Dichotomic classes, correlations and entropy optimization in coding sequences

Simone Giannerini[1]

[1]Università di Bologna, Dipartimento di Scienze Statistiche

Joint work with Diego Luis Gonzalez and Rodolfo Rosa
First International Conference in Code Biology, Paris, 20-24 May 2014

The mathematical model of the genetic code

Unique solution: $1, 1, 2, 4, 7, 8$

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

Matching the symmetries of the genetic code with those of the mathematical representation allows to assign the 64 binary strings to the codons and integer number from 0 to 23 to the amino acids.

Parity of the strings

Surprisingly,
the mathematical properties of the model have a counterpart on the genetic code.

The parity of a binary string, denoted as c_1, is defined as the parity of its sum:

$$c_1 = \left(\sum_{i=1}^{6} d_i \right) \mod 2; \quad \text{e.g. } 1\ 1\ 0\ 0\ 0\ 1 \text{ has 3 ones} \rightarrow \text{odd}$$

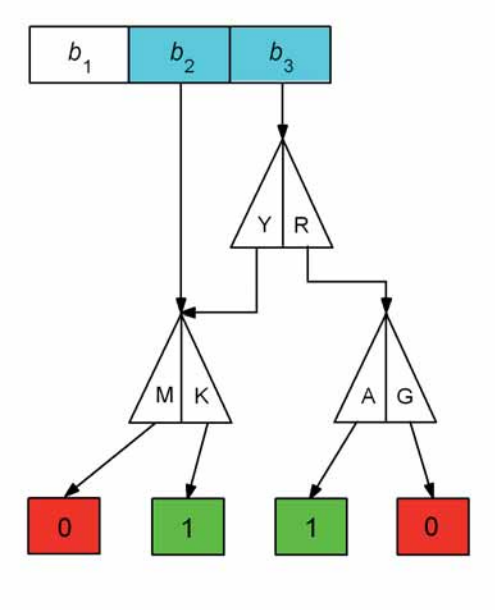
#	8 7 4 2 1 1	8 7 4 2 1 1	8 7 4 2 1 1	8 7 4 2 1 1	D	Amino acids pairs	8 7 4 2 1 1	8 7 4 2 1 1	8 7 4 2 1 1	8 7 4 2 1 1	#
0	0 0 0 0 0 0				1	W Trp M Met				1 1 1 1 1 1	23
1	0 0 0 0 1 0	0 0 0 0 0 1			2	S Ser 2 F Phe			1 1 1 1 1 0	1 1 1 1 0 1	22
2	0 0 0 1 0 0	0 0 0 0 1 1			2	Ter K Lys			1 1 1 1 0 0	1 1 1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

Dichotomic classes: parity

- Each base — T,C,A,G — can be classified according to chemical classes:

$$\begin{array}{rll} \{\textit{Purine; Pyrimidine}\} & \{\mathrm{R} = \mathrm{A},\ \mathrm{G}; & \mathrm{Y} = \mathrm{C},\ \mathrm{T}\ \} \\ \{\textit{Keto; Amino}\} & \{\mathrm{K} = \mathrm{T},\ \mathrm{G}; & \mathrm{M} = \mathrm{A},\ \mathrm{C}\ \} \\ \{\textit{Strong; Weak}\} & \{\mathrm{S} = \mathrm{C},\ \mathrm{G}; & \mathrm{W} = \mathrm{A},\ \mathrm{T}\ \} \end{array}$$

- The parity of the strings can be described in terms of the biochemical properties of the codons.



