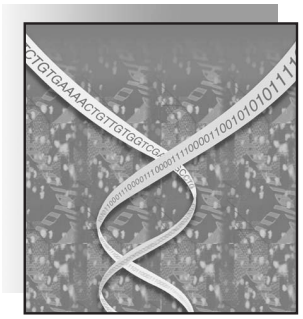


Detecting Structure in Parity Binary Sequences

Error Correction and Detection in DNA

BY DIEGO LUIS GONZALEZ, SIMONE GIANNERINI, AND RODOLFO ROSA

In this article, we investigate the possible existence of error-detection/correction mechanisms in the genetic machinery by means of a recently proposed coding strategy [11]. On this basis, we numerically code exons, creating binary parity strings and successively we study their dependence structure by means of rigorous statistical methods (moving block bootstrap, and a new entropy-based method). The results show that parity sequences display complex dependence patterns enforcing the hypothesis of the existence of deterministic error-correction mechanisms grounded on this particular parity coding.

Introduction

The rules governing the translation of RNA sequences into proteins were discovered some 40 years ago [1]. The universal biochemical translation table known as the genetic code connects two different chemical worlds: that of the nucleic acids with that of the biologically active proteins. The identification of this table with a code (a man-made system implementing symbolic representation for communication purposes) represents an early perception of the deep connection between biological information and coding theory. Coding theory is a research area intimately related to information and communication systems theory; for this reason, the introduction of tools from these related fields has represented a natural and inevitable step in the study of genetic information flow [2].

Information theory aims at giving a theoretical framework to processes characterized by some kind of information exchange. The main practical problem faced by information theory, and, in particular, by communication systems theory, is that of the transmission of reliable information through unreliable channels. To this aim, error detection and correction using appropriate coding and decoding methods represent a crucial step. All the methods of error-control coding are based on the adding of redundancy to the transmitted information. As the genetic information is redundant (evident, for example, in the case of short repeats inside introns), and since the genetic code is also redundant itself (intrinsic redundancy in the coding of amino acids), the possible existence of error-control mechanisms represents a somehow natural hypothesis [2]–[9] related to the biological task of ensuring a high degree of reliability in the transmission and expression of genetic information. In literature, such a hypothesis has been explored from a constructive point of view by proposing possible coding strategies that may be used in the organization of genetic information [6]–[8] and also from a statistical point of view, that is, by studying the dependence between symbols for revealing the existence of underlying coding mechanisms [5].

In this article we investigate this problem from a point of view that takes into account, to some extent, the two above mentioned approaches. The article is composed of two main parts. In the first part ("Parity Coding, Error Corrections, and DNA" and "A Mathematical Theory for the Genetic Code"), we present a new mathematical theory of the genetic code that leads to a coding strategy of codons and amino acids showing very interesting mathematical properties from the point of view of coding theory. This first part represents a constructivist part of the article in the sense that the natural coding we propose suggests the existence of error-control/correction mechanisms operating on the basis of its intrinsic mathematical structure.

In the second part ("Exploring the Structure of Binary Sequences" and "Results: Analysis of DNA Binary Sequences"), we study the statistical properties of real coding sequences with methods tailored for the study of binary sequences. The output of the aforementioned coding is a binary string obtained by sequentially appending 6-b words (binary words of 6 bits length) representing codons. As it will be shown below, these 6-b words possess well-defined parity properties; therefore, a simplified version of the coding can be obtained by replacing the 6-b words with 1-b words containing only the parity information (recall that the parity of the 6-b string can be defined as the parity of the summation of its symbols: an even number of ones leads to an even string, an odd number of ones to an odd string). The possible existence of an error-correction mechanism based on this parity coding is explored on an empirical basis. This hypothesis necessarily implies that the binary symbols exhibit some structure of dependence. The search for such structure of dependence by means of rigorous statistical methods is thus the scope of the second part of this article.

The sections are structured as follows: in the second section we discuss the goal of the article and where the work fits into the state of the art on coding theory applied to the study of biological information flow.

0739-5175/04/$20.00©2006IEEE

The new theory for the genetic code is based on a nonunivocal representation of whole numbers by means of the so-called nonpower binary bases.

In the third section, we describe in some detail the new mathematical theory of the genetic code [10], [11]; this theory is based on the representation of natural numbers by means of the so-called nonpower binary number representation systems. The theory describes the degeneracy of the genetic code and allows a 6-b binary characterization of codons. Moreover, hidden symmetry properties of the genetic code are highlighted. Also, the role of the parity of the 6-b representation of codons is shown. On this basis, we describe how to numerically code exons (protein coding regions of DNA) and obtain parity strings, i.e., strings formed by attaching a parity symbol to every codon in a given sequence. This parity coding represents the starting point for the statistical analysis performed in the following sections.

In the fourth section, we describe two advanced statistical methods for the study of the dependence structure in binary strings, i.e., the moving block bootstrap (MBB) [12] and an entropy-based dependence metric (the normalized Bhattacharya-Hellinger-Matusita distance) [13].

In the fifth section, we show the results of applying the methods described in section four to binary parity strings. These strings are obtained from the coding described in section three to the protein-coding sequences x80497 (phosphorylase kinase, *Homo sapiens*) and AF017114 (glycogen synthase mRNA, *Oryctolagus cuniculus*).

The possible significance of these statistical results is discussed in the last section. The results suggest that the intrinsic redundancy present in the genetic code, i.e., the existence of synonymous codons, can be used to encode additional information for error control and correction. Should this hypothesis be proven correct, the theoretical description of redundancy in the actual genetic code would represent a key point in elucidating the genetic mechanism acting on this basis. In the conclusion section, we report on implications and suggestions together with future research directions prompted by the results obtained by combining the novel mathematical theory for the genetic code and rigorous statistical methods.

Parity Coding, Error Correction, and DNA

We have mentioned the analogies between the genetic machinery and communication processes. Basically, a communication process is characterized by three main subprocesses: the coding of the information to be communicated, the transmission of the information along the communication channel, and the decoding of the information at the receiver. Usually, it is in the communication channel that unwanted errors are introduced. In man-made communication systems, the coding and decoding steps are tailored in such a way that detection and correction of the errors introduced in the communication channel can be achieved. Indeed, this is the main purpose of communication theory, i.e., to transmit reliable information through unreliable channels.

Different authors have modeled the genetic information flow in the framework of communication theory ([2], [6], [14], [15]). For example, in [16] a detailed view is given including transcription, translation initiation, and translation elongation in the decoding step, while replication is considered as the main process related to the transmission channel. Even if a thorough analysis in terms of communication theory is not the main aim of this article, we make use of some general features that every communication system, including the genetic one, must possess. From this point of view, we need to remark that it is very difficult to identify a coding step in the genetic system. The information is transmitted along different genetic processes (replication, transcription, and translation) as it is determined at its source, the double helix of DNA (following the central dogma of molecular biology, no additional information is produced in these steps); that is, the information arises already in coded format. Ignorance about these ab initio coding rules implies ignorance about the constraints imposed by these rules on the decoding mechanism.

Our point of view in this regard is a pragmatic one. We know that genetic information is coded; exons can be decoded following the rules of the genetic code (introns or intergenic regions also may convey biological information, but we do not know a general decoding rule). Moreover, we know that the genetic code is redundant; that is, a given amino acid may be decoded starting from more than one different codon. Thus, we have a decoding table—the genetic code—and a redundancy associated to this table, the main ingredients needed to implement an error-correction mechanism. Hence, the question naturally arises: Can an error-correction mechanism be implemented on this basis? To this regard, we need to distinguish between robustness to errors and error correction. The robustness to errors of the genetic code has long been recognized; a random error produced in a particular codon leads to the same amino acid or to some similar one from the point of view of physico-chemical properties. But this robustness is not due to an error-correction mechanism. It must be remarked that the natural robustness of the genetic code against errors does not impose any constraint in the redundancy distribution. Instead, an error-correction mechanism, implies the organization of the redundancy in a mathematically structured way (usually following the properties of finite groups). Thus, one of the crucial points for the existence of error-correction mechanisms is the existence of a mathematical structure in the coded data or, equivalently for the genetic case, a mathematical structure in the genetic code evidencing the mathematical structure in the data to be decoded.

The other crucial point concerns the redundancy: How is redundancy encoded in the genetic information? The existence of an error-correction mechanism automatically implies the

existence of dependencies between the symbols representing the information.

In this sense, our work is based on these two premises: the founding in the genetic code of a strong mathematical organization and the study of the dependence of the data produced by coding real data on the basis of this mathematical organization.

In literature, a few different mathematical models of the genetic code, mainly dealing with the description of the first level of degeneracy (the distribution of the number of synonymous codons), have been proposed [17]–[20]. Our approach is radically different because it is the unique model that describes exactly the first level of degeneracy of the genetic code and gives a deep insight into the second level of degeneracy (the association between specific codons and specific amino acids). Moreover, the approach reveals many surprising numeric and symmetry properties of the genetic code, as explained in the next section. To the authors' knowledge, this is the unique mathematical model based on a nonpower binary representation of natural numbers. A different model, proposed in [19], is based on number representations and on the number of nucleons on amino acids' side chains and uses the digital system and a modulus equivalence.

