we show that the set of DNA's base transformations associated to the three dichotomic classes can be put in a compact group-theoretic framework. We use the dichotomic classes as a coding scheme for DNA sequences and study the mutual dependence between such classes. The same analysis is carried out also on the chemical dichotomies of DNA bases. In both cases, the statistical analysis is performed by using an entropy-based dependence metric possessing many desirable properties. We obtain meaningful tests for mutual dependence by using suitable resampling techniques. We find strong short-range correlations between certain combinations of dichotomic codon classes. These results support our previous hypothesis that codon classes might play an active role in the organization of genetic information.

DOI: 10.1103/PhysRevE.78.051918 PACS number(s): 87.10.−e, 87.14.gk, 87.15.Qt

I. INTRODUCTION

The study of correlation structures in DNA sequences is of great interest because it allows us to obtain structural and functional information on the genetic mechanisms. In particular, short-range correlations of three base pairs have been generically ascribed to the codon organization of protein coding regions (see [1], and references therein). However, a satisfactory theoretical explanation for such short-range correlations which, for instance, takes into account the importance of the relative position of the bases along the reading frame, is still lacking. In this paper, we investigate this problem by means of a recent mathematical model of the genetic code [2–4] together with rigorous statistical methods applied to protein coding sequences of DNA. The mathematical model is used as a front end for the binary coding of such sequences.

The genetic code is a translation table connecting different biochemical words: the world of nucleic acids, molecules that store the relevant biological information, and the world of proteins, the essential chemical bricks for cellular metabolism. Every group of three contiguous bases (a codon) in messenger RNA (mRNA) is assigned to a specific amino acid by the genetic code; this determines the linear assembling ordering of such amino acids in the forming polymeric chain of a specific protein. The four bases used for coding genetic information in the single helix of mRNA are uracil (U), cytosine (C), adenine (A), and guanine (G). The actual sequence of bases in a mRNA is obtained through the substitution of the thymine base (T) by uracil in the coding segments of DNA. This process is called transcription, while protein synthesis performed by the ribosome complex following mRNA instructions is called translation. Since the information content of both coding DNA and mRNA is the same, in the following we use either U or T interchangeably.

The total possible number of codons in mRNA is 64, i.e., all the combination of four objects (the four bases U, C, A, G) in groups of three (the number of bases in a codon). As the amino acids used for proteins synthesis are only 20 (21 different symbols if we include the *stop* signal marking the end of protein synthesis), degeneracy and redundancy therefore follow. In fact, one of the main topics related to the research on the genetic code has been the study of its degeneracy properties. One important fact related to such degeneracy was noted early by the Russian theoretical physicist Yu. B. Rumer in the 1960's [5]. Rumer showed that exactly one-half of the quartets of the genetic code specifies amino acids with degeneracy 4 (a family), while the other half specifies amino acids with degeneracy 1, 2, or 3. We recall that a quartet is a group of four codons sharing the first two letters, as for example, [UUx]=[UUU,UUC,UUA,UUG].

Rumer's class is a dichotomic class in that a codon can assume either the value 4 or the value (1, 2, or 3); see Table I. Rumer's key observation was that a global transformation acting on the bases, i.e., $\mathrm{U,C,A,G}\leftrightarrow\mathrm{G,A,C,U}$, transforms a codon of class 4 into a codon of class 1, 2, or 3, and vice versa. In this respect, Rumer's transformation reveals the existence of an intrinsic antisymmetric property of the genetic code. In the following we show that, by using a recently developed mathematical theory for the genetic code, two new codon classes can be defined, i.e., the parity and the hidden classes. These classes are also antisymmetric with respect to the other two global transformations of the bases. We show also that the three global transformations of the

*Corresponding author. simone.giannerini@unibo.it

TABLE I. (Color online) Graphical representation of the classification of triplets in Rumer's classes. Green (light gray) boxes indicate triplets belonging to the class {1,2,3}, red (dark gray) boxes indicate triplets belonging to the class {4}.

	T	C	A	G	
T	TTT Phe	TCT Ser	TAT Tyr	TGT Cys	T
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys	C
	TTA Leu	TCA Ser	TAA Stop	TGA Cys	A
	TTG Leu	TCG Ser	TAG Stop	TGG Trp	G
C	CTT Leu	CCT Pro	CAT His	CGT Arg	T
	CTC Leu	CCC Pro	CAT His	CGC Arg	C
	CTA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CTG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser	T
	ATC Ile	ACC Thr	AAC Asn	AGC Ser	C
	ATA Ile	ACA Thr	AAA Lys	AGA Arg	A
	ATG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GTT Val	GCT Ala	GAT Asp	GGT Gly	T
	GTC Val	GCC Ala	GAC Asp	GGC Gly	C
	GTA Val	GCA Ala	GAA Glu	GGA Gly	A
	GTG Val	GCG Ala	GAG Glu	GGG Gly	G

bases, together with the identity, form a Klein V group of symmetry. Moreover, dichotomic codon classes can be obtained by means of nonlinear operators that act on matrices built from four consecutive bases of the DNA sequence.

