MATHEMATICAL BIOSCIENCES
AND ENGINEERING
Volume **10**, Number **1**, February **2013**

doi:10.3934/mbe.2013.10.199

pp. **199–219**

GENOME CHARACTERIZATION THROUGH DICHOTOMIC CLASSES: AN ANALYSIS OF THE WHOLE CHROMOSOME 1 OF *A. THALIANA*

Enrico Properzi and Simone Giannerini

Dipartimento di Scienze Statistiche, Università di Bologna
Via delle Belle Arti 41, 40126, Bologna, Italy

Diego Luis Gonzalez

CNR-IMM, UOS di Bologna
Via Gobetti 101, 40129 Bologna, Italy
and
Dipartimento di Scienze Statistiche, Università di Bologna
Via delle Belle Arti 41, 40126 Bologna, Italy

Rodolfo Rosa

Dipartimento di Scienze Statistiche, Università di Bologna
Via delle Belle Arti 41, 40126 Bologna, Italy
and
CNR-IMM, UOS di Bologna
Via Gobetti 101, 40129 Bologna, Italy

Abstract. In this article we show how dichotomic classes, binary variables naturally derived from a new mathematical model of the genetic code, can be used in order to characterize different parts of the genome. In particular, we analyze and compare different parts of whole chromosome 1 of *Arabidopsis thaliana*: genes, exons, introns, coding sequences (CDS), intergenes, untranslated regions (UTR) and regulatory sequences. In order to accomplish the task we encode each sequence in the 3 possible reading frames according to the definitions of the dichotomic classes (parity, Rumer and hidden). Then, we perform a statistical analysis on the binary sequences. Interestingly, the results show that coding and non-coding sequences have different patterns and proportions of dichotomic classes. This suggests that the frame is important only for coding sequences and that dichotomic classes can be useful to recognize them. Moreover, such patterns seem to be more enhanced in CDS than in exons. Also, we derive an independence test in order to assess whether the percentages observed could be considered as an expression of independent random processes. The results confirm that only genes, exons and CDS seem to possess a dependence structure that distinguishes them from i.i.d sequences. Such informational content is independent from the global proportion of nucleotides of a sequence. The present work confirms that the recent mathematical model of the genetic code is a new paradigm for understanding the management and the organization of genetic information and is an innovative tool for investigating informational aspects of error detection/correction mechanisms acting at the level of DNA replication.

2010 *Mathematics Subject Classification.* 92B05, 92D20, 62P10.

Key words and phrases. Dichotomic classes, genetic code, Arabidopsis thaliana, statistical tests.

1. **Introduction.** A modern working definition of a gene is *a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and/or other functional sequence regions*[16, 17]. Usually we refer to a gene as a region of DNA that encodes for a polypeptide or for an RNA chain that has a function in the organism.

Information transfer processing from a gene to a protein requires two different steps: first, the DNA information is transcribed into messenger RNA (mRNA); then mRNA is translated into a protein on the basis of the genetic code, a universal translation table that links the world of nucleic acids to the world of proteins. The discovery of the genetic code led scientists to focus on sequencing the entire genomes of different organisms. The Human Genome Project succeeded in sequencing the whole human genome in 2001 [13, 24] and triggered a strong hype on the possibility of diagnosing and treating many serious diseases. However, after ten years, it looks like the expectations have not been met. The recent article by S.S. Hall published on Scientific American: "Revolution Postponed: Why the Human Genome Project Has Been Disappointing" is emblematic: In fact its subtitle states: "*The Human Genome Project has failed so far to produce the medical miracles that scientists promised. Biologists are now divided over what, if anything, went wrong - and what needs to happen next*".

Genomes are composed also by noncoding DNA, those sequences that do not encode for any protein. For instance, more than 98% of the human genome is composed by noncoding DNA [2, 15]. The majority of noncoding DNA sequences have no known biological function and are sometimes referred to as *junk DNA*. However, some types of noncoding DNA sequences have biological functions such as the regulation of gene expression (i.e. promoters and enhancers). Some other noncoding sequences are transcribed but not translated (introns). Other noncoding sequences are highly conserved: between them we can distinguish regulatory regions, transposable elements and pseudogenes (vestiges of once-functional genes disabled by sequence deletions, insertions or mutations) [14].

The whole genetic information is passed from a parent cell to two or more daughter cells through the process of cell division. The main concern of cell division is the maintenance of the genome of the original cell. Before division can occur, the genomic information must be replicated and the duplicated genome is separated cleanly between cells. During DNA replication several errors may occur. Some of these errors have no effect on the life of the cell, while others can result in growth defects, cell death or cancer.

Cancer is a genetic disease although it is not usually an inherited disease. Cancer development in the body is due to a combination of events. Mutations occasionally occur within cells as they divide and can affect the behavior of cells, sometimes causing them to grow and divide more frequently. Several biological mechanisms can stop this process: biochemical signals can cause inappropriately dividing cells to die. Sometimes additional mutations make cells ignore these messages. Most dangerously, a mutation may give a cell a selective advantage, allowing it to divide more vigorously than its neighbours and to become a founder of a growing mutant clone. Eventually, mutations can accumulate within cells to promote their own growth, creating a tumour. Since mutations may occur because of errors during DNA replication, the study of error detection/correction mechanism in such process could be of key importance for understanding the onset of cancer.

The genetic code it is a surjective mapping between the set of 64 possible three-base codons and the set of 20 amino acids (plus the stop signal). Many slight variants of the standard genetic code have been discovered, including various alternative mitochondrial codes. Moreover in certain proteins, non-standard amino acids can substitute standard stop codons. For example, UGA can code for Selenocysteine, and UAG can code for Pyrrolysine. Selenocysteine is then seen as the 21^{st} biologically functional amino acid, and Pyrrolysine is seen as the 22^{nd} [3].

In biology, a reading frame is a way of breaking a sequence of nucleotides in DNA or RNA into three letter codons which can be translated in amino acids. There are 3 possible reading frames in an mRNA strand: each reading frame corresponds to starting at a different alignment. Usually, there is only one correct reading frame. Moreover, error detection and correction mechanisms are strictly involved with frame recognition. Coding sequences possess a local informational structure that can be related to frame synchronization processes [11]. But is frame important for non-coding sequences too? Are regulatory sequences, introns and intergenic sequences related to the frame and or to the existence of codons?