In literature, the structure of dependence has been studied mainly with the aim of identifying protein-coding regions, that is, discerning them from intronic and intergenic noncoding regions of DNA (see, for example, [21]). The main aim of this article is somehow different from existing approaches since, as previously stated, we want to investigate the existence of error-correction mechanisms suggested by the strong mathematical structure found in the genetic code. However, the results can be also interesting in relation to the study of long-range correlations in genetic data. In fact, as shown in the next section, the 6-b coding (and, consequently, also the parity coding) is not a fixed binary coding, that is, any of the four bases is not represented by a fixed 2-b number.

In fact, the four bases of DNA are usually coded by a two-digit binary number [22]. For example, we can assign to thymine/uracil (T/U) the binary string (0,0), to cytosine (C) (0,1), to adenine (A) (1,0), and to guanine (G) (1,1) (see also [23]). This assignation is necessarily arbitrary as there is no reason to assign to T the string (1,0) instead of (0,0) and so on. Some researchers assign such numbers taking into account the chemical properties of the base (such as the purine or pyrimidine character) sometimes reducing the binary dimension of the representation [21]. Also, in this case, the assignation is arbitrary since this only shifts the problem: Why assign 1 to purine and 0 to pyrimidine and not the reverse? or why not use the other possible partition of bases as keto and amino or strong and weak? For the coding of the four bases (T, C, A, G) in a triplet, there is a total of 24 possible fixed 2-b different choices or six different choices if the representation is 1-b (in fact robustness against fixed code choice has been tested for some alternatives regarding this last case [21]). Our approach, on the contrary, provides a natural strategy for the numerical coding of bases because it takes into account the degeneracy properties of the genetic code. Moreover, codons can be characterized by a parity bit in a nontrivial way, i.e., this characterization cannot be obtained with any of the fixed assignations mentioned above; the numerical assignation in the nonpower binary number representation is not fixed but context dependent. Because of the uniqueness of the numerical values of the basis describing the code degeneracy in the nonpower binary representation, we denote it for simplicity as the *genetic code like binary representation* (GCL binary representation).

The existence of correlations associated to this coding has not been investigated before. A comparative study between this and former approaches may give a deeper view about the origin of correlations in the genomic data.

In literature, the use of parity coding at the genetic level has been suggested in different contexts [3], [7], [8], [10], [11]. In the approach described in [7], it is shown that a parity coding is actually working at a chemical level for the selection of complementary bases in DNA. Such an approach is relevant in an evolutionary context in order to explain the actual use of complementary bases in modern DNA. Our approach suggests a similar coding acting along the double helix of DNA, where biologically meaningful information is encoded; of course, the same coding can be applied to both the DNA and the messenger RNA (mRNA) by simply swapping T and U.

Clearly, if some error-control mechanism is present, the redundant information cannot be stochastically independent; crucial information about the relative dependence of different symbols can be obtained by statistical methods. The GCL coding gives a natural parity assignation for codons; therefore, in the second part of the article, we study the dependence properties of binary parity sequences obtained from sequences of DNA or mRNA. This statistical information represents a necessary step in order to crack any eventual error-control mechanism based on this specific coding and may be useful also in order to gain understanding about the organization of the genetic information along nucleic acid molecules (for example, the existence of long-range correlations).

A Mathematical Theory for the Genetic Code

From a mathematical point of view, the new theory for the genetic code is based on a nonunivocal representation of whole numbers by means of the so-called nonpower binary bases. Contrary to the usual number representation systems, that is, systems in which the positional values grow as the powers of some base (for example, the powers of 10 in the usual decimal power representation), nonpower refers to the fact that the positional values of the representation system grow more slowly than the power of some basis (the powers of 2 in the nonpower binary systems). It is shown that there exists a unique nonpower set of positional bases (called a *genetic code like nonpower binary representation system*) capable of explaining all the logical properties associated with the degeneracy of the genetic code viewed as a generic correspondence or mapping between two sets with a different number of elements (the 64 possible codons formed by all the combinations of four letters, T/U, C, A, G, and the 20 amino acids plus the stop signal).

Table 1. Degeneracy distribution for the standard genetic code.

Degeneracy	Amino Acids (#)
6	3
4	5
3	2
2	9
1	2

If some error-control mechanism is present, the redundant information cannot be stochastically independent.

Due to the redundancy of the code, some elements of the codon set are necessarily mapped to the same element of the amino acids set. In Table 1 we show the actual distribution of the codons that codify the same amino acid in the standard genetic code. Such a table represents the first level of degeneracy.

Besides the description that refers to this first level of degeneracy, called the *degeneracy distribution,* it is important to account for the distribution of codons and amino acids inside the degeneracy distribution since each possesses a precise physicochemical and biological identity. The nonpower representation theory of the genetic code also provides a deep insight into this second level of degeneracy, that is, the specific codon and amino acid assignation inside the given degeneracy distribution. The theoretical representation proposed here is also able to disclose the presence of a hidden symmetry of the genetic code, the so-called palindromic symmetry. This symmetry is related to the existence of degeneracy-preserving transformation rules which associate amino acids in pairs (pairs of palindromic amino acids). Moreover, numerical elements in the nonpower representation can be associated with biochemical elements in the genetic code. This association uncovers another hidden property of the genetic code: individual codons are codified in parity, i.e., a parity bit can be assigned to every codon. This last property is very appealing since it can be related to the possible existence of error-correction mechanisms in the genetic machinery. Parity coding is one of the simplest and most widely used strategies for error control and correction in man made digital communication systems.

From the point of view of set theory, the genetic code is a correspondence or mapping between two sets of different cardinality: the 64 codons formed by all the possible combinations of the four bases, U, C, A, G, and the 20 amino acids plus the stop signal. The correspondence defines the starting and arriving sets; that is, an arrow points from a given element of the codons set (starting set) to a corresponding element in the amino acids set (arriving set). The direction of the arrows is compatible with the central dogma of molecular biology; that is, the genetic information flow is only from nucleic acids to proteins. The different cardinality of the starting and arriving sets implies the redundancy and degeneracy properties of the code. Our aim is to build a mathematical structure that possesses the degeneracy properties of the genetic code from a logical point of view. In other words, we aim at creating a structural isomorphism.

Table 2. Degeneracy distribution inside quartets for the Euplotid nuclear variant of the genetic code.

Degeneracy	Amino Acids (#)
4	8
3	2
2	12
1	2

First, we define the properties of the genetic code from the point of view of set theory: the code is a surjective, noninjective correspondence (that is, not one-to-many) between two sets of different cardinality. The surjective property assures that no elements in the arriving set are vacant (all amino acids and the stop signal are coded by at least one codon). The noninjective property refers to the fact that some elements in the arriving set are represented by more than one element in the starting set (some amino acids are represented by more than one codon). This property viewed from the point of view of the arriving set is called *degeneracy*: a given amino acid does not uniquely specify the codon that originated it. From a mathematical point of view, this implies that the correspondence is not invertible. In biological terms, this affirmation is equivalent to saying that, given a coding sequence, we can know the corresponding sequence of amino acids defining a particular protein; but given a particular protein, we do not know the specific sequence of codons that codify it at the mRNA or DNA levels. The noninjective property viewed from the starting set point of view implies the concept of redundancy: different elements of the starting set codify for the same element in the arriving set (different codons codify the same amino acid). The redundant elements are called in this biological case *synonymous codons*. The "not one-to-many" property means that a specific codon cannot codify for more than one amino acid; this is a true statement for a given variant of the code and excludes possible contex-dependent translation oddities.

Theoretically, the properties we have just defined can identify infinite correspondences between sets of different cardinality. We must now define a correspondence that takes into account the actual cardinal numbers of the genetic code. For this purpose, we first need to define *degeneracy distribution*, that is, a table (see Table 1) where we report all the degeneracy values actually found in the code (left) along with the corresponding number of amino acids that share such a degeneracy (right).

For different reasons (for example, the symmetry properties of the code or some oddities in the characteristics of the degeneracy-6 group of amino acids), different authors (see [11] and references therein) have considered the degeneracy distribution inside quartets of the genetic code, that is, inside groups of four codons sharing their first two letters (for example, the quartet UGU, UGC, UGA, UGG). From a mathematical point of view, this is equivalent to enlarging the arriving set of 21 elements to 24 elements as follows: the degeneracy-6 amino acids are split into two elements, the first represented

by four codons and the second by two of them. In this way the 20 amino acids plus the stop signal are represented by 24 different sets. For reasons that will become clear below, this is the representation which we aim at describing here. The degeneracy distribution obtained from the genetic code correspondence between the two sets of 64 and 24 elements is shown in Table 2 for the euplotid nuclear variant of the code, which differs from the standard one only in the assignation of the TGA codon (cysteine instead of stop). Now, we describe the mathematical model that enables us to build a structural isomorphism having the same mathematical structure as the genetic code. The theoretical framework is that of number theory and, in particular, the theory of integer number representation. Usual power positional representation systems are based on an additive process in which the powers of a given base are multiplied for the positional digits and added together in order to obtain a given integer. In the decimal representation system the powers of 10 are used. The digits can range from 0 to $n-1$, in this case, 0 to 9. For instance, the number 735 can be obtained as $75 = 5 \times 10^0 + 3 \times 10^1 + 7 \times 10^2$.