Some authors have proposed that error correction capabilities of the code in a coevolutive context (coevolution of the genetic information and the genetic codes) might be related to the actual shape of the genetic code [6–8]. We also conjecture that the structure of the code is tightly linked to its error detection or correction capabilities (as previously proposed in [3,9]). However, in our case, such possibility is investigated on the basis of a direct mathematical description of the genetic code. Also other approaches (not necessarily related to the error correction hypothesis) have highlighted in an evolutionary context the existence of mathematical structures in the genetic code. We mention, for example, those based on a Lie algebra [10], and, more recently, on p-Adic numbers [11]. A different approach based on a quantum algebra ($\mathcal{U}_{q\to 0}[sl(2)\oplus sl(2)]$) has been extensively investigated [12]. For recent developments and applications based on this model, see [13] and references therein. Notice that our approach differs from previous ones since it is based upon discrete symmetries related to *nonpower* integer number representations; our model describes all the degeneracy properties of the genetic code in a unified fashion [2–4]. Moreover, the new definition of the parity codon class reinforces the interesting possibility of error detection or correction mechanisms based on parity control. Such a possibility turns out to be even more attractive in the light of recent results showing that parity can be defined at a molecular level and represents a key element for chemical error correction and consequently for the selection of actual bases in present DNA and RNA molecules [14].

The paper is organized as follows. In the second section we briefly present the mathematical model of the genetic code. Section III is devoted to the definition of the two dichotomic codon classes that arise from the mathematical model. Section IV presents the statistical framework that allows the analysis of coding sequences of DNA in this context. In Sec. V we show and discuss the results pertaining to the statistical analysis of protein coding sequences. Such results motivate the theoretical definition of the third dichotomic class, the hidden class. Finally, Sec. VI presents conclusions and perspectives for future investigations.

II. MATHEMATICAL BACKGROUND

In this section we briefly present the mathematical model of the genetic code used for the definition of the codon dichotomic classes. The reader is referred to [2–4] for a comprehensive description. Such model is based on the so called nonpower representation of integer numbers [15] which allows to explain many structural properties of the degeneracy distribution and new symmetry transformations of the genetic code.

Usual positional number representations are called power representations because the numbers are additively decomposed following the powers of a number called the basis of the numeration system. Decimal systems use the basis $b=10$ and thus a number represented in this system is additively decomposed in the powers of 10. For example, the number 93 458 means that $93\,458=9\times10^4+3\times10^3+4\times10^2+5\times10^1+8\times10^0$. We know that in the decimal system the digits representing the number are limited between 0 and $(b-1)=9$. This ensures a one-to-one representation. However, we are interested in describing nonbijective (i.e., non-one-to-one) applications and thus we resort to nonpower number representations. In nonpower number representations the positional values grow more slowly than the powers of the system basis b. For example, for a binary system ($b=2$), such values need to grow more slowly than the powers of 2. The redundant Fibonacci number representation is a good example of this kind of systems. Fibonacci numbers, i.e., 1, 1, 2, 3, 5, 8, …, grow more slowly than the powers of 2, i.e., 1, 2, 4, 8, 16, 32, … . The Fibonacci representation is of little interest in this context because it does not possess the same degeneracy distribution of the genetic code. Nevertheless, it can be proved [2] that a unique set of nonpower bases, i.e., 1, 1, 2, 4, 7, 8 describes exactly the degeneracy of the genetic code. Such a description is a structural isomorphism between the genetic code and the nonpower representation; it can become a proper model by establishing correspondences between represented numbers and amino acids on the one side, and binary strings and chemical codons, on the other side. This is achieved by comparing the symmetry properties of the genetic code with those of the nonpower representation. Table II synthesizes the main features of the model (see caption).

III. DICHOTOMIC CODON CLASSES

The model represented in Table II associates a length-6 binary string to every codon of the genetic code and a whole number from 0 to 23 to every amino acid (including the stop

TABLE II. (Color online) Representation of the first 24 whole numbers (outer columns) in the nonpower representation defined by the positional weights [1 1 2 4 7 8] (length-6 binary strings, horizontal rows). The degeneracy number (D), number of binary strings that represent the same whole number, and the corresponding amino acids are shown in the center of the table. Notice that the table is symmetric (palindromic symmetry) and that the amino acids are associated in pairs (pairs of palindromic amino acids, e.g., Trp/Met). The colors indicate the parity of each string (green/light gray=odd, red/dark gray=even).