In this work we try to answer to the above questions by investigating the entire genome of chromosome 1 of *Arabidopsis thaliana*, one of the most popular model plant. *A.thaliana* has many advantages for genome analysis: a small size, a short generation time and relatively small nuclear genome. These advantages promoted the growth of a scientific community that has investigated the biological processes of *A. thaliana* and has characterized many genes [21]. Also, several studies have identified in *A. thaliana* genes homologous to human oncogenes or tumor suppressor genes [19, 23].We study the role of frame in coding and non coding sequences in the genome of *A. thaliana* by using a recently developed mathematical model for the genetic code [4, 5, 6]. In particular we use the information of dichotomic classes, binary variables naturally derived from the above mentioned model, in order to assess different behaviours between coding and non coding sequences. So far, the mathematical model of the genetic code has been used to investigate only some proteins of different origin [7, 8, 9, 10, 11]. Now we apply it to a whole chromosome of a single (and well known) organism. The finding of some local (or global) information structure related to dichotomic classes could be useful in order to develop alternative methods to understand error detection and correction mechanisms involved in the translation process.

The paper is organized as follows: in section 2 we describe the salient features of the mathematical model. In section 3 we describe the organism of *A. thaliana* and the steps followed to obtain the data set analyzed. In Section 4 we perform a descriptive statistical analysis on the data set; moreover, we implement and apply a test for independence based on dichotomic classes. In the last section we discuss the results.

2. **Mathematical model.**

The genetic code is the dictionary used by the cell to translate a sequence of codons (triplets or bases) of RNA in a sequence of amino acids during the translation process. Almost all living organisms use the same genetic code, called the standard genetic code, but many slight variants have been discovered. All known naturally-occurring codes are very similar and the coding mechanism is the same for all the organisms: it implies three-base codons, tRNA, ribosomes, reading the code in the same direction and translating the code three letters at a time into sequences of amino acids.

TABLE 1. Euplotid version of the genetic code

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

The DNA is made of four bases: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) (in RNA thymine is replaced by uracil (U)). Hence, there are $4^3 = 64$ possible codons; 61 of them encode amino acids, while the remaining three (TAA, TAG, TGA) encode stop signals that indicate the point where the assembly of the polypeptide chain should be stopped. Since the amino acids that contribute to the formation of proteins are only 20, some amino acids are necessarily encoded by more than one codon. This fact determines the properties of redundancy and degeneracy typical of the genetic code. Indeed, from a mathematical point of view, we can say that genetic code is a surjective (all amino acids are encoded by at least one codon) and non-injective (some amino acids are degenerate) function between two sets of different cardinality (amino-acids and codons). The euplotid version of genetic code is shown in table 1. The main difference with the standard version concerns the TGA codon: here it encodes the amino acid Cysteine while in the standard version of the code it is one of the stop signals. In table 1 codons are displayed into quartets: groups of four codons sharing the first two bases.

2.1. **Degeneracy and redundancy.** [4, 6] proposed a model that explains the degeneracy of the genetic code based on a non-power number representation system. This approach describes the structure of the genetic code from a mathematical point of view and allows the analysis of degeneracy and redundancy properties.

Degeneracy and redundancy are still described by the numerical quantities that define the respective sub-sets: Tyrosine is a degeneracy-2 amino acid because it is encoded by a set of two redundant codons (TAT, TAC). Table 2 shows the degeneracy distribution inside quartets of the euplotid genetic code.

The degeneracy distribution inside quartets is obtained by taking into account that the 3 degeneracy-6 amino acids (Arginine, Leucine and Serine) are divided into two subsets of degeneracy-2 and 4.

TABLE 2. Degeneracy distribution inside quartets of euplotid nuclear version of genetic code

number of amino acids sharing the same degeneracy	Degeneracy
2	1
12 (9+3)	2
2	3
8 (5+3)	4

2.2. **Non-power binary representation of the genetic code.** By using a binary string of length n we can represent 2^n different objects. For example, the 4 nucleotides (A, C, G, T) can be represented by a binary string of length 2. Consequently codons (groups of 3 nucleotides) can be represented by binary strings of length 6, in fact $2^6 = 64$. However, it is possible to show that fixed representation systems of this kind are not able to describe the degeneracies of the genetic code. Now, in [4, 6] it has been shown that a particular type of number positional representation, called non-power representation, can fully describe the degeneracy distribution of the genetic code. Usual number representation systems are positional power representation systems. In these systems numbers are represented by a combination of digits, from 0 to $n-1$, where n is the system base, that are weighted with values that grow following the power expansion of the base n. For example, if we want to represent number 476 in base 10 we have to use the digits as follows:

$$476 = 4 \times 10^2 + 7 \times 10^1 + 6 \times 10^0$$

while number 13 is obviously represented by

$$13 = 1 \times 10^1 + 3 \times 10^0$$

If we turn to the binary system, the power positional representation of number 13 is 1101:

$$13 = 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$$

In *non-power representation systems* the positional values grows more slowly than the powers of the system base. This implies that it is possible to represent redundantly all the numbers from 0 to the sum of all the positional weights. In other words, a given number can be represented by more than one string. Hence, non-power representation systems can be used to describe degeneracy distributions. Remarkably, there is a non-power binary representation that describes perfectly the degeneracy of the genetic code. This system is based on a specific sequence of positional weights: $(8, 7, 4, 2, 1, 1)$ and this solution is unique up to trivial equivalence classes [12, 4]. The solution is specific for the degeneracy inside quartets for the euplotid version of the code presented in table 2.

Table 3 shows the non-power representation of the first 23 integers by length-6 binary strings and positional weights $(8, 7, 4, 2, 1, 1)$. Notice the same degeneracy distribution of euplotid genetic code (see table 4).

We can state that each codon can be associated to a length-6 binary string representing a whole number. Thus, the genetic code and this specific non-power binary representation are linked by a structural isomorphism: they share the same structure. However this result does not represent a model of the genetic code *per se*.

TABLE 3. Non power representation of whole numbers from 0 to 23 by length-6 binary strings

Number	Positional weights: [8,7,4,2,1,1]
0	000000
1	000001 000010
2	000011 000100
3	000101 000110
4	000111 001000
5	001001 001010
6	001011 001100
7	010000 001101 001110
8	100000 010001 010010 001111
9	100001 100010 010100 010011
10	100011 100100 010101 010110
11	100101 100110 011000 010111
12	101000 100111 011001 011010
13	101001 101010 011100 010111
14	101100 101011 011100 011011
15	110000 101101 101110 011111
16	110001 110010 101111
17	110100 110011
18	110101 110110
19	111000 110111
20	111001 111010
21	111100 111011
22	111101 111110
23	111111

In order to build a model we need to establish links between aminoacids (defined by codons) and numbers from 0 to 23 (defined by 6-bit strings). The task can be accomplished by studying symmetry properties of both these mappings.

2.3. **A hierarchy of symmetries.**

2.3.1. *Pyrimidine ending codons.* If we analyze the genetic code and the mathematical model we can notice many symmetry properties. First, if we make a Pyrimidine (T vs C) exchange in the last base of each codon, the meaning of the codon remains the same. This implies the definition of 16 groups of two codons each that encode the same amino acid. So far we know 26 variants of the genetic code (10 nuclear and 16 mitochondrial) and all of these respect this symmetry. Remarkably, the non-power representation system shows an analogous symmetry. In fact, the 6-digit binary strings xxxx01 and xxxx10 always encode the same number and define 16 groups of two strings each. This is a property of this specific representation system because of the positional weights chosen. This means that we can associate binary strings ending in 01 or 10 with Pyrimidine ending codons. Notice that there is no biochemical reason for this degeneracy as codons ending in C or T can be recognized by different tRNAs [25].