The fact that the digits are limited to the value of $n-1$, ensures the one-to-one character of the representation: a number is represented by only one combination of digits and vice versa. However, we are interested in redundant representation systems; redundancy can be obtained in two ways, allowing for the digits to go over their range or decreasing to some extent the values of the positional numbers (the ordered powers of the given base for power representation systems). We use this second possibility starting with the power representation system with the lowest integer base, i.e., 2, which defines the binary positional system. Thus in our system we assign to the positional numbers different values that grow more slowly than the powers of two. In particular, by taking the following set of positional values: [1 1 2 4 7 8], we can reproduce exactly the degeneracy distribution of the genetic code presented in Table 2 (for details see [11]).

At this point we have found a structural isomorphism between two correspondences: on one side we have codons in the starting set and amino acids in the arriving set, and on the other side, we have six-digit binary strings in the starting set and 24 whole numbers in the arriving one. The scheme describes perfectly the degeneracy properties of the genetic code. Nevertheless, it is well known that the identity of biological elements matters: arbitrary permutations of codons inside a given degeneracy distribution are not equivalent from a biological point of view. Surprisingly though, on the basis of the analysis of the symmetry properties on both sides of the structural isomorphism, we can relate in a natural fashion numerical elements of the no-power representation to biochemical elements of the genetic code. In this way, codons are mapped into 6-b binary strings, just like amino acids are mapped into integer numbers. The details of this mapping have been reported in [11] and are summarized in Table 3. The table shows the representation of the first 24 whole numbers in the GCL nonpower binary system. Each whole number is represented by a set of length-6 binary strings, for example, the number 6 is represented by the strings (001011) and (001100). In fact, $0*8+0*7+1*4+0*2+1*1+1*1 = 0*8+0*7+1*4+1*2+0*1+0*1 = 6$. The amino acid corresponding to a given whole number is shown in the central columns of the table. Observe that the number of binary

Table 3. Representation of the first 24 whole numbers (outer columns) in the GCL nonpower representation (1 1 2 4 7 8) (length-6 binary strings, horizontal rows). The degeneracy number (the 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). The color (either light gray or dark gray) indicates the parity of each string (odd and even, respectively).

#	8	7	4	2	1	1	Degen. #	Pairs of Palindromic		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	2	F Phe	S Ser 2							1	1	1	1	1	0	1	1	1	1	0	1	22
2	0	0	0	1	0	0	2	Ter	K Lys							1	1	1	1	0	0	1	1	1	0	1	1	21
3	0	0	0	1	1	0	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	2	L Leu 2	R Arg 2							1	1	1	0	0	0	1	1	1	0	1	1	19
5	0	0	1	0	1	0	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	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	3	C Cys	I Ile							1	1	0	0	1	0	1	1	0	0	0	1	16
8	1	0	0	0	0	0	4	S Ser 4	T Thr	1	1	0	0	0	0	1	0	1	1	1	0	0	1	1	1	1	1	15
9	1	0	0	0	1	0	4	P Pro	A Ala	1	0	1	1	0	0	1	0	1	0	1	1	0	1	1	1	1	0	14
10	1	0	0	1	0	0	4	V Val	G Gly	0	1	1	1	0	0	1	0	1	0	0	1	0	1	1	0	1	1	13
11	1	0	0	1	1	0	4	L Leu 4	R Arg 4	1	0	1	0	0	0	1	0	0	1	1	1	0	1	1	0	0	1	12

strings representing a given whole number corresponds to the degeneracy of the assigned amino acid describing exactly the degeneracy distribution of the genetic code.

Two major consequences arising from this approach are in order: 1) the uncovering of a hidden symmetry inside the genetic code, i.e., the palindromic symmetry, and 2) the natural classification of codons in definite parity classes. In Table 4, we show graphically the palindromic symmetry that maps quartets, preserving the degeneracy distribution; in Table 5, we show the parity distribution of codons. Every codon is assigned a parity bit which corresponds to the parity of the length-6 binary string shown in Table 3. Note that, as remarked above, the parity of the binary string can be computed by summating its symbols: an even number of ones leads to an even string and an odd number of ones to an odd string. We can observe that the parity bit can be also derived easily on the biochemical side by means of two complementary rules applied to codons: if the codon ends with a purine base (A or G), the parity is determined by such base (that is, an A defines an odd codon, and a G defines an even codon); if instead, the codon ends with a pyrimidine base (U or C), the parity bit is determined by the second letter of the codon (i.e., C or A in the second position give an even codon, and U or G determines an odd codon).

As we have remarked above, in this article we focus mainly on the parity coding because of its connection with hypothetical error-control mechanisms. Hence, in the next section we try to fathom this hypothesis by analyzing the dependence structure of parity sequences obtained through the application of the nonpower parity bit assignation as shown in Table 5.

Table 4. Graphical representation of the palindromic symmetry. All the quartets—defined by the same two first letters (bases) of codons—are associated in pairs by the palindromic transformation. Arrows of the same color indicate a common operation at the triplet level.

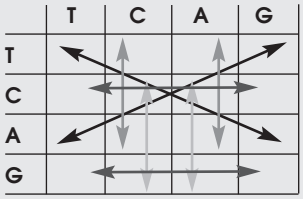


Table 5. Graphical representation of the parity classification of triplets. Light gray boxes indicate odd triplets, dark gray boxes indicate even triplets. The parity of a codon corresponds to the parity of the length-6 binary string that represents it in the GCL nonpower representation (see Table 3).

	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

Exploring the Structure of Binary Sequences

In the bioinformatics literature, the dependence structure of DNA sequences has been investigated in several studies, even though a rigorous statistical approach is not always followed. For an excellent statistics-oriented review on the topic see [24].

In order to investigate the dependence structure of DNA parity sequences, we exploit statistical methods that are appropriate for the analysis of dependent data. In the next section, we will discuss and motivate the use of such methods in our context. In particular, we will give a reliable estimate of the standard error and related confidence intervals for the proportion p of zeros in the sequences. Moreover, as shown above, we are able to assess quantitatively the presence of possible dependencies in the data. In order to investigate further such dependence, we will introduce a metric based on entropy as a relevant tool for characterizing the structure of DNA sequences.

Bootstrap Methods for Dependent Data

In this section, we outline a brief sketch about bootstrap methods. Such techniques introduced by Efron in 1979 [25] and described more fully in [26] are intensive computational procedures based on resampling from the observed data. Let the observed sample $\mathbf{x}_n = (x_1, x_2 \ldots, x_n)$ be a realization of random vector $\mathbf{X}_n = (X_1, X_2 \ldots, X_n)$ having an unknown underlying distribution function F. Let θ be the unknown parameter of interest (for example, the mean, median, correlation coefficient, etc.), $\theta(\mathbf{X}_n)$ be an estimator of θ, and $\theta(\mathbf{x}_n)$ be an estimate based on the observed sample $\mathbf{x}_n$. The bootstrap gives a somewhat "automatic" nonparametric method for providing an approximation of the unknown distribution of $\theta(\mathbf{X}_n)$, in particular for estimating its standard error, denoted with $\sigma(\theta)$. In the following, we will denote the estimate of $\sigma(\theta)$ with $\hat{\sigma}(\theta)$ or simply σ. Notice that even though F is known, assessing the accuracy of an estimate is often a difficult task, except for rather simple cases. The basic idea of the bootstrap is to resample the original data $\mathbf{x}_n$ and make inference from the resamples. All this requires the following steps: 1) estimate F by $\hat{F}$, the empirical distribution function, obtained by putting probability mass $1/n$ on each x_n; 2) generate a bootstrap sample $\mathbf{x}_n^* = (x_1^*, x_2^* \ldots, x_n^*)$ from $\hat{F}$ by making independent random draws with replacement from the data; 3) compute the bootstrap replication $\theta^* = \theta^*(\mathbf{x}_n)$, that is, the value of the statistics pertaining to the bootstrap sample $\mathbf{x}_n^*$; and 4) repeat the second and third steps B times to obtain B

bootstrap replications whose distribution approximates the distribution of $\theta(\mathbf{X}_n)$. The estimate of the standard error $\sigma(\theta)$ is approximated by the bootstrap standard error $\hat{\sigma}(\theta^*)$, or simply $\hat{\sigma}^*$, given by:

$$\hat{\sigma}^* = \left[\sum_{b=1}^{B} \frac{(\hat{\theta}_b^* - \bar{\theta}_b^*)^2}{B-1} \right]^{1/2}, \tag{1}$$

where

$$\bar{\theta}^* = \sum_{b=1}^{B} \frac{\hat{\theta}_b^*}{B}.$$

If the observations are correlated, that is, they can no longer be considered realizations of mutually independent random variables with the same distribution function F, the bootstrap is not applicable in the form outlined above, since the dependence structure of the data is disregarded. So, the "classical" bootstrap for independent and identically distributed (IID) variables, call it an *IID-bootstrap*, must be replaced by the moving block bootstrap (MBB), which resamples not individual observations but *blocks* of observations. The MBB allows one to assign measures of accuracy to statistical estimates for dependent observations in the form of finite time series. This problem is discussed also in [27], here we recall summarily some basic points.