#	8 7 4 2 1 1	8 7 4 2 1 1	8 7 4 2 1 1	8 7 4 2 1 1	D	Amino acids pairs	8 7 4 2 1 1	8 7 4 2 1 1	8 7 4 2 1 1	8 7 4 2 1 1	#
0	0 0 0 0 0 0				1	W Trp M Met				1 1 1 1 1 1	23
1	0 0 0 0 1 0	0 0 0 0 0 1			2	S Ser 2 F Phe			1 1 1 1 1 0	1 1 1 1 0 1	22
2	0 0 0 1 0 0	0 0 0 0 1 1			2	Ter K Lys			1 1 1 1 0 0	1 1 1 0 1 1	21
3	0 0 0 1 1 0	0 0 0 1 0 1			2	Y Tyr N Asn			1 1 1 0 1 0	1 1 1 0 0 1	20
4	0 0 1 0 0 0	0 0 0 1 1 1			2	L Leu 2 R Arg 2			1 1 1 0 0 0	1 1 0 1 1 1	19
5	0 0 1 0 1 0	0 0 1 0 0 1			2	H His D Asp			1 1 0 1 1 0	1 1 0 1 0 1	18
6	0 0 1 1 0 0	0 0 1 0 1 1			2	Q Gln E Glu			1 1 0 1 0 0	1 1 0 0 1 1	17
7	0 0 1 1 1 0	0 0 1 1 0 1	0 1 0 0 0 0		3	C Cys I Ile		1 0 1 1 1 1	1 1 0 0 1 0	1 1 0 0 0 1	16
8	1 0 0 0 0 0	0 1 0 0 1 0	0 1 0 0 0 1	0 0 1 1 1 1	4	S Ser 4 T Thr	1 1 0 0 0 0	1 0 1 1 1 0	1 0 1 1 0 1	0 1 1 1 1 1	15
9	1 0 0 0 1 0	1 0 0 0 0 1	0 1 0 1 0 0	0 1 0 0 1 1	4	P Pro A Ala	1 0 1 1 0 0	1 0 1 0 1 1	0 1 1 1 1 0	0 1 1 1 0 1	14
10	1 0 0 1 0 0	0 1 0 1 1 0	0 1 0 1 0 1	1 0 0 0 1 1	4	V Val G Gly	0 1 1 1 0 0	1 0 1 0 0 1	1 0 1 0 1 0	0 1 1 0 1 1	13
11	1 0 0 1 1 0	1 0 0 1 0 1	0 1 1 0 0 0	0 1 0 1 1 1	4	L Leu 4 R Arg 4	1 0 1 0 0 0	1 0 0 1 1 1	0 1 1 0 1 0	0 1 1 0 0 1	12

signal). Notice that the codon AUG codifies both the the *start* signal and the amino acid Methionine. In [3] it has been shown that the parity of a codon's binary string, defined as the parity of the digits sum, is related to the chemical properties of the two last bases of the codon. This parity index can be extracted by means of a nonlinear rule as shown in Fig. 1(a). For the definition of both the parity and other codon class dichotomies we use the unique three possible chemical binary classifications for the bases T, C, A, and G; the symbolic labeling is the following:

$$\{\text{purine;pyrimidine}\}\ \{R;Y\}\ \{A,G;C,T\},$$

$$\{\text{keto;amino}\}\ \{K;Am\}\ \{T,G;A,C\},$$

$$\{\text{strong;weak}\}\ \{S;W\}\ \{C,G;A,T\}.$$

In words, the rule of Fig. 1(a) can be described as follows. If the last letter of the codon is a purine (R=A,G), the parity of the binary string is immediately obtained: an A corresponds to an odd string and a G to an even string. If the last letter is instead a pyrimidine (Y=C,T), in order to determine the parity we need to observe the chemical character of the previous base in the codon, that is, the second or middle base. However, in such a case we have to consider a different chemical dichotomy: if the second base belongs to the amino class (Am=A,C), the corresponding string is even; if, instead, it belongs to the keto class (K=T,G), the corresponding string is odd. Now we show for the first time that a completely similar rule holds also for the determination of Rumer's degeneracy classes. In order to achieve this (i) shift the analysis window to the first two bases of the codon, (ii) consider the keto-amino dichotomy for the middle base [as suggested by the parity algorithm; see Fig. 1(a)], (iii) use the dichotomy class strong (S=C,G) or weak (W=A,T) for the first base. The algorithmic rule for Rumer's class is presented in Fig. 1(b). At this point it is important to notice that there exists a global transformation of the bases (Rumer's transformation), i.e., (T,C,A,G↔G,A,C,T), that flips the Rumer's class of a codon. This means that Rumer's class is antisymmetric with respect to Rumer's transformation. Notice that such transformation is a global transformation acting on all the three bases of the codon; however, the same effect on the Rumer's class is obtained if we consider only the first two bases of the codon (the third base is noninfluential). Moreover, Rumer's transformation can be seen as a composition between the other two possible global transformations. These transformations (T,C,A,G↔A,G,T,C, and T,C,A,G↔C,T,G,A) exchange bases inside the S-W and R-Y dichotomies, respectively. Observe that Rumer's transformation corresponds to the exchange inside the third possible chemical dichotomy (K-Am). Furthermore, following the parity rules defined before, it is easy to ascertain that the R-Y transformation changes for sure the parity of a codon. Thus we find that parity and Rumer's classes are antisymmetric with respect to the two global transformations K-Am and R-Y, respectively. In view of these findings, it is somehow natural to hypothesize that a third dichotomic class exists. Such class should be antisymmetric with respect to the third global transformation (S-W) and should be also deter-