TABLE 4. Palindromic representation of the degeneracy of the euplotid version of the genetic code and non-power representation of whole numbers

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

2.3.2. *Purine ending codons.* The former aspect determines an immediate consequence since the remaining 32 codons have to be associated with the remaining 32 strings representing whole numbers. Thus strings ending in 00 or 11 are necessarily associated to Purine ending codons. Since the only two degeneracy-1 strings (000000 and 111111) have to be associated with the only degeneracy-1 codons (ATG and TGG) a new concept raises naturally: the parity of a string, that is, the sum of its digits. So, we can assume that, in case of Purine ending codons, strings with even parity are associated to G-ending codons, while, by exclusion, strings with odd parity are associated to A-ending codons. All these aspects are summarized in table 5

2.3.3. *Degeneracy-3 elements.* We can notice that there are only two degeneracy-3 whole numbers (7 and 16) and amino acids (Cysteine and Isoleucine). Obviously these elements must be associated. So the group (ATT, ATC, ATA) and (TGT, TGC, TGA) are linked with binary strings representing numbers 7 and 16. These two groups of codons are linked by a degeneracy-preserving transformation: T $\leftrightarrow$ A in the first base and T $\leftrightarrow$ G in the second one. It is remarkable to notice that this transformation corresponds to a symmetry property from the model point of

TABLE 5. Equivalence between strings and Purine/Pyrimidine ending codons

Strings	Parity	Codons
x x x x 0 1	Even	N N C/T
x x x x 0 1	Odd	N N C/T
x x x x 1 0	Even	N N C/T
x x x x 1 0	Odd	N N C/T
x x x x 1 1	Even	N N G
x x x x 1 1	Odd	N N A
x x x x 0 0	Even	N N G
x x x x 0 0	Odd	N N A

view: the palindromic symmetry. In fact the first group strings can be obtained by the $0 \leftrightarrow 1$ exchange of the digits of the second group strings. This palindromy is observed also for the degeneracy-1 string (000000 and 111111) coding for amino acids Methionine (ATG) and Tryptophan (TGG). Notice that the numbers encoded by a palindromic couple sum up to 23.

Summarizing, we can state that degeneracy-3 and degeneracy-1 amino acids form two groups of quartets that show a *palindromic symmetry*. Notice that, differently from the euplotid code, in the standard genetic code we have TGA codon encoding a stop signal and not Cysteine. Palindromic symmetry involves all the quartets of the genetic code. It connects quartets with the same degeneracy distribution and strings related by the complement to one operation.

2.3.4. *Degeneracy-6 elements.* By analyzing table 6 we can see that there are two degeneracy-2 numbers that correspond to A-ending codons (4 and 19); but there are no amino acids with degeneracy 2 encoded by two A-ending codons. Therefore these numbers must be associated with the degeneracy-2 part of degeneracy-6 amino acids encoded by at least two A-ending codons. Looking at the tables it is easy to recognize these amino acids in leucine (Leu) and arginine (Arg). Both of them are encoded also by two G-ending codons that necessarily belongs to their degeneracy-4 part. The only degeneracy-4 numbers showing this feature are 11 and 12: both of them display two even strings ending with 00 or 11. It can be observed once more that this couples of numbers (4 and 19) and (11 and 12) are palindromic (their sum equals 23). So we can state that there is a symmetry of the role of Leu and Arg.

2.3.5. *Pyrimidine ending codons with odd parity.* We succeeded in linking binary strings with codons whose second letter is T or G. Moreover all the T or C ending codons so far associated show an odd parity. So we can state that amino acids with Pyrimidine ending codon and with a G or a T (Keto base) in the second position are encoded by an odd string We can find only two degeneracy-4 numbers (10 and 13) and only two degeneracy-2 numbers (1 and 22) satisfying this rule and, as a consequence, we can associate to them the amino acids valine (Val), glycine (Gly), phenylalanine (Phe) and the degeneracy-2 part of serine (Ser).

2.3.6. *The last associations: second base A or C.* Now it remains to associate only codons whose second base is an Amino-base (A or C). It is quite simple because codons with A in second position share all degeneracy-2, while codons with C in the second position have degeneracy-4. Following the rules described above, we can

try to give a place to all codons into the mathematical model. The result is shown in table 6

TABLE 6. Non-power model of the euplotid nuclear genetic code

	T			C			A			G			
T	1	000001	Phe	15	101101	Ser	18	110110	Tyr	16	110010	Cys	T
	1	000010	Phe	15	101110	Ser	18	110101	Tyr	16	110001	Cys	C
	4	001000	Leu	15	011111	Ser	2	000100	Stp	16	101111	Cys	A
	11	011000	Leu	15	110000	Ser	2	000011	Stp	23	111111	Trp	G
C	11	100101	Leu	14	011110	Pro	3	000101	Tyr	12	011010	Arg	T
	11	100110	Leu	14	011101	Pro	3	000110	Tyr	12	011001	Arg	C
	4	000111	Leu	14	101100	Pro	17	110100	Stp	19	111000	Arg	A
	11	010111	Leu	14	101011	Pro	17	110011	Stp	19	101000	Arg	G
A	7	001101	Ile	8	010010	Thr	5	001001	Asn	22	111110	Ser	T
	7	001110	Ile	8	010001	Thr	5	001010	Asn	22	111101	Ser	C
	7	010000	Ile	8	100000	Thr	21	111011	Lys	19	110111	Arg	A
	0	000000	Met	8	001111	Thr	21	111100	Lys	12	100111	Arg	G
G	13	101001	Val	9	100001	Ala	20	111010	Asp	10	010110	Cys	T
	13	101010	Val	9	100010	Ala	20	111001	Asp	10	010101	Cys	C
	13	011100	Val	9	010011	Ala	6	001011	Glu	10	100011	Cys	A
	13	011011	Val	9	010100	Ala	6	001100	Glu	10	100100	Trp	G

It is easy to notice how palindromy preserve degeneracy within quartets. From a mathematical point of view palindromy is represented by the complement to one operation of all the binary digits of a given string. From a biochemical point of view palindromy is given by different base transformations depending on the quartet considered. By looking at table 6 we succeeded in assigning a binary string to each codon but of course the solution is not unique. For instance, it is possible to exchange the full set of strings of a quartet with the set assigned to the palindromic quartet. This assignation is one of the most probable, given all the symmetry properties presented.