Dichotomic classes: Rumer's class - 1

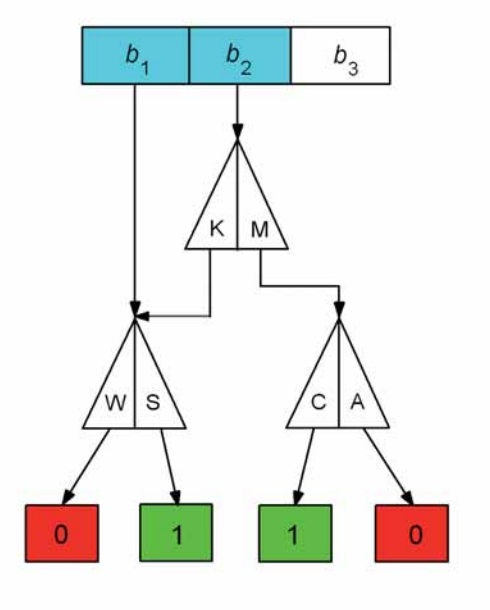
	U			C			A			G			
U	1	000001	Phe	14	011110	Ser	5	001001	Tyr	7	001101	Cys	U
	1	000010	Phe	14	011101	Ser	5	001010	Tyr	7	001110	Cys	C
	4	000111	Leu	14	101100	Ser	21	111011	Ter	7	010000	Cys	A
	11	011000	Leu	14	101011	Ser	21	111100	Ter	0	000000	Trp	G
C	11	100101	Leu	8	010010	Pro	3	000101	His	12	011010	Arg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

Discovered in the 60s by the Russian physicist Rumer.

- Green = degeneracy 4
- White = degeneracy $\neg 4$.

Dichotomic classes: Rumer's class - 2

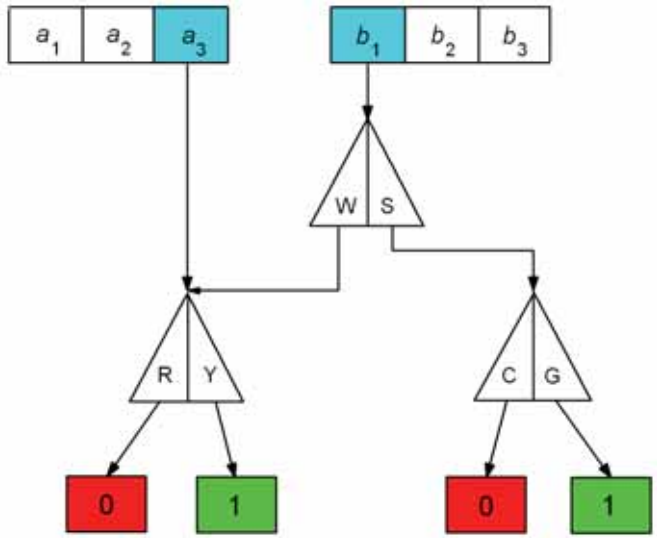
- Also Rumer's class can be derived with a similar algorithm. The first two bases of the codons are involved.



- Rumer's class can be derived from the parity of the first 5 digits of the string.