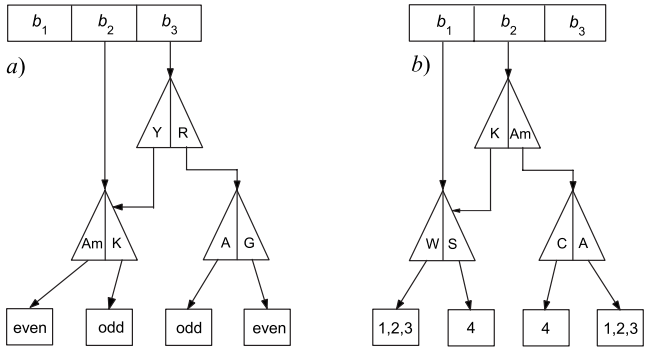

FIG. 1. Algorithmic representation of the codon dichotomic classes: (a) parity class, (b) Rumer's class; 1…4 indicate the degeneracy.

mined by a nonlinear rule similar to that shown in Figs. 1(a) and 1(b). Remarkably, such framework defines a Klein V group structure. In order to prove such a statement, define the bases as four-dimensional column vectors as follows:

$$\mathrm{T}' = (1000), \quad \mathrm{C}' = (0100), \quad \mathrm{A}' = (0010), \quad \mathrm{G}' = (0001), \tag{1}$$

where the prime symbol denotes transposition. The possible global transformations of the bases are implemented by the usual matrix product together with the following permutation matrices:

$$L = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad M = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$N = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \tag{2}$$

which are associated to the transformations

$$\mathrm{T,C,A,G} \leftrightarrow \mathrm{G,A,C,T}; \quad \mathrm{T,C,A,G} \leftrightarrow \mathrm{C,T,G,A};$$

$$\mathrm{T,C,A,G} \leftrightarrow \mathrm{A,G,C,T},$$

respectively. If we include the identity matrix, I, the set (2) form an Abelian group, indeed the Klein V group. In fact, it is easy to ascertain that L,M,N are orthogonal and the following identities hold:

$$LM = ML = N, \quad LN = NL = M, \quad MN = NM = L.$$

Moreover, if we define the infinite order matrix norm for a $m \times m$ square matrix Q as

$$\|Q\|_\infty = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{m} |q_{ij}|$$

we can obtain the following operators:

$$O_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 2 & 2 & 1 \\ 0 & 0 & 3 & 4 \end{pmatrix}, \quad O_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 2 \\ 0 & 4 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

These operators act on a 4×4 square matrix built up with four consecutive vectors or bases. The values of the two classes c_1=parity, c_2=Rumer can be obtained through the following operation:

$$c_i = \|O_i \odot Q'\|_\infty \bmod 2, \quad i = 1,2, \tag{3}$$

where $\odot$ denotes the matrix Hadamard product.

Now, the hypothesized third dichotomic codon class (we call it the *hidden* class) by analogy with Rumer's and parity classes needs to be defined with a further shift of the analysis window by one base position to the 5′ end of the sequence. Hence the hidden class is determined by the first letter of a codon and the last letter of the previous one. Moreover, if we maintain the choice of the S-W dichotomy induced by Rumer's class on the first letter of the codon, the chemical dichotomic class for the last letter of the previous one will be necessarily of one of the two possible types, R-Y or K-Am. In this way the antisymmetric property of the third global transformation (S-W), i.e., $\mathrm{T,C,A,G} \leftrightarrow \mathrm{A,G,T,C}$, is kept. In the following sections we will show that the right choice is the K-Am dichotomic arising naturally from the statistical analysis of protein coding DNA sequences.

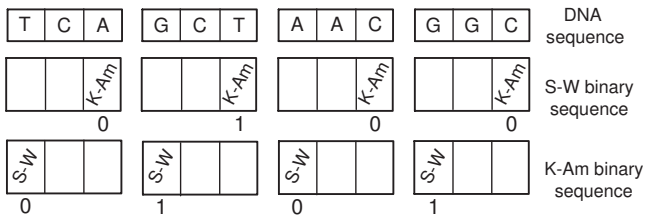

FIG. 2. Scheme of the position dependent coding for the K-Am class in the third codon position and the S-W class in the first codon position.

IV. METHODS

In a previous work [3] we have analyzed the univariate dependence structure of parity sequences generated from protein coding DNA regions, that is, we focused on the statistical properties of single sequences. Now, the theoretical framework developed here requires a multidimensional, position dependent approach, e.g., we compare pairwise chemical or dichotomic codon classes. An example of the position dependent chemical coding is illustrated in Fig. 2. The first row reports the DNA sequence, while the second and the third rows show the K-Am coding performed on the third codon bases and the S-W coding performed on the first codon bases, respectively.