2.4. **Dichotomic classes.** By studying the degeneracy properties of the genetic code we can classify di-nucleotides into three dichotomic classes: parity, Rumer and hidden. For the definition of these classes it is necessary to introduce the unique three possible chemical classification of the bases (T, C, A, G):

Purine (R)	vs	Pyrimidine (Y):	{A,G} vs {C,T}
Keto (K)	vs	Amino (Am):	{G,T} vs {A,C}
Strong (S)	vs	Weak (W):	{C,G} vs {A,T}

2.4.1. *Parity class.* According to the mathematical model described so far, each codon is associated to a binary string. The parity of a codon corresponds to the parity of the sum of all the digits of the associated string. We can observe that the parity of a binary string can be obtained simply by counting the number of ones: an even number of ones leads to an even string while an odd number of ones leads to an odd string. It is important to underline that palindromic symmetry preserves parity; in fact, the complement to one operation does not change the parity of the string since the string length is even. The parity bit of a string can be determined also by its biochemical composition: first, any codon ending with A (G) is represented by an odd (even) string. Instead, if the codon ends with a Pyrimidine (T or C) then we have to look at the second base of the codon: when it is an Amino-base then the codon is even, while a Keto-base in the second position

leads to an odd codon. Also notice that the R-Y transformation changes the parity of a string. Now, it is possible to build an algorithm in order to define the parity of a codon from its biochemical composition. This algorithm involves the last two bases of the codon and it is shown in Figure 1

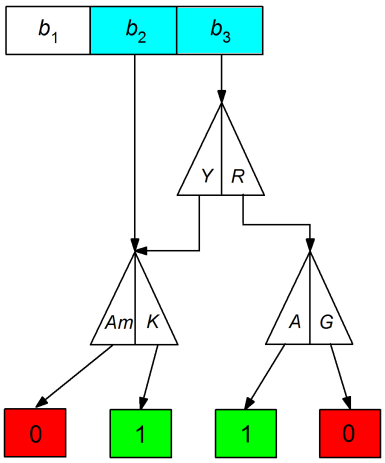

FIGURE 1. Algorithmic definition of the parity class.

2.4.2. *Rumer's class.* Y.U.B. Rumer was a theoretical physicist who first noticed a regularity of the degeneracy distribution within quartets in the standard genetic code. He observed that exactly one half of the quartets showed degeneracy-4 while the other half showed degeneracy 1, 2 or 3. Thus, each codon can be assigned to a dichotomic class named Rumer's class depending on whether it belongs to a degeneracy-4 or degeneracy 1, 2 or 3 quartet. Moreover, Rumer observed that a specific transformation, called Rumer's transformation, links the two halves of the genetic code: T,C,A,G $\leftrightarrow$ G,A,C,T. Rumer's transformation converts a codon of class 1 2 or 3 in a codon of class 4 and vice-versa; it breaks the degeneracy of the code since it reveals an antisymmetric property of the degeneracy distribution. Rumer's transformation is global i.e. it acts univocally on the 4 mRNA bases.

By looking at the chemical properties of the bases we can create an algorithm for determining Rumer's class (see Figure 2). First, we can take into account the second base of a codon: if it is an Amino-base we can immediately determine the class (class 4 if it is C, class 1,2,3 if it is A). If the second base is a Keto-type base (G or T) we need to consider the Strong/Weak character of the first base of the codon. If the first base is a Strong type base (C or G) then the codon has class 4. Otherwise it has class $\neg 4$ (1,2 or 3).

2.4.3. *Hidden class.* We observed that Y-R transformation changes the parity of a codon, while the K-Am transformation changes the Rumer's class. Now, we can postulate the existence of a third class, called hidden class that is antisymmetric with respect to the global transformation S-W. This class can be defined by an algorithm similar to those proposed for Rumer and parity classes as shown in Figure 3. In this case we have to consider the bases of two different codons: the first base of a codon and the third base of the previous one. If the first base is a Weak base

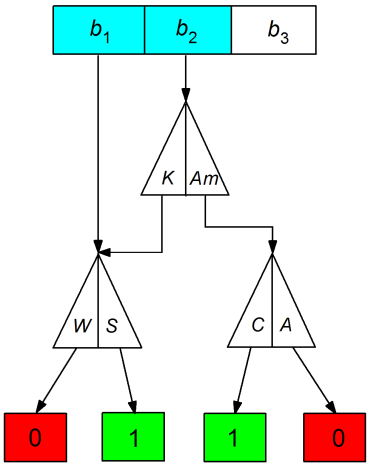

FIGURE 2. Algorithmic definition of the Rumer's class.

(A or T) then the hidden class is 0 for A and 1 for T. In case of Strong first base (C or G) we have to consider the last base of the previous codon: if it is a Pyrimidine base the hidden class is 0 otherwise it is 1.

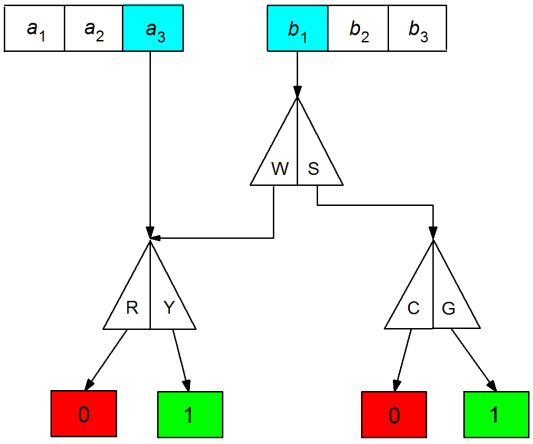

FIGURE 3. Algorithmic definition of the hidden class.

The three global transformations described above, together with the identity transformation, define a Klein V group structure as shown in table 7.

3. The chromosome 1 of *A. thaliana*.

3.1. *A. thaliana* as a model organism. *Arabidopsis thaliana* is a small flowering plant native to Europe, Asia, and northwestern Africa. A spring annual with a relatively short life cycle, *A. thaliana* is popular as a model organism in plant biology and genetics. *A. thaliana* has a rather small genome, only 157 megabase pairs (Mbp) and five chromosomes [22]. Arabidopsis was the first plant genome to be sequenced, and is a popular tool for understanding the molecular biology of

TABLE 7. Product table of the Klein V group as implied by the three global transformations plus the identity.

	I	**K-Am**	**S-W**	**Y-R**
I	I	K-Am	S-W	Y-R
K-Am	K-Am	I	Y-R	S-W
S-W	S-W	Y-R	I	K-Am
Y-R	Y-R	S-W	K-Am	I

many plant traits. By the beginning of 1900s, *A. thaliana* began to be used in some developmental studies. It plays the role in plant biology that mice and fruit flies (Drosophila) play in animal biology. Although *A. thaliana* has little direct significance for agriculture, it has several traits that make it a useful model for understanding the genetic, cellular, and molecular biology of flowering plants.