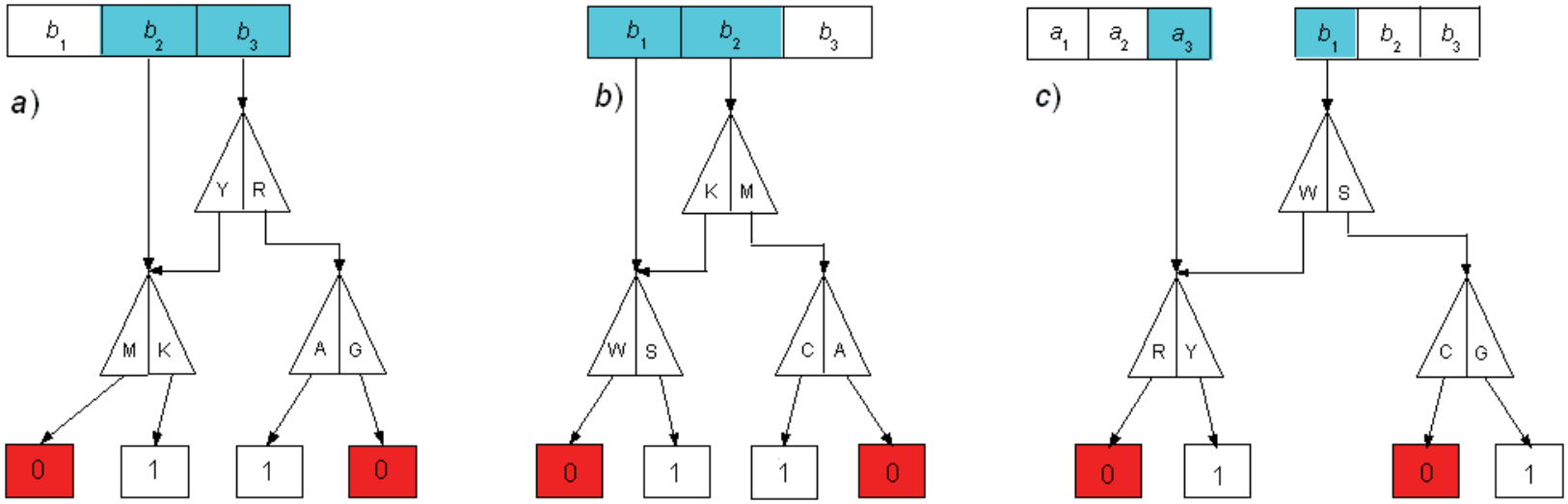
Dichotomic classes: hidden class

If we apply the same reasoning and shift the algorithm we obtain another class: the hidden class



- The hidden class connects two adjacent codons.

Dichotomic classes and transformations



There are $3 + 1$ possible global transformations of a codon:

	from	to	class
KM	T,C,A,G	G,A,C,T	Rumer
YR	T,C,A,G	C,T,G,A	parity
SW	T,C,A,G	A,G,C,T	hidden
I	T,C,A,G	T,C,A,G	

Each transformation is antisymmetric w.r.t. a specific dichotomic class.

Dichotomic classes: a group framework

Denote the bases with the vector notation:

$$\mathrm{T}' = (1000) \qquad \mathrm{C}' = (0100) \qquad \mathrm{A}' = (0010) \qquad \mathrm{G}' = (0001)$$

The transformations of the bases can be implemented by the usual matrix product together with the following permutation matrices:

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

$\{\Gamma, *\}$, where $\Gamma = \{L, M, N, I\}$, is an Abelian (commutative) group isomorphic to the Klein V group $(Z_2 \otimes Z_2)$.

In fact, for each $x, y, z \in \Gamma$ we have

1. I is the neutral element
2. $x * x = I$ (indeed, L, M, N, I are orthogonal);
3. $x * (y * z) = (x * y) * z$ (associativity)
4. $x * y = y * x = z$ (commutativity and closure)

Dichotomic classes as nonlinear operators

define the following matrices:

$$O_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 2 & 2 & 1 \\ 0 & 0 & 3 & 4 \end{pmatrix}; \quad O_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 2 \\ 0 & 4 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad O_3 = \begin{pmatrix} 2 & 1 & 1 & 2 \\ 0 & 4 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The classes $c_1 =$ parity, $c_2 =$ Rumer, $c_3 =$ hidden can be obtained as follows:

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

where:

- Q is a 4×4 matrix that represents 4 contiguous bases
- $\odot$ denotes the matrix Hadamard product
- $\|Q\|_\infty$ is the infinite order matrix norm for a $m \times m$ square matrix Q: $\|Q\|_\infty = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{m} |q_{ij}|$