As concerns the dichotomic codon classes, notice that the coding is computed on two contiguous bases and depends on the open reading frame. The coding is illustrated in Fig. 3 where the bases involved are highlighted in cyan (gray); in this case, both classes are coded "out of frame 1," namely, the coding starts from the second base.

The statistical analysis of binary sequences obtained by means of our coding framework is based on the implemen-

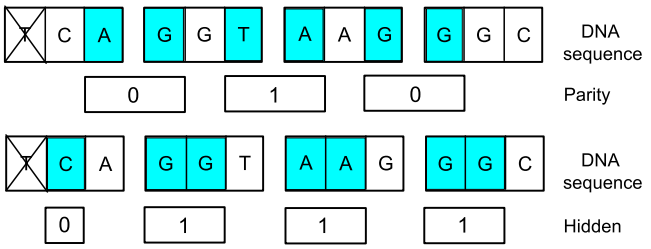

FIG. 3. (Color online) Scheme of the position dependent coding for the parity and the hidden classes, both of them out of frame 1, i.e., the first base is discarded. The coding involves a window of two consecutive bases in cyan (gray).

tation of a bivariate version of the metric entropy measure S_ρ, a normalized variant of the Bhattacharya-Hellinger-Matusita distance. This version corresponds to a two-dimensional implementation of the methods employed in [3].

The measure S_ρ is defined as follows:

$$S_\rho(k) = \frac{1}{2}\int\int [\sqrt{f_{(X_t,Y_{t+k})}(x,y)} - \sqrt{f_{X_t}(x)f_{Y_{t+k}}(y)}]^2 dxdy,$$

where $f_{X_t}(\cdot)$ and $f_{(X_t,Y_{t+k})}(\cdot,\cdot)$ denote the probability density function of X_t and of the vector (X_t, Y_{t+k}) respectively. In our case $X_t(\cdot)$ is a random process that measures which nucleotide appears at position t whereas (X_t, Y_{t+k}) is the bivariate random process that measures the joint appearance of $X_t(\cdot)$ at position t and $Y_{t+k}(\cdot)$ at position $t+k$. S_ρ is in precise mathematical relationship with other entropy functionals such as Shannon entropy and Kullback-Leibler divergence and can be interpreted as a nonlinear crosscorrelation function possessing many desirable properties not present in other entropy functionals. For a detailed discussion on the definition, implementation and estimation issues of S_ρ, see e.g., [16]. The measure has been proven to have impressive and robust power for characterizing nonlinear processes. In particular, it has been shown that tests based upon S_ρ have very good performances in terms of power and size (see [17]). In the binary case the double integral reduces to summation and probabilities $[Pr(\cdot)]$ are estimated through relative frequencies:

$$S_\rho(k) = \frac{1}{2}\sum_{i=0}^{1}\sum_{j=0}^{1}[\sqrt{Pr(X_t = i, Y_{t+k} = j)} - \sqrt{Pr(X_t = i)Pr(Y_{t+k} = j)}]^2.$$

In order to obtain appropriate confidence bands for $S_\rho(k)$, several issues have been considered: (i) the null hypothesis we test is that of independence between binary sequences, that is, the absence of an informational organization between codons; (ii) such test has to take into account the different proportion of bases across DNA sequences, i.e., the possible correlation found does not have to depend from the GC or other biologically meaningful contents; (iii) when comparing dichotomic classes the test does not have to depend on correlations induced by their definition; in fact, some specific combinations of dichotomic classes and reading frames induce nonzero spurious correlations even in random sequences. The above requirements can be satisfied by resorting to suitable nonparametric bootstrap or permutation schemes. The original DNA base sequence is randomly permuted. On this new sequence, the chemical (or dichotomic) classes are computed and the measure S_ρ is estimated. The procedure is repeated B times (say $B=5000$) as to obtain the bootstrap distribution of $S_\rho(k)$ under the null hypothesis. Clearly, each permutation of the original data preserves the original proportion of bases and this fulfils requirement (ii). Also, the computation of the measure S_ρ on two binary sequences obtained from the same random permutation of DNA bases automatically accounts for coding-induced correlations and fulfils point (iii).

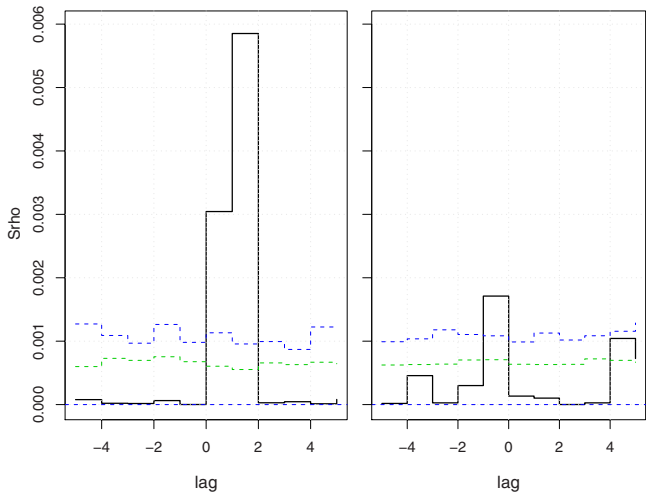

FIG. 4. (Color online) $S_\rho(k)$ computed on K-Am (3rd bases) and S-W (1st bases) (left), and on R-Y (3rd bases) and S-W (1st bases) (right). Dashed lines: confidence bands at levels 95% (green/light gray) and 99% (blue/dark gray).