The small size of its genome makes *A. thaliana* useful for genetic mapping and sequencing. It was the first plant genome to be sequenced, completed in 2000 by the Arabidopsis Genome Initiative [21]. The most up-to-date version of the *A. thaliana* genome is maintained by the Arabidopsis Information Resource (TAIR). Much work has been done to assign functions to its 27,000 genes and the 35,000 proteins they encode [22].

3.2. **Dataset.** We consider seven groups of sequences (see Fig. 4) from the chromosome 1 of *A. thaliana*, that is composed by a long DNA sequence of 30.427.671 base pairs as follows:

1. **Genes**: regions of a genomic sequence corresponding to a unit of inheritance. They are formed by regulatory regions, transcribed regions, and/or other functional sequence regions.
2. **Exons**: portions of a gene that are transcribed into mRNA and then translated into a protein. Each gene can contain one or more exons.
3. **CDS**: portions of a gene that encode for a given protein. It is formed by joining exons (one or more) within a gene.
4. **Introns**: portions of a gene that are transcribed but not translated.
5. **Intergenes**: sequences between a gene and the following one.
6. (**UTR**): portions of mRNA that precede the codon that begins translation (AUG) (5'UTR) and follow the termination codon (3' UTR)
7. **Regulatory regions**: portions of a gene, with regulatory function, that precede (upstream) and follow (downstream) the fragment transcripted into mRNA

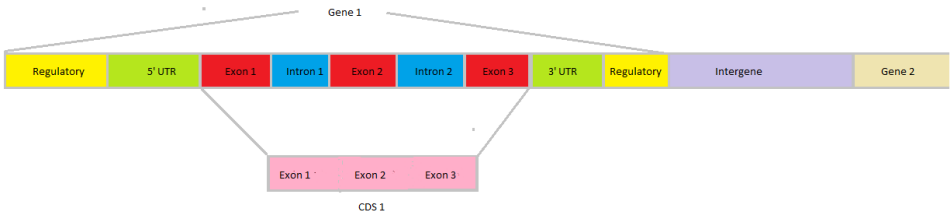

FIGURE 4. Definition of type of sequences within a fragment of DNA

First of all we have extracted the complete sequence of *A. thaliana* chromosome 1 from Genbank. This dataset allows to extract four kind of sequence data in fasta format: the complete sequence of the entire chromosome, a list of the genes sequences, a list of CDS, a list of mRNA sequences. We imported and processed the data by the means of R [20]. Then we created specific routines that, using the information of annotation, allowed us to extract the remaining group of sequences of interest: exons, introns, intergenes, 5' and 3' untranslated regions (UTR), and upstream and downstream regulatory regions. The annotation file, in fact, contains useful information for this purpose such as the nucleotide position of the beginning and the end of each gene, CDS and mRNA. The procedure led to the creation of seven datasets, one for each sequence group.

Once the data have been imported, we removed from the datasets those sequences that display undefined bases (different from A, C, G, T) or that are shorter than 6 bases. The seven different dataset together with the number of records are shown in table 8.

TABLE 8. Number of records and percentages of bases for each type of sequence analyzed from *A. thaliana* chromosome 1

Type	***Records***	A	C	G	T
Genes	8428	28.47	18.77	21.33	31.43
Exons	37549	29.00	19.94	23.73	27.33
CDS	9262	28.61	20.48	23.87	27.04
Introns	30663	26.93	15.72	16.68	40.68
Intergenes	8350	34.01	15.92	16.04	34.03
UTR	14427	30.42	17.76	16.78	35.10
Reg	2037	31.08	18.34	16.38	34.19

4. **Statistical analysis.** In this section we perform a statistical analysis on the dichotomic classes computed on the seven groups of sequences of the chromosome 1 of *A. thaliana* described in the previous section. As mentioned above, the aim is to study whether the information conveyed by dichotomic classes can characterize different portions of the genome. In order to accomplish the task, we code all the sequences into the three dichotomic classes and study the distributions of such binary sequences. In particular, we focus on their mean value, that is, the percentage of ones. Thus, for each sequence, we obtain 22 variables as reported in table 9:

The first results are reported in tables 10 and 11 where we show the means over the set of records defined in table 8 computed respectively on sense and antisense strands. The median values (not shown here) are very similar to the means and lead to the same conclusions.

From the two tables we can notice that coding and non-coding sequences show a different behaviour. In fact, in noncoding sequences (except for Exons and CDS) the mean values do not vary with frame (for example $p0 \simeq p1 \simeq p2 \simeq 60\%$). Such similarity is very high for UTR and regulatory sequences. On the contrary, Exons and CDS (the only sequence that undergo the transcription and translation processes) show different mean values in different frames. For example, if we consider parity, we can see that $p0$ is similar to $p1$ ($\sim 50\% - 52\%$) but both of them are lower than $p2$ (56.26% for exons and 58.67% for CDS, respectively). This kind of analysis suggests that the frame plays a role only for coding sequences; moreover,

TABLE 9. Variables included in each dataset

Name	**Description**
$p0, r0, h0$	mean value for parity, Rumer, hidden classes, in frame
$p1, r1, h1$	mean value for parity, Rumer, hidden classes, out of frame 1
$p2, r2, h2$	mean value for parity, Rumer, hidden classes, out of frame 2
$p0a, r0a, h0a$	mean value for parity, Rumer, hidden classes, antisense strand in frame
$p1a, r1a, h1a$	mean value for parity, Rumer, hidden classes, antisense strand out of frame 1
$p2a, r2a, h2a$	mean value for parity, Rumer, hidden classes, antisense strand out of frame 2

TABLE 10. Mean values of the percentages of dichotomic classes computed in sense strand

	$p0$	$p1$	$p2$	$r0$	$r1$	$r2$	$h0$	$h1$	$h2$
Genes	55.19	55.86	56.84	38.87	38.66	38.56	48.00	48.18	47.17
Exons	52.13	52.71	56.26	42.17	39.93	39.47	50.58	51.44	47.41
CDS	50.13	51.63	58.67	44.09	41.74	38.81	51.88	54.78	45.90
Introns	61.21	60.40	59.29	34.85	33.06	32.82	38.74	38.33	37.83
UTR	60.22	60.01	60.15	35.27	35.24	35.25	44.55	44.23	44.31
Regulatory	59.67	59.72	60.19	35.86	35.84	36.53	44.08	43.56	43.45
Intergenes	60.62	60.53	60.42	31.48	31.43	31.49	49.19	49.07	49.13

TABLE 11. Mean values of the percentages of dichotomic classes computed in antisense strand

	$p0a$	$p1a$	$p2a$	$r0a$	$r1a$	$r2a$	$h0a$	$h1a$	$h2a$
Genes	56.29	56.33	56.01	38.60	39.33	39.09	51.04	51.01	50.59
Exons	54.85	55.92	53.29	40.75	45.28	42.76	49.09	49.27	47.06
CDS	54.61	56.33	51.06	39.05	49.32	43.64	48.66	49.39	44.14
Introns	60.45	61.58	61.74	29.38	26.75	28.64	61.01	61.68	61.36
UTR	60.12	59.70	59.71	32.16	32.40	32.19	52.20	52.30	52.65
Regulatory	60.39	59.74	59.18	32.61	33.28	33.15	49.50	48.97	49.13
Intergenes	60.56	60.47	60.54	31.56	31.56	31.53	49.12	49.09	49.14

the information contained in the dichotomic classes can reveal such role. In the following we summarize the results for the three dichotomic classes. Note that the values for genes can be considered as the weighted mean of coding and non-coding sequences.