The dichotomic classes c_i are nonlinear functions of Q

Dichotomic classes: an example of coding

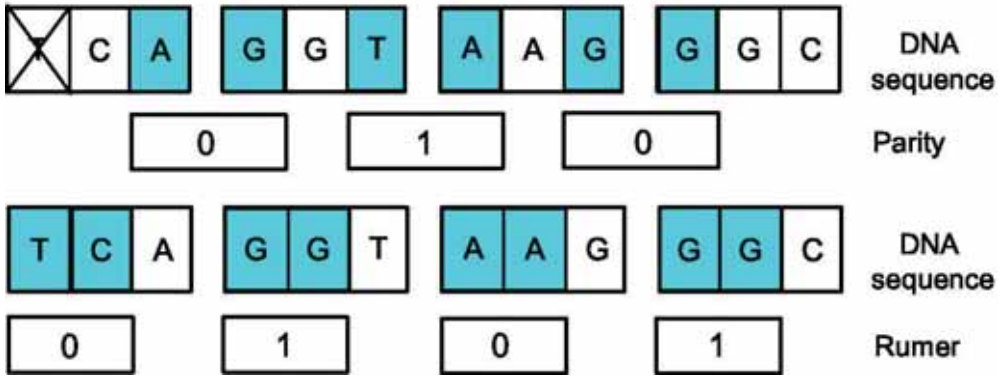
Given the sequence

TCA GGT AAG GGC

we have three possible reading frames:

frame 0	TCA GGT AAG GGC
frame 1	CA GGT AAG GGC
frame 2	A GGT AAG GGC

below we compute the parity on the frame 1 sequence and Rumer's class in frame:



The same analysis can be applied to the complementary reversed sequence

GCC CTT ACC TGA

Error correction and time series

- Redundancy and parity coding are the main ingredients of man made error detection and correction systems;
- The existence of a coding mechanism for error correction/detection implies some kind of dependence inside data;
- If all the genetic information share a common error correction machinery this should imply the emergence of common structures.

- 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.

From a time series perspective this poses several issues:

1. Are there such (universal) correlations that can be found in every sequence?
2. Is the mathematical structure playing a role?

Dichotomic classes and dependence

Is there a dependence structure in the dichotomic classes?

Given two sequences X_t and Y_t we have:

$$\begin{cases} H_0 : X_t \text{ and } Y_{t+k} \text{ are independent} \\ H_1 : X_t \text{ and } Y_{t+k} \text{ are not independent} \end{cases} \qquad \text{for } k \in \mathbb{Z}$$

Problem: build a valid test. We need:

- A suitable measure of dependence;
- A scheme for testing H_0 by taking into account repeated testing issues.

A cross entropy metric

We use a normalized version of the Bhattacharya-Hellinger-Matusita distance:

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

- $f_{X_t}(x)$ pdf of X_t;
- $f_{Y_{t+k}}(y)$ pdf of Y_{t+k};
- $f_{(X_t, Y_{t+k})}(x,y)$ joint pdf of (X_t, Y_{t+k});

- Reduces to a measure of serial dependence if $Y_t = X_t$;
- $S_\rho(k)$ possesses many desirable theoretical properties;

The testing scheme

Issues

- The dichotomic classes are naturally correlated because they can be computed on the same bases.
- Spurious correlations due to nonstationarity/different GC content.

Because of such issues simple nonparametric bootstrap schemes that resample the binary sequences are not appropriate.

Solution: a modified permutation scheme

Given a nucleotide sequence Z_t

1. on Z_t compute the two dichotomic classes X_t and Y_t
2. compute the measure on X_t and Y_{t+k}: $\hat{S}_k$
3. draw Z_t^*, a random permutation of Z_t
4. on Z_t^* compute the two dichotomic classes X_t^* and Y_t^*
5. compute the measure on X_t^* and Y_{t+k}^*: $\hat{S}_k^*$
6. repeat steps $3-5$ B times.
7. compare $\hat{S}_k$ with the quantiles of the distribution of $\hat{S}_k^*$.

The single test case

We wish to test a single null hypothesis H_0.
We set the significance level α and reject H_0 if the p-value of the test is smaller than α.

		Test		
		H_0	H_1	
Truth	H_0	$1-\alpha$	α	1
	H_1	β	$1-\beta$	1

- $\alpha = P(\text{reject } H_0 | H_0 \text{ is true})$ Type I error
- $\beta = P(\text{accept } H_0 | H_0 \text{ is false})$ Type II error

The multiple test case

We wish to test N null hypotheses H_{0i} , $i = 1 \ldots, N$.
N can be of the order of tens of thousands.