Consider the stationary time series $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$. Let $\mu, \sigma_0^2, \gamma_k$ and $\rho_k\, k = (1, \ldots, n-1)$ be the mean, variance, covariance, and autocorrelation function of $\mathbf{X}_n$, respectively. Note that $\gamma_0 = \sigma_0^2$ and $\rho_k = \gamma_k/\gamma_0$ The variance of the estimator $\bar{X}_n$ of μ, is given by

$$\sigma^2 = Var[\bar{X}_n] = \frac{\sigma_0^2}{n} + 2\sum_{k=1}^{n-1} \frac{n-k}{n^2}\gamma_k$$

$$= \frac{\sigma_0^2}{n}\left[1 + 2\sum_{k=1}^{n=1} \frac{n-k}{n}\rho_k \right] \tag{2}$$

Since it will be useful in the following, we recall that in statistical mechanics the variance of $\mathbf{X}_n$ is usually written as (see, e.g., [28])

$$\sigma^2 = \sigma_0^2\left(1 + \frac{2\tau}{\delta t}\right), \tag{3}$$

where τ is the integrated correlation time

$$\tau = \int_0^{\infty} \rho(t)dt,$$

and δt is the time interval between two successive observations. We will show through the MBB how it is possible to estimate σ^2 directly and derive from it an estimate for τ through (3). For a comparison of various approaches for estimating σ^2, one can see also [29].

In order to estimate σ^2, the MBB considers in a chain of n observations all possible contiguous blocks of length l such that observations more than l apart are nearly statistically independent. In this way $q = n - l + 1$ "moving blocks" are obtained $(\mathbf{Q}_1, \ldots, \mathbf{Q}_q)$, where the ith block $\mathbf{Q}_i$ with starting point X_i contains l observations: $\mathbf{Q}_i = (X_i, X_{i+1} \ldots, X_{i+l-1})$. From these q blocks $\mathbf{Q}_i (i = 1, \ldots, q)$, we draw at random with replacement h blocks, with $h \times l = n$. The h selected blocks, placed one after the other, form the new sequence $\mathbf{Q}^* = (\mathbf{Q}_1^*, \ldots, \mathbf{Q}_h^*)$. Analogously to the IID-bootstrap, we can form a suitable number of MBB replications $\mathbf{Q}^*$ from each of which the statistic of interest is computed, and the bootstrap estimate of the standard error $\hat{\sigma}^*$ is derived through (1). The idea of the MBB is due to [30] and was studied theoretically in [12] and [31]. In practice, by varying l, one sees that when l is small the MBB estimate $\hat{\sigma}^*$ is close to σ_0 because the scheme does not manage to reproduce the correlation structure at lag $> l$, which is present in the original data. With increasing l, the data belonging to different blocks become more and more independent of one another until the blocks are actually IID random variables under the MBB scheme, and at the same time, inside each block the correlation is retained. In the presence of a positive (negative) correlation in the series, the plot of $\hat{\sigma}^*$ *vs* l shows an increase (decrease) of $\hat{\sigma}^*$ [see (2)] up to a region, call it a plateau, in which the variations are less pronounced (see Figure 1 below for the application to DNA series). The reaching of the plateau indicates: 1) a suitable choice for l, 2) the MBB estimate $\hat{\sigma}^*$ of σ, and 3) the "strength" of the correlation, as derived from (3).

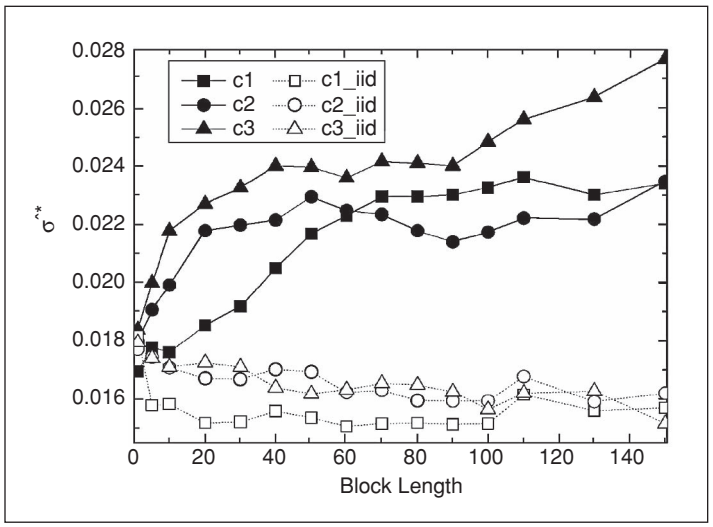


Fig. 1. Moving block bootstrap estimates $\hat{\sigma}^*$ of the standard error of p as a function of the block length l, computed for the sequence AF017114 (coding region) ($n = 736$, filled symbols): $c1$: in frame; $c2$, $c3$: out of frame. Results pertaining to IID binomial sequences having the same proportion p as the original data are reported in empty symbols ($c1\_iid$, $c2\_iid$, $c3\_iid$).

A Dependence Metric Based on Entropy

In the literature, there are many proposals of dependence measures, each of them motivated by different needs and built to characterize a specific aspect of the process under study. An important class of such measures is based on entropy functionals developed within information theory (see, for example, [32] and the references therein). For

instance, Shannon mutual information has spread widely in the context of nonlinear dynamics [33] as well as time series analysis [34]. However, none of these entropies define a metric since either they do not obey the triangular inequality or they are not commutative operators. Also, there have been recent studies in the statistics community with the aim of describing the properties that an ideal measure of dependence should possess (see, for example, [13] and the references therein). For these reasons, we have adopted the metric entropy measure $S_\rho(k)$, a normalized version of the Bhattacharya- Hellinger-Matusita distance, defined as follows:

$$S_\rho(k) = \frac{1}{2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(\sqrt{f_{(X_t,X_{t+k})}(x_1,x_2)} - \sqrt{f_{X_t}(x_1)f_{X_{t+\kappa}}(x_2)}\right)^2 dx_1 dx_2, \quad (4)$$

where $f_{X_t}(\cdot)$ and $f_{(X_t,X_t+k}(\cdot,\cdot)$ denote the probability density function of X_t and of the vector (X_t, X_{t+k}), respectively. The measure is in precise relation to other entropy functionals, such as Shannon entropy and Kullback-Leibler divergence, and can be interpreted as a nonlinear autocorrelation function. $S_\rho(k)$ satisfies many desirable properties: 1) it is a metric and is defined for both continuous and discrete variables; 2) it is normalized and takes the value 0 if X_t and X_{t+k} are independent and takes the value of 1 if there is a measurable exact (nonlinear) relationship between the variables; 3) it reduces to the linear autocorrelation function in the case of Gaussian variables; and, notably, 4) it is invariant with respect to continuous, strictly increasing transformations. Among other things, [13] addresses the issues of nonparametric kernel estimation of $S_\rho(k)$ and of its utilization in the context of hypothesis testing of serial dependence. The measure has been proven to have impressive and robust power for characterizing nonlinear processes. In the case of binary series the measure becomes

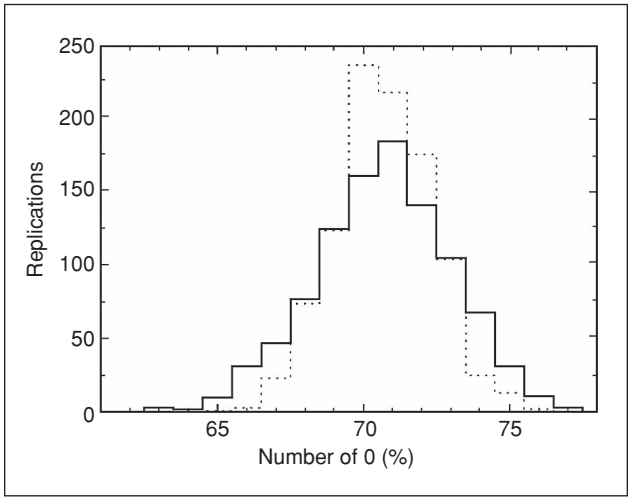


Fig. 2. Histograms of 1,000 moving block bootstrap replications of the proportion p for the $c1$ codon sequence AF017114 (coding region) (continuous line) and for IID binomial sequences (dotted line). The block length has been set to $l = 80$.

$$S_\rho(k) = \frac{1}{2}\sum_{i=0}^{1}\sum_{j=0}^{1}\left(\sqrt{\Pr\{X_t = i, X_{t+k} = j\}} - \sqrt{\Pr\{X_t = i\}\Pr\{X_{t+k} = j\}}\right)^2. \quad (5)$$

Here, the probabilities have been estimated in a nonparametric fashion by means of relative frequencies, and the confidence bands at the 95% level under the null hypothesis of independence have been obtained through Monte Carlo simulation.