Parity. The percentages of parity are similar for introns, intergenes, UTR and regulatory sequences (the range is from 59% to 61%). There are not big differences for the three frames. Also, the genes show the same behaviour as non coding sequences but the mean values are different (55%-56%). CDS and Exons show a lower value of

$p0$ and $p1$ (CDS: 50%-51%; Exons: 52%) and an higher value of $p2$ (CDS: 58.67%; Exons: 56.26%). We can summarize the last result as follows:

$$p0 < p1 < p2$$

Rumer. As concerns Rumer's class, non-coding sequences (introns, IG, UTR and regulatory) show almost the same values in and out of frame. UTR and regulatory sequences show values between 35% and 36%, while introns vary within 32.82% and 34.85%. Intergenes seems to have a specific Rumer value around 31.5%. Genes show the same behaviour as non coding sequences but the values are around 39%. We can summarize the result of CDS as follows:

$$r2 < r1 < r0$$

Note that the values for CDS are remarkably higher than those of non-coding sequences.

Hidden. Non-coding sequences (introns, IG, UTR and regulatory) show almost the same values both in and out of frame but with specific values for each sequence class: introns ($\sim 38\%$) , IG ($\sim 49.1\%$), UTR and regulatory($\sim 44\%$). Once again, coding sequences show different patterns across the frames:

$$h2 < h0 < h1$$

For what concerns exons we can see that $h0$ is very similar to $h1$ (50.58% and 51.44%) but higher than $h2$ (47.41%). As for the Rumer's class, genes show the same behaviour as non coding sequences (they don't change with frame).

If we study the dichotomic classes computed on the antisense strand we can observe that coding and non-coding sequences have a similar behaviour. Genes and non coding sequences do not vary with frame while exons and CDS do. The values are different from those of the sense strand except for intergenes. In fact intergenes, surprisingly, show the same percentages observed on the sense strand. As concerns the other non-coding sequences (introns, UTR and regulatory) we can see that parity values are similar to those computed in sense strand, while Rumer and hidden values differ when passing from sense to antisense strand. Finally, CDS and exons show again a frame-related behaviour:

$$p2a < p0a < p1a$$

$$r0a < r2a < r1a$$

$$h2a < h0a < h1a$$

The above analysis has been repeated on the set of transformed sequences:

- complementary sequences
- reverted sequences
- sequences undergone to Keto/Amino global transformation
- sequences undergone to Purine/Pyrimidine global transformation

The results obtained show a similar behaviour for all the kind of genome portions considered. The whole analysis can be found in [18].

TABLE 12. Partition of the 16 dinucleotides into parity groups.

i	1^{st} base (b_1)	2^{nd} base (b_2)	Dinucleotide group (D)	parity
1	A	A	S_1	1
2	C	A		
3	G	A		
4	T	A		
5	G	T		
6	T	T		
7	G	C		
8	T	C		
9	A	G	S_2	0
10	C	G		
11	G	G		
12	T	G		
13	A	T		
14	C	T		
15	A	C		
16	C	C		

4.1. **Independence test.** In this section we test the hypothesis that the dichotomic classes are expression of an underlying organization of the genetic information. This would imply the existence of a correlation structure in the nucleotide sequences so that the percentages observed are not compatible with the hypothesis of an independent random process. Hence, under the hypothesis of stochastic independence of the sequence, the percentages of dichotomic classes would depend only on the proportions of bases. For instance, if we consider parity, define the random variable X as the parity of a dinucleotide for a given sequence. Then, X follows a Bernoulli distribution with parameter $\pi = P(X = 1)$, that is: $E(X) = \pi$ and $V(X) = \pi(1 - \pi)$. Then, we have the following null hypothesis:

$$H_0 : \pi = \pi_0$$

$$H_1 : \pi \neq \pi_0$$

where $\pi_0 = P(X = 1)$ under the assumption that the DNA sequence is a realization of a *i.i.d.* process. Now, we can see that:

$$\pi_0 = P(X = 1) = P(D) \in S_1 = \sum_{i=1}^{8} P(b_1)_i * P(b_2)_i \tag{1}$$

where $P(b_1)$ and $P(b_2)$ are the probability of occurrences of the 4 nucleotides (T,C,A,G) in the first and second base, respectively. The association scheme for the parity is presented in table 12. Therefore, the possible differences observed between the original and *i.i.d.* sequences are not due to the proportion of bases and all the quantities derived from it (e.g. the GC content and the like).
For example, if we have the following probability distribution for the nucleotides:

base	P
A	0.30
C	0.25
G	0.20
T	0.25

then

$$\pi_0 = 0.30 \times 0.30 + 0.25 \times 0.30 + 0.20 \times 0.30 + 0.25 \times 0.30 + 0.20 \times 0.25 + \\ + 0.25 \times 0.25 + 0.20 \times 0.25 + 0.25 \times 0.25 = 0.525$$

Now, if we take the usual sample mean $\hat{\pi}$ as the estimator of π we have that under the null hypothesis, $E(\hat{\pi}) = \pi_0$ and $V(\hat{\pi}) = \frac{\pi_0(1-\pi_0)}{n}$ where n is the length of binary sequence. Thus, we can use the test statistic Z

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} .$$

Z converges in distribution to a standard normal random variable so that the usual critical values can be used.

We have computed the p-values associated to the test for each sequence and for each dichotomic class. The results are shown in Figures 5 and 6, where we present the histograms of the p-values for the sequences analyzed. By looking at the histograms, we can see that only coding sequences and genes show patterns which indicate the departure from the *i.i.d.* hypothesis. Only these classes, in fact, show an important proportion of p-values lower than 0.05 (see Tab. 13). This suggests that only genes, exons and CDS show an informational structure that depends on the frame but is not related to the proportion of bases.