V. RESULTS

In order to investigate whether the mathematical structure implied by the model finds a correspondence in real genetic sequences, we have analyzed the whole set of 88 different protein coding DNA sequences listed in [18]. We have computed the cross entropy measure on all the possible nontrivial combinations of chemical classes. For each sequence, we have 36 different cases (27 concerns comparison between two different chemical classes, see, e.g., Fig. 2).

A representative example of the results for the chemical classes is shown in Fig. 4. The figure can be interpreted similarly to crosscorrelograms: at a given lag k, $S_\rho(k)$ shows the dependence between X_t and Y_{t+k}. The dashed lines correspond to confidence bands at levels 95% (green/light gray) and 99% (blue/dark gray), obtained under the null hypothesis of independence as described before. We found very strong short range correlations (indeed, at lags −1, 0, and 1) for particular position configurations. In detail, we observe strong correlations for the R-Y dichotomy in the third position and the K-Am one in the second. These correlations are clearly related to the parity class [see Fig. 1(a)]; a further key result is that K-Am (3rd bases) and S-W (1st bases) classes present strong correlations at lags 0 and 1. This finding supports the hypothesis of the existence of the hidden class. Furthermore, consistently with the group structure and the antisymmetric properties described above, the former results suggest to choose the K-Am class for the last base of the previous codon in the definition of the hidden class. Hence the resulting algorithmic representation of the hidden class is shown in Fig. 5 and the associated matrix operator is

$$O_3 = \begin{pmatrix} 2 & 1 & 1 & 2 \\ 0 & 4 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

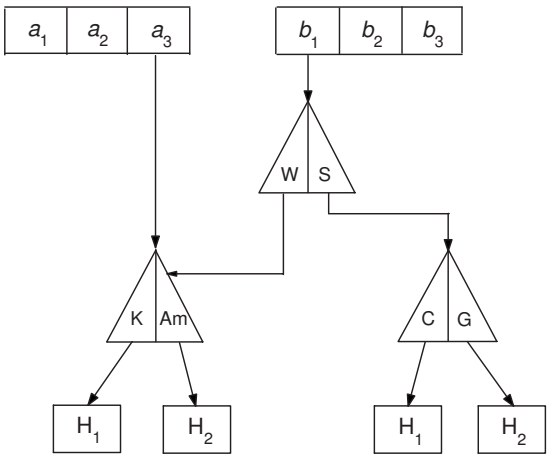

FIG. 5. Algorithmic representation of the coding of the hidden class; H_1 and H_2 denote the two categories of the class.

Observe that the existence of the hidden class implies some surprising consequences from both chemical and informational points of view. In particular, the genetic information contained in a codon is not simply that corresponding to the coded amino acid; indeed, there exists a complex structure that correlates the information content of a given codon with that of neighboring ones. Such informational properties seem to be associated with the possibility of error correction mechanisms as previously reported [3]. In fact, this hypothesis implies the existence of short range correlations between these newly defined dichotomic codon classes. Codon classes are clearly linked to chemical classes; however, as we have shown above, the mapping between them is nonlinear and involves two contiguous bases. Hence the correlation pattern of dichotomic codon classes is essentially different from that of chemical classes and provides a new way to gain insights on the informational structure of protein coding sequences. Thus as a further step in the statistical analysis, we have computed the cross entropy between the three codon classes (parity, Rumer, and hidden). As for chemical classes, we have computed the cross entropy measure on all the nontrivial combinations of dichotomic codon classes (by taking also into account reading frame shifts and complementary antisense computation). For each sequence we have 51 different cases (36 concerns comparison between two different codon classes).

A representative example is shown in Fig. 6 where we report $S_\rho(k)$ computed on the three dichotomic classes (all of them with frame shift 1): parity vs Rumer's (left), parity vs hidden (center), Rumer's vs hidden (right).

The results are again particularly informative. In fact, we found a strong correlation between the parity and the hidden classes at lag -1, frame shift 1 (see Fig. 6, center). We have observed such a remarkable correlation in 66 out of 88 sequences analyzed. Moreover, the dependence is much stronger (up to ten times) than that observed for chemical classes. In Table III we summarize the most interesting results pertaining to the whole set of combinations. The Table reads as follows: the second column indicates the combination of dichotomic codon classes where P=parity, R=Rumer, H =hidden; the shift with respect to the reading frame is indicated with a number: 0=no shift; 1=one base shift; 2=two base shift; the suffix "a" indicates the computation of the dichotomic class upon the complementary strand in reverse sense. The third column indicates the lag at which the cross entropy exceeds the confidence bands at 95%. The fourth column reports the frequency of exceedances. Finally, the last column shows the bases involved in the computation of the codon classes: the numbers 1, 2, 3 and 4, 5, 6 indicate the ordered bases of two contiguous codons. The overbar indicates the complementary base in the corresponding position. For instance, row number 4 refers to the correlation between the hidden class computed in the reading frame and the Rumer's class computed on the reverse complementary sequence with frame shift 1. In this case we observe 78 out of 88 sequences having a significant correlation at lag −1. The bases involved are the last of a codon and the first of the following one (hidden class) and the first and the second of the following one subject to the complementary transformation and reversed (Rumer's class).