		Test		
		H_0	H_1	
Truth	H_0	$N_0 - a$	a	N_0
	H_1	$N_1 - b$	b	N_1
		$N - R$	R	N

- Of the N_0 null cases a are rejected incorrectly (false discoveries);
- Of the N_1 non-null cases b are rejected correctly (true discoveries);
- a/R is the *false discovery proportion*;

Solutions:

- Bonferroni bound: controls FWER:

$$\text{FWER} = P(\text{reject any true } H_{0i}) = P(a > 0)$$

- Benjamini and Hochberg prodcedure: controls Fdp

$$E(\text{Fdp}) = E\left(\frac{a}{R}\right)$$

The multiple test case - 2

1. Bonferroni bound: given a significance level α reject those hypotheses for which:

$$p_i \leq \alpha/N$$

   A theorem assures that FWER $\leq \alpha$. Problem: too conservative.

2. Benjamini and Hochberg's FDR control algorithm BH(q):
   - we have a decision rule that produces a p-value p_i for each test, $i = 1, \ldots, N$.
   - If H_{0i} is true then: $p_i \sim \mathcal{U}(0, 1)$
   - order the p-values:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(i)} \leq \cdots \leq p_{(N)}$$

   - for a fixed value of $q \in (0, 1)$ let $i_{\max}$ the largest index for which

$$p_{(i)} \leq \frac{i}{N} q \tag{2}$$

     reject H_{0i} if $i \leq i_{\max}$
   - Under the hypothesis of independence of the p-values we have:

$$E(\text{Fdp}) = \pi_0 q \leq q$$

     where $\pi_0 = N_0/N$

The empirical Bayes interpretation of the BH(q) procedure

Consider the p-values $p_i, i = 1, \ldots, N$:

$$p_i = F_0(z_i) \quad \text{left tail} \tag{3}$$

$$p_i = 1 - F_0(z_i) \quad \text{right tail} \tag{4}$$

where $F_0(z_i)$ is the cdf under the null. In our case F_0 is $\mathcal{U}(0,1)$ so that $p_i = z_i$.
Order the z-values:

$$z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(i)} \leq \cdots \leq z_{(N)}$$

Note that the empirical cdf satisfies:

$$\bar{F}(z_{(i)}) = i/N$$

We can write the BH rule (2) as

$$\frac{F_0(z_{(i)})}{F(z_{(i)})} \leq q \tag{5}$$

$$\overline{\text{Fdr}}(z_{(i)}) = \pi_0 \frac{F_0(z_{(i)})}{F(z_{(i)})} \leq \pi_0 q \tag{6}$$

The BH rule can be rewritten as follows: reject H_i if $z_i > z_{\text{max}}$ where

$$z_{\text{max}} = \sup_z \left\{ \overline{\text{Fdr}}(z) < q \right\} \tag{7}$$

The dataset: KOGs clusters of predicted orthologs

We have analyzed 458 KOG sequences for each of the six genomes. KOGs are clusters of predicted orthologs (eukaryotic orthologous groups).

In other words, sequences of different species associated to the same KOG are functionally homologous.

Table: Classes of organisms analysed. The third column reports the number of kilobases (kb) of each class.

	Organism	kb
1	Homo sapiens	553.901
2	Drosophila melanogaster	557.970
3	Arabidopsis thaliana	561.582
4	C. elegans	552.873
5	Saccharomyces cerevisiae	564.831
6	Schizosaccharomyces pombe	551.130

We have grouped the data

- by KOG $\rightarrow$ 458 sequences of average length 7.3 kb.