Results: Analysis of DNA Binary Sequences

In the following we show the results obtained from the application of the methods described in the previous section to several DNA parity sequences, derived from both codons and anticodons, each of them considered both in-frame and out-frame. Here, by anticodon we mean the complementary triplet in the Watson-Crick sense. Since parity is defined by the second and/or the third letter in the codon, the parity sequence associated to the anticodons, which are read in reverse order, carries completely different information. For this reason, we have chosen to analyze the anticodon sequences also. Moreover, as redundant information can be codified along the sequences in unknown ways, it is also interesting to study the out of frame versions of both codon and anticodon sequences. Hence, from each codon we obtain and analyze six sequences. In the following *c1* (frame) and *c2* (+1), *c3* (−1) (out-frame) refer to the codon sequence, whereas *a1* (frame) and *a2* (+1), *a3* (−1)(out-frame) refer to the anticodon sequence. Notice that the length of the sequence n refers to the codon sequence so that it has to be multiplied by three to obtain the length of the base sequence.

As remarked above, the studies where DNA sequences are rendered dichotomous rely on somehow arbitrary choices for the mechanism of dichotomization, without an underlying model. In our case, such a procedure is encompassed naturally within the approach presented in the previous sections so that we expect the results to be well informative. In the following, we will always refer to the protein-coding part of the DNA sequences. Also, in this article, we will mainly concentrate on the statistical aspects arising from the analysis of parity sequences. Further investigations, including a comparison with noncoding portions, are in progress and will be reported in a future work.

The Moving Block Bootstrap

First, we apply the MBB in order to 1) obtain confidence intervals for the proportion p of zeros in the sequence and 2) investigate the dependence in the sequence. Recall that, for independent data, i.e., realizations of IID binomial variables, the standard error of the estimator of p is given by $\sqrt{\hat{p}(1-\hat{p})/n}$, where $\hat{p}$ denotes an estimate of p. In this instance, both the IID-bootstrap and the MBB give the same results. However, if there is some form of dependence, the MBB is able to reveal it and, at the same time, estimate the "true" standard error σ.

Figure 1 (filled symbols) shows the behavior of the MBB estimates $\hat{\sigma}^*$ of the standard errors of the estimator of p as a function of the block length l for the parity sequence AF017114 [35] (coding region), codon in frame

(*c1*) and out of frame (*c2* (+1) , *c3* (−1)). We also report the results obtained by applying the MBB on binomial IID sequences having the same proportions p of zeros as the observed data (empty symbols). The bootstrap replications B are 1,000 in all cases.

If the data were independent there would be no statistical difference in the standard errors estimates between observed and IID sequences. The values of $\hat{p}$ are 0.702,0.603,0.448 for *c1*, *c2*, and *c3*, respectively. The values of $\hat{\sigma}_0$ are 0.017 (*c1*) and 0.018 (*c2*, *c3*). At first sight, there is a clear difference between the results from the observed and the IID sequences, revealing the presence of a kind of dependence in the data. For the IID sequences, in fact, $\hat{\sigma}^*$ always remains close to $\hat{\sigma}_0$ while significant increases of $\hat{\sigma}^*$ are displayed for the observed sequences. Let us follow the trend referring, as an example, to *c1*. At the beginning $\hat{\sigma}^*$ is very close to $\hat{\sigma}_0$ as expected. With increasing l, $\hat{\sigma}^*$ grows. After $l \approx 70$, $\hat{\sigma}^*$ reaches a plateau. On the plateau, the actual dependence structure of data is captured, and the value found for $\hat{\sigma}^*$ may be retained as an estimate for the standard error σ. Here, it results $\hat{\sigma}^* = 0.023$ By replacing $\hat{\sigma}_0$ and $\hat{\sigma}^*$ in (3), it follows that the integrated correlation time is $\hat{\tau} = 0.41$. Similar trends hold also for *c2* and *c3*, for which $\hat{\tau}$ values are 0.25 and 0.39, respectively. It has been proved [12], [31] that $\hat{\sigma}^*$ is a consistent estimator of σ if l grows to infinity with n, provided that $l/n \to 0$**.** In practice, as shown in [29], the MBB enables one to assign accuracy even though the number of blocks is rather small, say, $n/l \approx 10$, so that the last points of Figure 1, corresponding to 4-6 blocks, are not reliable.

As remarked above, bootstrap methods can assess more than standard errors. For instance, we report in Figure 2 the bootstrap distributions of the estimator of p of the proportion of zeros with $B = 1,000$ for the IID-sequence (dotted line) and the observed sequence (continuous line) taken at $l = 80$ for the *c1* sequence. As expected from the central limit theorem, both distributions are Gaussian with the same mean, but the difference in the variance is clearly visible. Such a difference can be assessed easily through a test on the variances that results significant. The trends of Figures 1 and 3 show clearly that the difference between IID and MBB is due to the autocorrelation of the sequence [the second term between square brackets in (2)]. Notice that, in general, one can build confidence intervals from bootstrap distributions without having to make normal theory assumptions (see [26] for a complete discussion on this point).

We have tested several DNA sequences. In most cases, the trends of $\hat{\sigma}^*$ *vs* l are qualitatively similar to those reported in Figure 1, that is, a rise of $\hat{\sigma}^*$ as l increases. However, some sequences reveal no dependence; that is, $\hat{\sigma}^*$ remain always close to $\hat{\sigma}_0$, while other sequences display a decrease of $\hat{\sigma}^*$ as l increases, as shown in Figure 3 for the anticodon *a1* of the AF017114 sequence (coding region). The decrease of $\hat{\sigma}^*$ up to the plateau around $l \approx 80 - 100$ indicates that in this instance the correlation is negative. In analogy with Figure 2, we report in Figure 4 the MBB distribution of the estimators of p ($l = 90$). In this instance, the MBB distribution has a variability that is smaller than that of IID case, so the MBB confidence interval will be more accurate.

In Table 6 we summarize the results obtained by the MBB for the sequence AF017114 (coding region). In the columns are reported the sequence name, the estimate $\hat{p}$, $\hat{\sigma}_0$, $\hat{\sigma}^*$ and 95% confidence interval for p, obtained under the IID binomial hypothesis ($[\text{C.I. } 95]_0$) and, correctly, through the MBB [C.I. 95]*. It is important to note that even though in this instance the confidence intervals under the IID assumption do not differ markedly from the MBB intervals, the latter approach is the correct one for assessing the accuracy in the presence of dependent data without making distributional assumptions. For instance, assume we wish to test the hypothesis that the proportion p is the same for *c*1 and *c*2 at the 99% significance level. On the basis of an IID confidence interval, one would erroneously reject such a hypothesis. On the contrary, the hypothesis is not rejected if the MBB is employed.

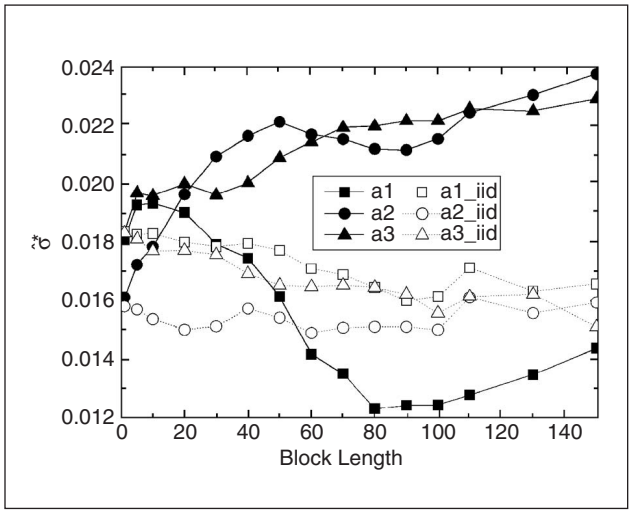


Fig. 3. Moving block bootstrap estimates $\hat{\sigma}^*$ of the standard error of p as a function of the block length l, computed for the anticodon sequence AF017114 (coding region) ($n = 736$, filled symbols): *a*1: in frame; *a*2, *a*3: out of frame. Results pertaining to IID binomial sequences having the same proportion p as the original data are reported in empty symbols (*a*1_iid, *a*2_iid, *a*3_iid).

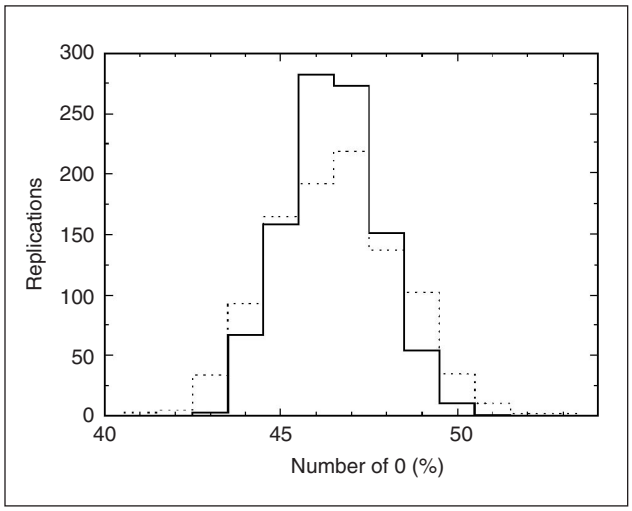


Fig. 4. Histograms of 1,000 moving block bootstrap replications of the proportion p for the *a*1 codon sequence AF017114 (coding region) (continuous line) and for IID binomial sequences (dotted line). The block length has been set to $l = 80$.

Table 6. Summary statistics and confidence intervals ((C.I. 95)) for the sequence AF017114 obtained under the IID binomial hypothesis ($_0$) and through the MBB (*).