TABLE 13. Percentages of p-values lower than 0.05

	parity	Rumer	hidden
Genes	17.94	13.24	8.93
CDS	28.28	14.96	5.80
Exons	11.45	5.80	4.00
Introns	4.31	1.50	1.20
Intergenes	9.09	3.13	1.70
UTR	6.27	1.98	1.43
Regulatory	5.65	2.85	1.96

The results for parity seem to be more informative than those for other classes. In fact, they show a higher rate of low p-values. This trend can be observed also in non-coding sequences where parity p-values seem to follow an uniform distribution while Rumer and hidden show fewer low p-values. Finally, if we compare the histograms of exons and CDS we can see that the latter show an higher rate of low p-values. This effect could be related to the sample size. In order to investigate the matter in detail, it might be interesting to join all the sequences of the same kind as to give more discriminatory power to the test and reduce the probability of false rejections due to multiple testing. The latter issue might be also studied by resorting to adjusted p-values and to false discovery rate estimation (see e.g. [1]). The overall results confirm the findings of [8] regarding the presence of correlations between dichotomic classes in coding sequences.

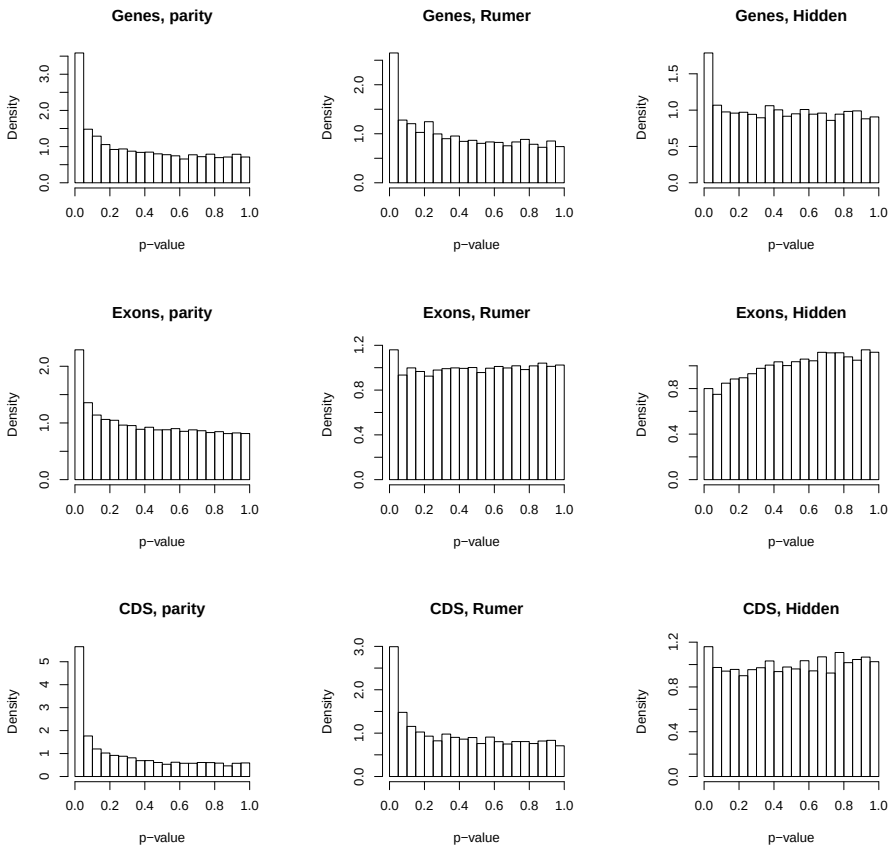

FIGURE 5. Histograms of the p-values associated to the independence test for coding sequences

5. **Conclusions.** In this work we have shown how a recently developed mathematical model for the genetic code can be used as a tool to characterize different parts of the genome. In the first part, we have reviewed the main mathematical features of the model and its symmetry properties. Then, we have analyzed the whole chromosome 1 of *Arabidopsis thaliana* by creating specific routines that, by using genome annotations, extract and build seven groups of sequences: genes, exons, introns, coding sequences (CDS), intergenes, untranslated regions (UTR) and regulatory sequences. The sequences have been encoded according to the definitions of dichotomic classes, binary variables that derive naturally from the mathematical structure and that are related to the chemical properties of the sequences. We studied the percentages of the three classes in and out of frame for the whole dataset and used this information to discriminate between the seven group of sequences.

Since the percentages of dichotomic classes vary with frame only for coding sequences, we can conjecture that frame is important only for these kind of sequences and dichotomic classes can be useful to recognize them. All the dichotomic classes can distinguish between coding and non-coding sequences; in fact, the mean values are always different. In particular, we can see that parity could discriminate

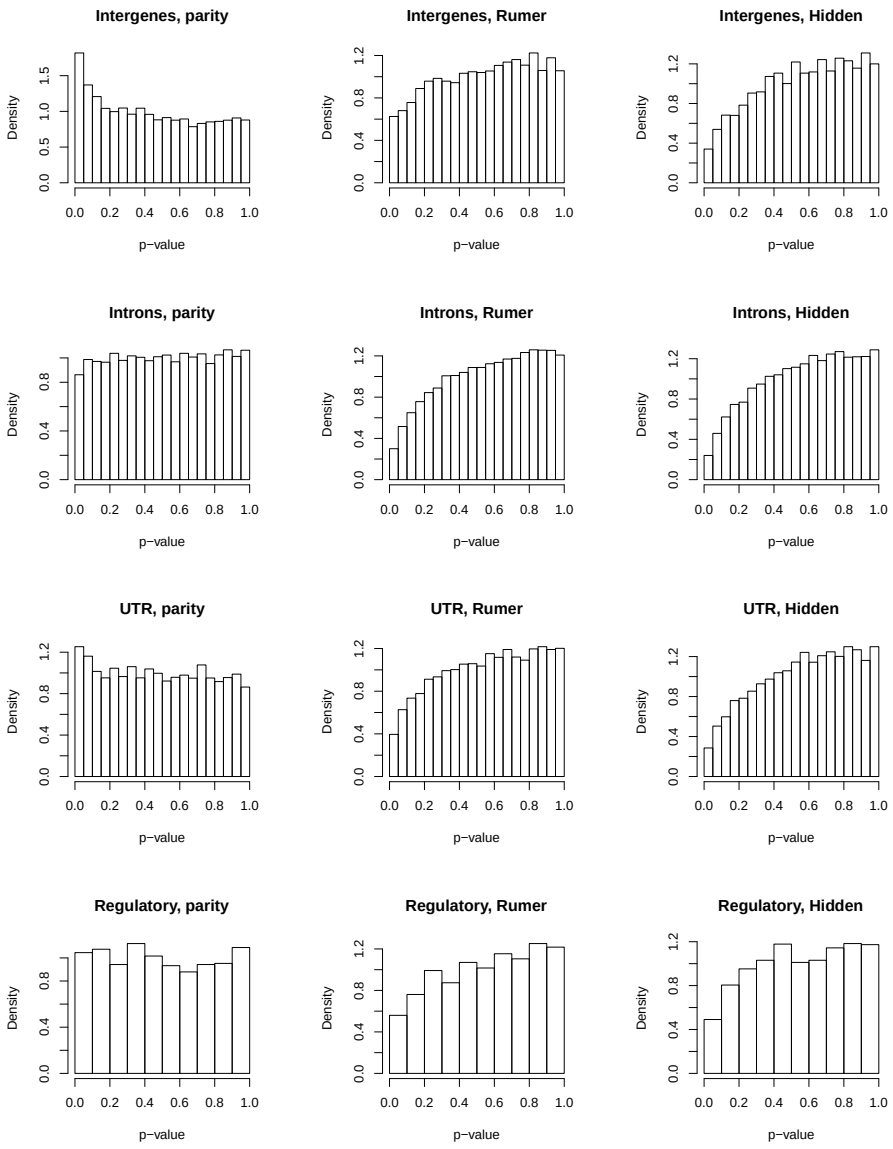

FIGURE 6. Histograms of the p-values associated to the independence test for noncoding sequences