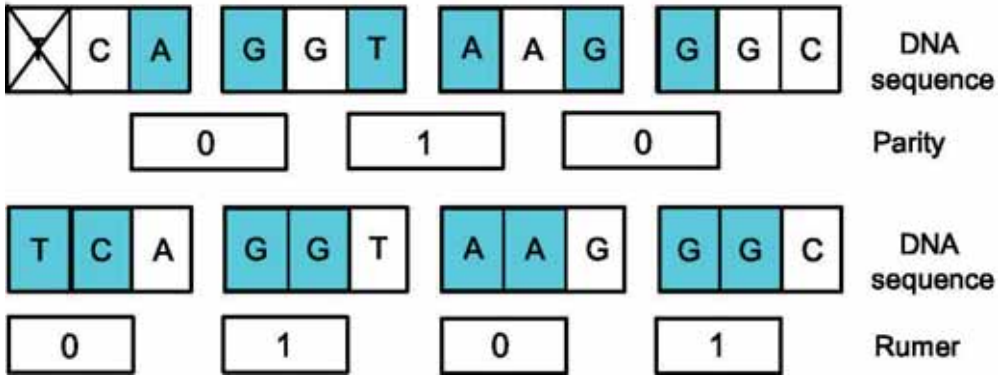
The dataset: some notation

For each sequence we have 18 dichotomic classes in the 3 reading frames.

Table: Legend

class	frame	anticodon
p = parity r = Rumer h = hidden	frame 0 frame 1 frame 2	a = reversed complement