Seq. AF017114	$\hat{p}$	$\hat{\sigma}_0$	$\hat{\sigma}^*$	$(C.I.\ 95)_0$	$(C.I.\ 95)^*$
c1 (0)	0.702	0.017	0.023	(0.670, 0.736)	(0.654, 0.744)
c2 (+1)	0.602	0.018	0.022	(0.566, 0.637)	(0.557, 0.643)
c3 (−1)	0.448	0.018	0.024	(0.409, 0.483)	(0.401, 0.492)
a1 (0)	0.470	0.018	0.012	(0.435, 0.505)	(0.446, 0.494)
a2 (+1)	0.739	0.016	0.021	(0.706, 0.770)	(0.697, 0.781)
a3 (−1)	0.458	0.018	0.022	(0.423, 0.491)	(0.417, 0.503)

The Entropy-Based Dependence Metric

The plot of Figure 5 shows the entropy-based dependence metric $S_\rho(k)$ versus the lag $k = 1, \ldots, 300$ computed on the coding sequence of the gene X80497 [36], $n = 1,236$ codons, in frame. We recall that the measure can be interpreted as a nonlinear autocorrelation function; that is, if $S_\rho(k)$ exceeds the confidence band at lag k, then there is a significant correlation between symbols that are distant k steps in the sequence. As in the previous section, by significant we mean consistently different from IID processes. Here and in the following, in order to obtain good estimates for $S_\rho(k)$, we chose a number of lags which is approximately one quarter of the length of the sequence as is well known in time series analysis. Two remarkable aspects emerge clearly from the inspection of the figure. First, the appearance of a kind of long-range dependence, starting from about lag 100. Second, the presence of several peaks that extend well over the confidence band, indicating the possible presence of periodicities. The strength of the peaks and the distribution of the distances between them seem to conform to a kind of regularity that may be gene specific, as also shown in a recent study on human chromosomes [37]. The long memory content seems to be less pronounced in out-frame sequences, as shown in Figure 6. Notice, however, that the peaks show a significant dependent structure also in this instance.

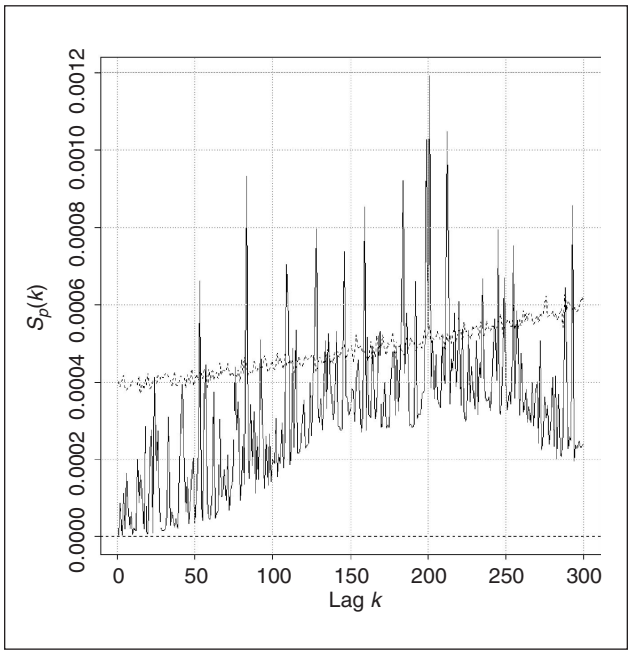


Fig. 5. $S_\rho(k)$, $k = 1, \ldots, 300$ for the codon sequence X80497 in frame $c1$, $n = 1,236$. The confidence band at the 95% level (dashed line) was obtained through Monte Carlo generation of 3,000 Bernoulli IID sequences.

In Figures 7 and 8, we show the results of the computation of $S_\rho(k)$ upon the anticodon sequence X80497 in frame and out frame, respectively. In this instance, the situation seems somehow reverted if compared with the results for the codon. In fact, the long-range dependence here is more evident for out-frame than for in-frame sequences. Notice also the large peak at lag 1 for the first of the two out-frame sequences [Figure 8)]. In any case, the anticodon parity series also shows a significant dependence structure at several lags that cannot be attributed to statistical fluctuations.

The findings reported above show clearly that the entropy-based metric has been able to disclose the existence of a nontrivial dependence structure in DNA parity sequences. Further investigations along this line will include testing for nonlinearity by exploiting, for instance, surrogate data methods, a class of Monte Carlo tests aimed at building distribution-free hypothesis testing for nonlinear time series (for a review on the topic see [38]). Another important topic is to assess how the correlation structure of DNA sequences depends on the repetition of certain patterns throughout the sequence. Also in this case, it is possible to employ a suitable modification of surrogate data methods in order to build several statistical hypotheses in a straightforward manner. A similar task is pursued in [37], although the authors do not build statistical tests; rather, they seem to make comparisons on the basis of a single sequence rather than building a randomization distribution.

Conclusions

In this article, we have employed a novel mathematical theory for the genetic code in order to test the hypothesis that some error-control mechanism based on parity coding may be active inside the genetic machinery.

We have used this particular model mainly because

- it describes completely the degeneracy distribution of the genetic code
- it uncovers strong numeric and symmetry properties (TC degeneracy in the third letter, complement to 1 palindromy, coding of the third letter, etc.)
- it gives a natural coding method for parity characterization of codons
- there are not alternative nonpower representations describing the code degeneracy (the set of nonpower basis, [1 1 2 4 7 8], is unique).

Returning to error correction, it is known that different biochemical error-control systems are actually working at the level of amino acid translation in the ribosome, for

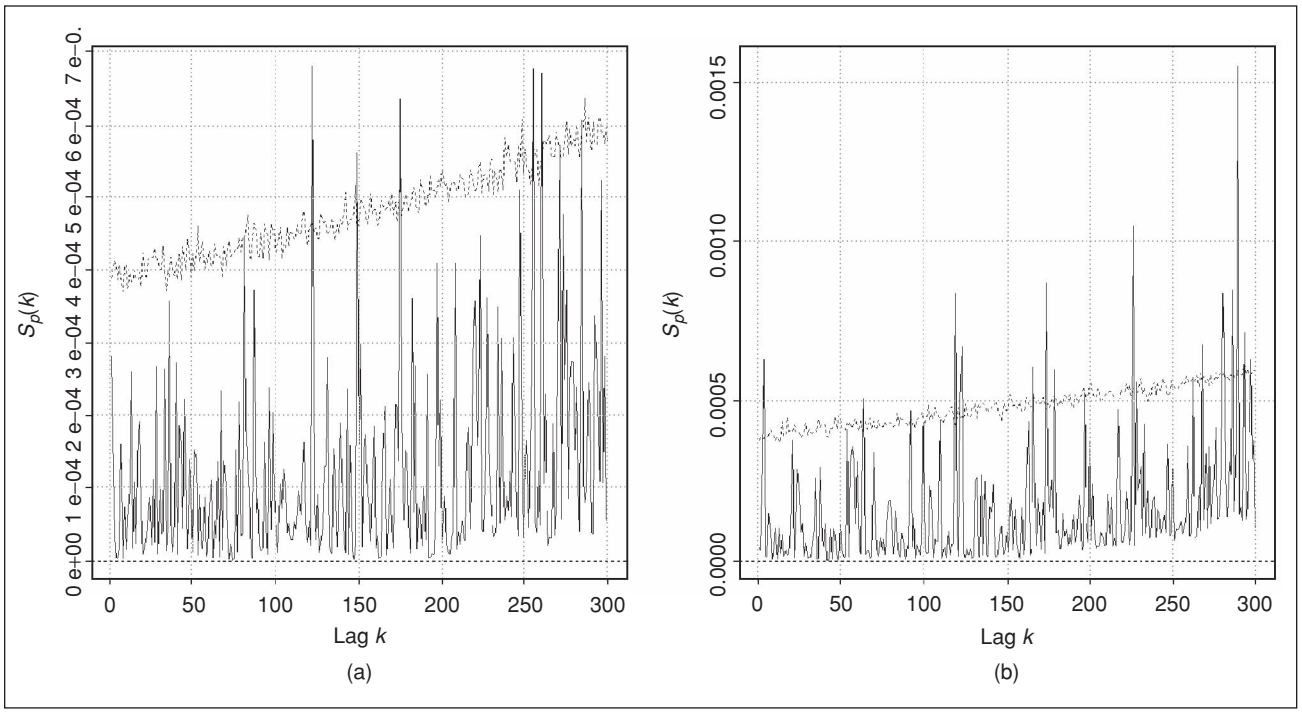


Fig. 6. $S_\rho(k), k = 1, \ldots, 300$ for out-frame sequences of the gene X80497 c2 (a) and c3 (b), $n = 1,235$. The confidence band at the 95% level (dashed line) was obtained through Monte Carlo generation of 3,000 Bernoulli IID sequences.

example, stopping a frame-shift reading [39]. However, many aspects of error control are poorly understood and the very low error rates associated with different genetic processes are difficult to be described theoretically outside a mathematical framework. Error discrimination and correction can be performed only if some mathematical template is available (of course, in the case of genetic processes, this template may represent some privileged state in terms of chemical energy exchange). Moreover, it is somehow accepted that the genetic code itself has gained its actual form thanks to its self-correcting capabilities [40]. However, this error-minimizing ability of the genetic code is usually studied within a probabilistic approach since random mutations either do not modify or minimally modify a synthesized protein. In this context, the decoding of an exon is kept immune to random mutations due to the particular structure of the genetic code itself (see [41] for some controversial related issues).