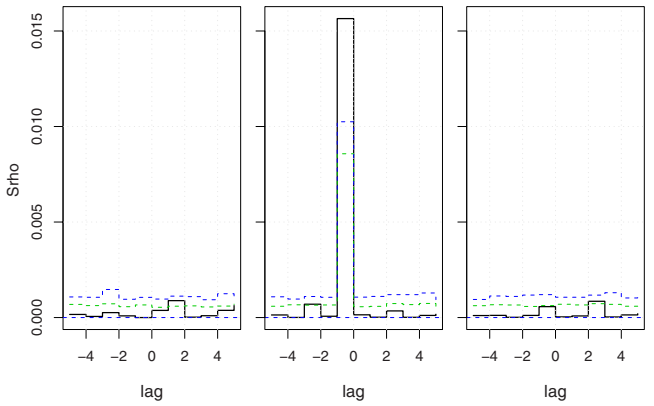

FIG. 6. (Color online) $S_\rho(k)$ computed on the three dichotomic classes (all of them with frame shift 1): parity vs Rumer's (left), parity vs hidden (center) see also Fig. 3, Rumer's vs hidden (right). Dashed lines: confidence bands at levels 95% (green/light gray) and 99% (blue/dark gray).

TABLE III. Summary of the most significant results regarding correlations between dichotomic codon classes (see text for the description).

#	Combination	Lag	Frequency	Bases
1	R1–R1a	0	49/88	$23-\overline{45}$
2	R1–R2a	0	78/88	$23-\overline{34}$
3	H1–P1	−1	66/88	45–34
4	H0–R1a	−1	78/88	$34-\overline{45}$
5	H0a–R1	0	84/88	$\overline{34}-23$
6	H0a–R2a	0	60/88	$\overline{34}-\overline{34}$
7	H1–R1a	−1	88/88	$12-\overline{12}$
8	H1a–R1	1	79/88	$\overline{23}-23$
9	H2–R1	0	82/88	23–23
10	H2a–R0a	1	58/88	$\overline{12}-\overline{23}$
11	H2a–R1a	0	84/88	$\overline{12}-\overline{12}$

The main points that emerge from Table III are (i) only in two combinations (3, 9) we observe significant correlations involving bases belonging to the sense strand. Both of them include the third base of a codon, the base related to redundancy in the amino-acid coding. This fact is compatible with our hypothesis on error correction [3]. However, differently from combination 9, combination 3 involves bases of two different codons. This correlation is particularly important because is not the consequence of either the coding scheme or the proportion of bases, as explained in the previous section. Hence, we can assert that there is a mechanism of short range information transmission along codons based upon the redundancy of amino acid coding. (ii) Significant correlations between the same dichotomic class computed out of its definitory position are observed only for the Rumer's class (see rows 1 and 2). (iii) Remarkably, all the remaining significant correlations involve only the hidden-Rumer combination and include a complementary antisense computation. In particular, we observe three combinations (5, 7, 11) with significant lags in more than the 95% of the sequences analyzed. Note that, combination 5 is analogous to combination 3 in that it involves bases of two different codons and include the wobble base in one of them. Moreover, combinations 7 and 11 involve the same couple of bases and exclude the wobble base. Thus we may conjecture that these last two correlations may be related to the protein structure.

VI. CONCLUSION

In this work, we have further developed a theoretical approach for the description of the genetic code [2,3] which has relevant consequences for the understanding of the organization of genetic information in coding sequences of DNA. On the basis of such theory the existence of a dichotomic codon class, i.e., the parity class, emerges in a natural way. This class has a clear mathematical meaning (the parity of binary strings of digits) and can be derived by means of a nonlinear algorithm acting on the two last bases of a codon. On the same ground we have shown that Rumer's class [5], i.e., the codon class describing the degeneracy of the genetic code, can be obtained by applying a similar algorithm to the first two bases of a codon. Both classes exhibit an antisymmetric behavior under the action of two specific global transformations of the bases. Since a third global transformation exists, i.e., the complementary transformation, it is reasonable to conjecture the existence of a third class, the hidden class, which is antisymmetric with respect to such transformation. Thus the dichotomic codon classes, together with their associated antisymmetric transformations, can be put in an elegant unified mathematical framework possessing the group property.