also between sense and antisense coding sequences as the mean percentage is always around 60% for non-coding sequences (in both strands), while it is remarkably lower and strand dependent for CDS and exons. Since Rumer and hidden percentages vary both between non-coding sequences and between sense and antisense strand, they are useful to discriminate between the different classes of non-coding sequences (i.e: introns, intergenes, UTR and regulatory).

Finally, as for coding sequences, we have seen that the mean values of CDS differ with frame more than those of exons. For example, if we consider parity, we can

see that $p0$ and $p1$ are almost the same for exons ($p0 = 52.1\%$, $p1 = 52.7\%$) while for CDS we have $p0 = 50.1\%$, $p1 = 51.6\%$. Moreover the mean for $p2$ is higher for CDS than for exons (58.7% and 56.3% respectively). This kind of difference can be observed in Rumer and hidden proportions too. The results obtained from the independence tests show that the framework suggested by dichotomic classes is able to uncover the existence of significant correlations in those sequences that are involved in protein synthesis.

Indeed, the existence of a mechanism for error correction/detection linked to the replication and translation processes imply some kind of dependence inside DNA sequences. Several studies have highlighted the presence of fractal long-range correlations in nucleotide sequences. However, error detection and correction should act at a local level and should discriminate between different portions of a gene.

No doubt, further studies are needed in order to assess how the information carried by dichotomic classes could discriminate between coding and noncoding sequence and, therefore, contribute to unveil the role of the mathematical structure in error detection and correction mechanisms. Still, we have shown the potential of the approach presented for the understanding the management of genetic information. We believe that this approach could help to keep the promises and hopes related to molecular biology and the Human Genome Project.

Acknowledgments. We would like to thank the Editor Alberto d'Onofrio for his support. This work has been partially funded by MIUR.

REFERENCES

[1] B. Efron, "Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction," Cambridge University Press, Cambridge, 2010.

[2] G. Elgar and T. Vavouri, *Tuning in to the signals: Noncoding sequence conservation in vertebrate genomes*, Trends in genetics, **24** (2008), 344–352.

[3] A. Elzanowski and J. Ostell, *The genetic codes*, National Center for Biotechnology Information (NCBI), (2008-04-07). Retrieved 2010-03-10.

[4] D. L. Gonzalez, *Can the genetic code be mathematically described?*, Medical Science Monitor, **10** (2004), 11–17.

[5] D. L. Gonzalez, *Error detection and correction codes*, in "The Codes of Life: the Rules of Macroevolution, volume 1 of Biosemiotics. Chapter 17" (eds. M. Barbieri and J. Hoffmeyers), Springer Netherlands, (2008), 379–394.

[6] D. L. Gonzalez, *The mathematical structure of the genetic code*, in "The Codes of Life: the Rules of Macroevolution, volume 1 of Biosemiotics. Chapter 8" (eds. M. Barbieri and J. Hoffmeyers), Springer Netherlands, (2008), 111–152.

[7] D. L. Gonzalez, S. Giannerini and R. Rosa, *Detecting structures in parity binary sequences: Error correction and detection in DNA*, IEEE Engineering in Medicine and Biology Magazine, **25** (2006), 69–81.

[8] D. L. Gonzalez, S. Giannerini and R. Rosa, *Strong short-range correlations and dichotomic codon classes in coding DNA sequences*, Physical review E, **78** (2008), 051918.

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

[10] D. L. Gonzalez, S. Giannerini and R. Rosa, *Circular codes revisited: A statistical approach*, Journal of Theoretical Biology, **275** (2011), 21–28.

[11] S. Giannerini, D. L. Gonzalez and R. Rosa, *DNA, frame synchronization and dichotomic classes: a quasicrystal framework*, Philosophical Transactions of the Royal Society. Series A, **370** (2012), 2987–3006.

[12] D. L. Gonzalez and M. Zanna, *Una nuova descrizione matematica del codice genetico*, Systema Naturae, Annali di Biologia Teorica, **5** (2003), 219–236.

[13] International Human Genome Sequencing Consortium, *Initial sequencing and analysis of the human genome*, Nature, **409** (2001), 860–921.

[14] A. G. Jegga and B. J. Aronow, *Evolutionary conserved noncoding DNA*, in "Encyclopedia of Life Sciences," John Wiley & sons, (2006).
[15] S. Ohno, *So much "junk" DNA in our genome*, Brookhaven Symposia in Biology, **23** (1972), 366–370.
[16] H. Pearson, *Genetics: What is a gene?*, Nature, **441** (2006), 398–401.
[17] E. Pennisi, *Genomics. DNA study forces rethink of what it means to be a gene.*, Science (New York, N. Y.), **316** (2007), 1556-1-557.
[18] E. Properzi, "Genome Characterization Through the Mathematical Structure of the Genetic Code: An Analysis of the Whole Chromosome 1 of *A. Thaliana*," PhD Thesis, University of Bologna.
[19] M. Quimbaya, K. Vandepoele, E. Rasp, M. Matthijs, S. Dhondt, G. T. Beemster, G. Berx and L. De Veylder, *Identification of putative cancer genes through data integration and comparative genomics between plants and humans*, Cell. Mol. Life Sci., **69** (2012), 2041–2055.
[20] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, (2012), http://www.R-project.org/.
[21] The Arabidopsis Genome Initiative, *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana*, Nature, **408** (2000), 796–815.
[22] TAIR, *Genome Annotation*, http://www.arabidopsis.org/
[23] O. Trapp, K. Seeliger and H. Puchta, *Homologs of breast cancer genes in plants*, Front. Plant Sci., **2** (2011).
[24] J. C. Venter et al., *The sequence of the human genome*, Science, **291** (2001), 1304–1351.
[25] K. Watanabe and T. Suzuki, "Genetic Code and its Variants," in "Encyclopedia of Life Sciences," John Wiley & sons, 2006.

Received May 09, 2012; Accepted September 04, 2012.

E-mail address: enrico.properzi3@unibo.it
E-mail address: simone.giannerini@unibo.it
E-mail address: gonzalez@bo.imm.cnr.it
E-mail address: rodolfo.rosa@unibo.it