For a given protein, mutations can be viewed as little deviations from a coding template. Robustness of the protein synthesis is related to the fact that mutations usually modify a codon either into a synonymous codon (without noticeable changes in the protein) or into a codon codifying for a chemically similar amino acid (introducing minimal changes in the protein structure). However, the most striking fact is the existence of a given template, that is, the reason why nature has preferred a particular sequence of codons in order to codify a particular protein. A relatively short protein 100 amino acids long can be represented in 3^{100} (the average degeneracy per codon is approximately 3 because there are 61 codons representing 20 amino acids) or roughly 10^{47} different manners due to the freedom of choice between synonymous codons. In fact, an important theoretical question we have tried to address here and will continue to study in the future is how one or a few sequences are selected as the good ones in the ocean of equally valid possibilities. Is this choice related to some organizational principles in the genetic information such as error control and correction? For different reasons, our approach seems to point to an affirmative answer to this question.

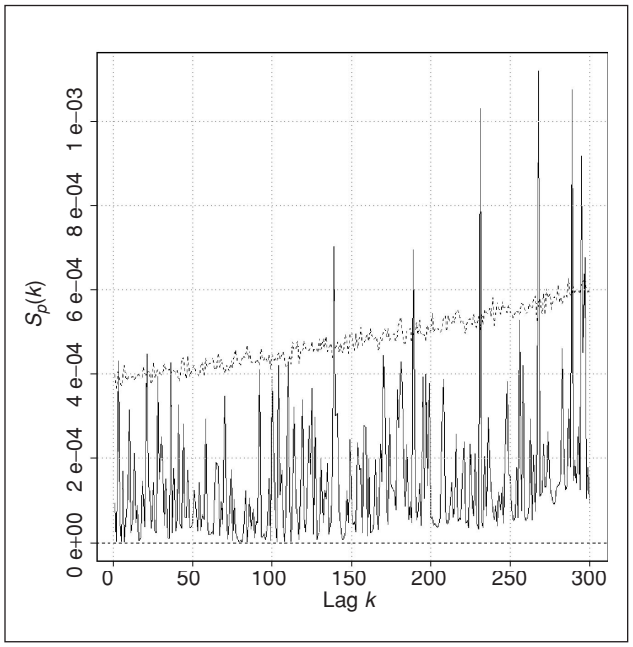


Fig. 7. $S_\rho(k), k = 1, \ldots, 300$ for the anticodon sequence X80497 in frame $a1$, $n = 1,236$. The confidence band at the 95% level (dashed line) was obtained through Monte Carlo generation of 3,000 Bernoulli IID sequences.

- The genetic code exhibits a strong mathematical structure that is difficult to put in relation with biological advantages other than error correction. It must be remarked that the probability of random generation of a similar but less restrictive mathematical ordering has been calculated to be 3.09^{-32}, that is, practically zero [11].
- Interestingly, this mathematical structure implies that codons are parity coded (parity coding represents the simplest and most widely used system for error checking and correction in man-made digital data communication systems).
- The statistical analysis performed in this work clearly shows that parity symbols exhibit strong and complicated dependence patterns (a necessary condition for the existence of mechanistic constraints).

From the point of view of the statistical analysis of the sequences, we have introduced two methods for a rigorous study of the serial dependence. In fact, the existence of error-correction mechanisms implies the presence of a correlation in the sequences. Both the approaches show the existence of a significant dependence and prompt us to pursue further investigation on the topic. As we have remarked above, the MBB is a tool to assign the accuracy to estimates in presence of dependent data. As a byproduct, it provides also a measure of the correlation of the sequence through (3). In particular, we have focused on the proportion p of even codons as defined through the theoretical approach presented above. Thus, the study of the variance associated to p gives substantive information about the dependence of the data in this coding framework. In addition, this analysis can be interesting for a comparative study. In fact, as can be desumed from Table 5, the proportion p does not depend explicitly on the GC content of the sequences, potentially allowing for a non-GC-biased comparison between organisms or regions of the same genome that differ in the GC content.

The entropy-based metric $S_\rho(k)$ is tailored to explore the dependence structure of a sequence and can be seen as a nonlinear autocorrelation function. Since its relation with several existing entropy-based measures and due to its good properties, $S_\rho(k)$ has revealed a powerful and informative tool in this context. The computation of $S_\rho(k)$ upon several parity sequences has highlighted the existence of a long-range dependence together with high peaks that might be associated to gene-specific periodicities. Even though a detailed phenomenological discussion is out of the scope of this article, it is clear that the results obtained through the two methods are coherent and complement each other. Hence, our approach appears to have a great potential in different applied fields related to genomics and bioinformatics.

These different findings prompt us to investigate the issue further under different hypotheses. A matter we intend to pursue is to assess whether the dependence we have observed is of a nonlinear nature. This hypothesis can be tested directly by means of nonparametric tests for nonlinearity based on surrogate data. Since the entropy-based measure $S_\rho(k)$ was shown to have considerable power against nonlinear alternatives, it is possible to employ it as a test statistic and build the Monte Carlo distribution of $S_\rho(k)$ under the null hypothesis that the series we have observed is a realization of a correlated linear process. Our guess is that a nonlinear dynamical system may represent a very efficient decoding system for the management of nonlinearly correlated information. Moreover, the observed dynamic complexity of such kind of systems (including chaotic behavior and also self-correcting capabilities) can

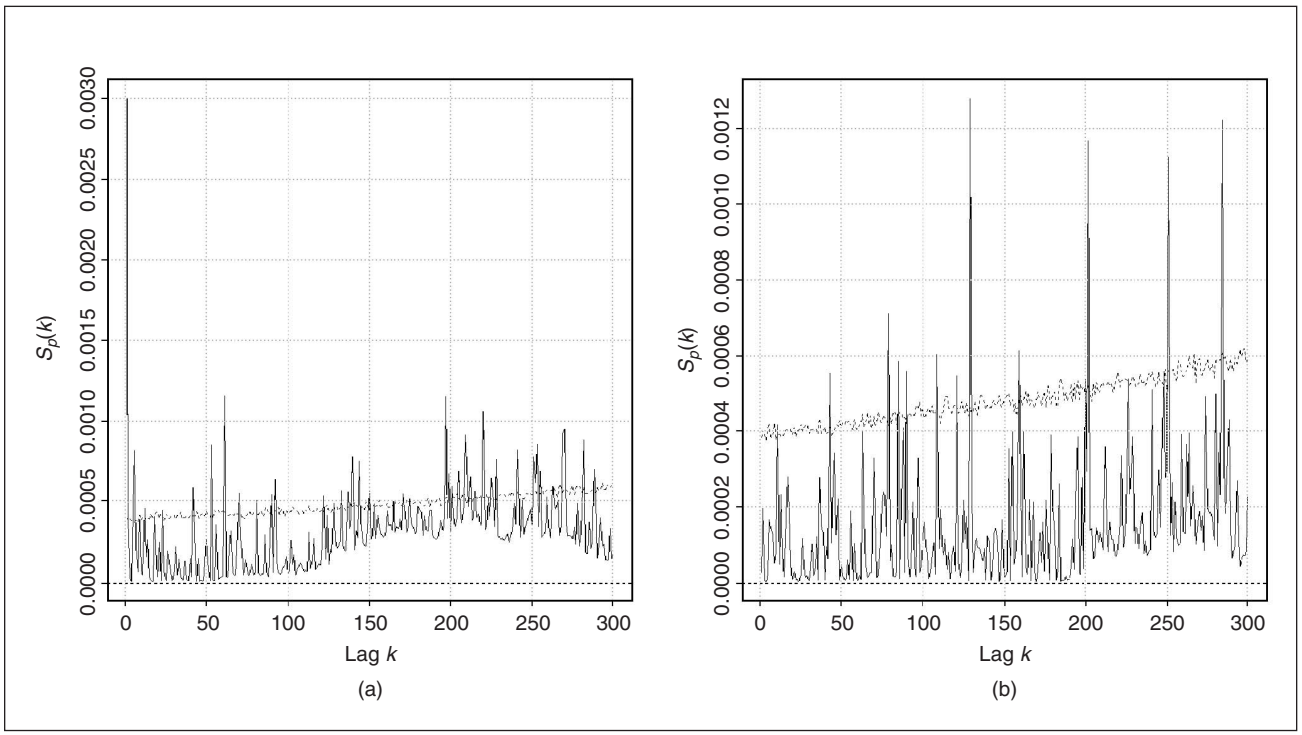
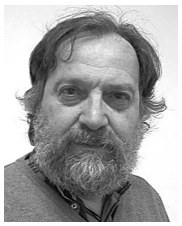


Fig. 8. $S_\rho(k)$, $k = 1, \ldots, 300$ for out-frame sequences of the gene X80497 $\alpha 2$ (a) and $\alpha 3$ (b), $n = 1,235$. The confidence band at the 95% level (dashed line) was obtained through Monte Carlo generation of 3,000 Bernoulli IID sequences.

offer an interesting possibility for determining the complex and elusive rules underlying the biological encoding of genetic information.