On the basis of this theoretical development, new insights about the informational structure of real sequences of protein coding DNA regions can be derived. In fact, with the aid of this new coding strategy, we can generate from DNA sequences binary sequences corresponding to the codon classes. In order to study the mutual dependence of such binary sequences we have implemented a metric-entropy measure S_ρ that can be interpreted as a nonlinear crosscorrelation function. The measure possesses many desirable properties not present in other entropy functionals [16,17]. By means of these methods, we have analyzed a set of 88 different protein coding DNA sequences. For each sequence we have computed the cross entropy measure on all the possible nontrivial combinations of both chemical classes and codon classes. The results of the analysis are summarized in Table III. The main points to be remarked regarding chemical classes are (i) we found very strong short range correlations for the R-Y dichotomy in the third position and the K-Am one in the second. These correlations are clearly related to the parity class [3]. (ii) Significant correlations are found also for the K-Am dichotomy (3rd bases) and the S-W one (1st bases). Consistently with the group structure and the antisymmetric properties described above, this result confirms the existence of the hidden class and emphasizes the K-Am role of the third base of a codon for its definition.

Regarding the statistical analysis of dichotomic codon classes, the main results are (i) we found remarkable short-range correlations one order of magnitude greater than those of the chemical classes. (ii) These correlations neither depend on the proportion of bases of the sequence, nor are a consequence of the coding algorithm. This is granted by the resampling methods employed. (iii) We observe significant correlations that involve bases of two different codons. This implies the existence of a mechanism of short range information transmission along the coding sequence. (iv) Those correlations emerging from antisense computations involve mainly the hidden-Rumer combination. The bases involved in one of these combinations exclude the wobble base and pertains to a single codon suggesting that these last correlations may be related to the protein structure. Remarkably, many of the correlations found seem to be universal: the same behavior is observed on almost all the protein coding sequences of the sample set (see Table III, rows 5, 7, 9, 11).

As previously suggested [3,4,9], the results support the idea that the genetic information is organized according to a nonlinear mathematical structure that allows error detection and correction. Notice that a dynamical system framework has been also proposed in a context related to the structure of the dynamics of chromatin in order to account for observed long range correlations [19]. This nonlinear dynamics paradigm is very appealing and deserves indeed a great deal of attention and further research. In this line of reasoning, the present work represents a further step in connecting a mathematical theoretical framework with the informational structure of DNA and mRNA molecules. One of the main aims is the uncovering of error detection and correction mechanisms based on nonlinear coding and decoding of genetic data.

ACKNOWLEDGMENTS

This work has been partially supported by MIUR 60% and by MIUR 40% funds.

[1] D. Holste, I. Grosse, S. Beirer, P. Schieg, and H. Herzel, Phys. Rev. E **67**, 061913 (2003).
[2] D. Gonzalez, Med. Sci. Monit. **10**, 11 (2004).
[3] D. Gonzalez, S. Giannerini, and R. Rosa, IEEE Eng. Med. Biol. Mag. **25**, 69 (2006).
[4] D. Gonzalez, in *The Codes of Life: The Rules of Macroevolution*, Biosemiotics, No. 1, edited by M. Barbieri and J. Hoffmeyer (Springer, Amsterdam, 2008), Chap. 8, pp. 111–152.
[5] Yu. B. Rumer, Dokl. Akad. Nauk SSSR **167**, 1393 (1966) (in Russian).
[6] G. Sella and D. Ardell, J. Mol. Evol. **63**, 297 (2006).
[7] T. Tlusty, J. Theor. Biol. **249**, 331 (2007).
[8] T. Tlusty, Phys. Biol. **5**, 016001 (2008).
[9] D. Gonzalez, in *The Codes of Life: The Rules of Macroevolution* (Ref. [4]), Chap. 17, pp. 379–394.
[10] José Eduardo M. Hornos and Yvone M. M. Hornos, Phys. Rev. Lett. **71**, 4401 (1993).
[11] B. Dragovich and A. Dragovich, SFIN A **20**, 179 (2007).
[12] L. Frappat, P. Sorba, and A. Sciarrino, Phys. Lett. A **250**, 214 (1998).
[13] L. Frappat, A. Sciarrino, and P. Sorba, J. Biol. Phys. **27**, 1 (2001).
[14] D. Dónall Mac Dónaill, Chem. Commun. (Cambridge) **18**, 2062 (2002).
[15] S. Wolfram, *A New Kind of Science* (Wolfram Media, Inc., Champaign, IL, 2002).
[16] C. W. J. Granger, E. Maasoumi, and J. Racine, J. Time Ser. Anal. **25**, 649 (2004).
[17] S. Giannerini, E. Maasoumi, and E. Bee Dagum, in *Bulletin of the International Statistical Institute, 56th Session* (ISI, 2007).
[18] A. Som, S. Chattopadhyay, J. Chakrabarti, and D. Bandyopadhyay, Phys. Rev. E **63**, 051908 (2001).
[19] S. Nicolay, F. Argoul, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, and A. Arneodo, Phys. Rev. Lett. **93**, 108101 (2004).