Diego Luis Gonzalez was born in Buenos Aires, Argentina, in 1951. He received his degree in physics from the University of La Plata in Argentina in 1981 and a Ph.D. in theoretical physics from the same university in 1987. His Ph.D. focused in the study of synchronization and chaos in nonlinear oscillators. He worked from 1988 for the National Research Council of Italy at LAMEL Institute in the field of microelectronics and microsystems. Since 1999 he has collaborated with the Acoustical Lab of the St. George School Foundation and the National Research Council in Venice, Italy. His main research interests are the theory of nonlinear dynamics and chaos and their application to the modeling of complex dynamic systems, with particular emphasis on biological systems.

Simone Giannerini was born in 1970 in Castiglione dei Pepoli (Bologna, Italy). He received his degree in statistics in 1998 and a Ph.D. in statistics in 2002 at Bologna University. He also obtained an M.Sc. in statistics at the London School of Economics in 2001. Since 2005, he has been a researcher at the statistics department at Bologna University. His research interests include chaos theory, nonlinear time series analysis, stochastic processes, and epidemiology.



Rodolfo Rosa was born in Bologna, Italy, in 1944. He received his degree in physics in 1968 and in philosophy in 1977. From 1969–1992 he has been a researcher at the National Research Council-LAMEL institute in Bologna. Since 1992 he has been a professor at the Faculty of Statistics, Bologna University, where he teaches courses on statistics for experimental research, chaos and complexity, and stochastic processes. His research interests include philosophy of science, Monte Carlo methods applied to atomic interactions in solids, statistical mechanics and, more recently, advanced statistical methods and chaos theory.

Address for Correspondence: Diego Luis Gonzalez, Laboratorio di acustica musicale e architettonica, CNR-Fondazione Scuola di S. Giorgio, Isola di San Giorgio Maggiore, Venezia, I-30124, Italy. E-mail: diego.gonzalez@cini.vecnr.it.

References

[1] B. Hayes, "The invention of the genetic code," *Comput. Sci.*, vol. 86, pp. 8–14, no. 14, 1998.
[2] H. Yockey, *Information Theory and Molecular Biology*. New York: Cambridge University Press, 1992.
[3] D.R. Fordsike, "Are introns in-series error detecting sequences?" *J. Theoret. Biol.*, vol. 93, no. 4, pp. 861–866, 1981.
[4] J. Rzeszowska-Wolny, "Is genetic code error-correcting?" *J. Theoret. Biol.*, vol. 104, no. 4, pp. 701–702, 1983.
[5] L. Liebovitch, Y. Tao, A. Todorov, and L. Levine, "Is there an error correcting code in the base sequence in DNA?" *Biophys. J.,* vol. 71, no. 3, pp. 1539–1544, 1996.
[6] G. Battail, "Is biological evolution relevant to information theory and coding?" in *Proc. ISCTA'01,* Ambleside, 2001, pp. 343–351.
[7] D. MacDónaill, "A parity code interpretation of nucleotide alphabet composition," *Chem. Commun.*, vol. 18, pp. 2062–2063, 2002.
[8] E. May, "Analysis of coding theory based models for initiating protein translation in prokaryotic organisms," Ph.D. dissertation, NC State Univ., Raleigh, NC, 2002.
[9] G. Rosen and J. Moore, "Investigation of coding structure in DNA," in *Proc. ICASPP*, 2003.
[10] D. Gonzalez and M. Zanna, "Una nuova descrizione matematica del codice genetico," *Systema Naturae, Annali di Biologia Teorica*, vol. 5, pp. 219–236, 2003.
[11] D. Gonzalez, "Can the genetic code be mathematically described?," *Med. Sci. Monitor*, vol. 10, no. 4, pp. 11–17, 2004.
[12] H. Künsch, "The jackknife and the bootstrap for general stationary observations," *Annals Stat.*, vol. 17, no. 3, pp. 1217–1241, 1989.
[13] C.W. Granger, E. Maasoumi, and J. Racine, "A dependence metric for possibly nonlinear processes," *J. Time Series Anal.*, vol. 25, no. 5, pp. 649–669, 2004.
[14] L.L. Gatlin, *Information Theory and the Living System*. New York: Columbia Univ. Press, 1972.
[15] E.E. May, M.A. Vouk, D.L. Bitzer and D.I. Rosnick, "A coding theory framework for genetic sequence analysis," *J. Franklin Instit.*, vol. 341, no. 1–2, pp. 89–109, 2004.
[16] E.E. May, "Towards a biological coding theory discipline," *New Thesis*, vol. 1, no. 1, pp. 19–38, 2004.
[17] J.E.M. Hornos and Y.M.M. Hornos, "Algebraic model for the evolution of the genetic code," *Phys. Rev. Lett.*, vol. 71, no. 26–27, pp. 4401–4404, 1993.
[18] V.A. Karasev and V.E. Stefanov, "Topological nature of the genetic code," *J. Theor. Biol.*, vol. 209, no. 3, pp. 303–317, 2001.
[19] V.I. Shcherbak, "Arithmetic inside the universal genetic code," *BioSyst.*, vol. 70, no. 3, pp. 187–209, 2003.
[20] A. Patel, "Quantum algorithms and the Genetic Code," *Pramana- J. Physics*, vol. 56, no. 2–3, pp. 367–381, 2001.
[21] A. Arneodo, Y. D'Aubenton-Carafa, E. Bacry, P.V. Graves, J.F. Muzi and C. Thermes, "Wavelet based fractal analysis of DNA sequences," *Physica D*, vol. 96, no. 1, pp. 291–320, 1996.
[22] M.A. Jimenez-Montaño, C.R. de la Mora-Basañez, and T. Pöschel, "The hypercube structure of the genetic code and non-conservative amino acid substitutions in vivo and in vitro," *BioSyst.*, vol. 39, no. 2, pp 117–125, 1996.
[23] N. Stambuk, P. Konjevoda, and N. Gotovak, "Binary Coding, mRNA Information and Protein Structure," *J. Comput. Information Technol.*, vol. 12, no. 2, pp. 73–81, 2004.
[24] R. Durbin, S. Eddy, A. Krogh, and G. Mitchinson, Eds., *Biological Sequence Analysis*. Cambridge, UK: Cambridge Univ. Press, 1998.
[25] B. Efron, "Bootstrap methods: Another look at the jackknife," *Annals Stat.*, vol. 7, no.1, pp. 1–26, 1979.
[26] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.
[27] B. Ripley, *Stochastic Simulation*. New York: Wiley, 1987.
[28] K. Binder, "Introduction," in *The Monte Carlo Method in Condensed Matter Physics*, K. Binder, Ed. Berlin: Springer-Verlag, 1982, pp. 1–22.
[29] S. Mignani and R. Rosa, "Markov chain Monte Carlo in statistical mechanics: The problem of accuracy," *Technometrics*, vol. 43, no. 3, pp. 347–355, 2001.
[30] S. Gottlieb, P. Mackenzie, H. Thacker, and D. Weingarten, "Hadronic coupling constants in lattice gauge theory," *Nuclear Physics B*, vol. 263, no. 3–4, pp. 704–730, 1986.
[31] R. Liu and K. Singh, "Moving blocks jackknife and bootstrap capture weak dependence," in *Exploring the Limits of Bootstrap*, R. LePage and L. Billard, Eds. New York: Wiley, 1992, pp. 225–248.
[32] J. Crutchfield and D. Feldman, "Regularities unseen, randomness observed: Levels of entropy convergence," *Chaos*, vol. 13, no. 1, pp. 25–54, 2003.
[33] H. Abarbanel, *Analysis of Observed Chaotic Data*. New York: Springer Verlag, 1996.
[34] H. Joe, "Relative entropy measures of multivariate dependence," *J. Amer. Stat. Assoc.* vol. 84, no. 405, pp. 157–164, 1989.
[35] Oryctolagus cuniculus glycogen synthase mRNA [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=2384761.
[36] PHKLA gene; phosphorylase kinase, Homo Sapiens [Online]. Available: http://www. ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=663009
[37] D. Holste, I. Grosse, S. Beirer, P. Schieg, and H. Herzel, "Repeats and correlations in human DNA sequences," *Phys. Rev. E*, vol. 67, 061913, 2003.
[38] T. Schreiber and A. Schmitz, "Surrogate time series," *Physica D*, vol. 142, no. 3–4, pp. 346–382, 2000.
[39] V. Marquez, D. Wilson, and K.H. Nierhaus, "RNA-protein machines," *Biochem. Soc. Trans.*, vol. 30, no. 2, pp. 133–140, 2002.
[40] D. Ardell, "On error minimization in a sequential origin of the standard genetic code," *J. Mol. Evol.*, vol. 47, no. 1, pp. 1–13, 1998.
[41] S. Freeland and L. Hurst, "Load minimization of the genetic code: History does not explain the pattern," in *Proc. R. Soc. Lond. B*, 1998, vol. 265, pp. 2111–2119.