Example: the combination `p1-r0`

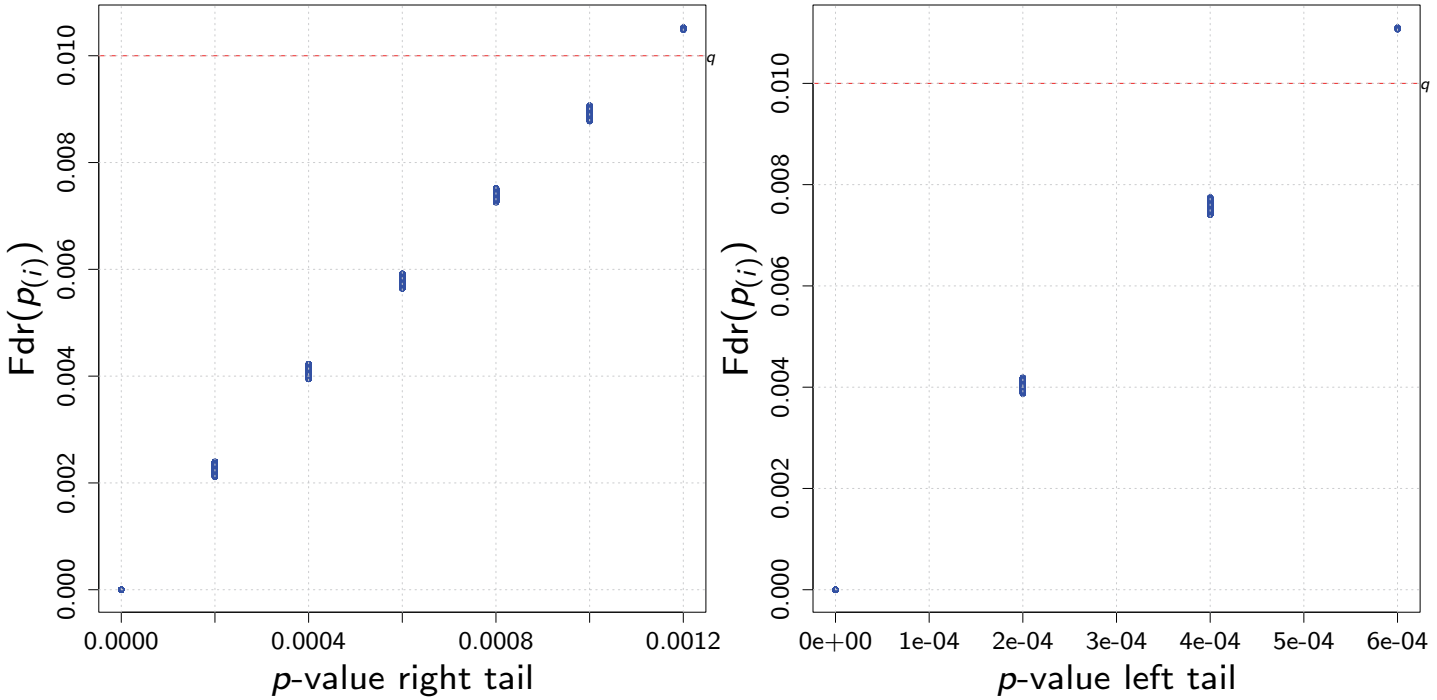


- `p1-r0` at lag 0 involves bases 34 and 12
- `p1-r0` at lag 1 involves bases 34 and 56
- `p1-r0` at lag -1 involves bases 67 and 12

Results: bivariate (cross entropy)

- we set $q = 0.01$. Is the estimate of the Bayes probability that a rejected null is is actually null.
- The number of valid combinations of dichotomic classes is 153.
- The lags tested are three: $-1, 0, 1$
- overall, we have $N = 153 \times 3 \times 458 = 210222$ simultaneous tests.
- B = 5000 bootstrap replications.

Plot of the estimated Fdr vs p-values



- BH(q) threshold p-values: 0.001 (right tail) and 0.0004 (left tail)
- Independence of the tests is not required and affects only the accuracy of the estimation of the Fdr. $\overline{Fdr}$ is still an unbiased estimator of $Fdr(z)$.

Results: bivariate (cross entropy) – right tail rejections

- Percentages of rejections over the 458 KOG sequences.

cnames	lags -1	0	1
h0a-h2a	76.4	16.6	1.3
p0a-p2	69.2	2.0	3.9
h0-h1a	7.2	69.4	3.1
h0-p1	3.7	88.4	1.3
h0-r2a	7.4	93.2	2.0
h0a-h1a	9.0	63.8	2.4
h0a-r1	2.0	97.4	29.5
h1-h2	2.2	66.6	1.5
h1a-p1	3.3	84.5	2.0
h1a-p2	2.8	83.8	2.6

cnames	lags -1	0	1
h1a-r2a	6.3	88.6	5.0
h2-h2a	1.3	87.3	1.7
p0-p0a	2.4	62.0	1.3
p1-r2a	7.2	97.4	2.2
r1-r1a	7.2	80.6	0.9
h1-r0	4.1	5.9	61.6
h1-r1	0.9	1.7	86.5
h2-r1	0.4	1.7	97.8
h2a-r1	1.5	0.7	64.0
h2a-r1a	1.3	24.7	86.5

Results: bivariate (cross entropy) – right tail rejections 2

- Example:

`h0-h1a` at lag 0
involves bases 34 and 5'6'

```
$`3506`

      0    1
  0 20.3 23.2
  1 25.6 30.9

$`1596`

      0    1
  0 25.7 34.7
  1 21.5 18.2

$`1758`

      0    1
  0 35.2 23.3
  1 24.4 17.0
```

`p1-r2a` at lag 0
involves bases 34 and 3'4'

```
$`1727`

      0    1
  0 24.2 31.9
  1 29.3 14.6

$`3449`

      0    1
  0 29.0 37.7
  1 21.8 11.4

$`1762`

      0    1
  0 26.6 39.4
  1 22.5 11.5
```

More random than random? (1)

Two binary random variables X and Y are stochastically independent iff:

$$P(X, Y) = P(X)P(Y) \quad \text{or } P(Y|X) = P(X)$$

		Y		
		0	1	
X	0	$p_{0\|0}$	$p_{1\|0}$	1
	1	$p_{0\|1}$	$p_{1\|1}$	1
		p_0	p_1	1

- where $p_{i|j} = P(Y = i|X = j)$
- Independence implies that $p_{i|0} = p_{i|1} = p_i$, that is the conditional distributions by row are equal

More random than random? (2) – left tail rejections

- Percentages of rejections over the 458 KOG sequences.

cnames	lags -1	0	1
p0a-r1a	65.7	0.0	0.0
p0a-r2a	86.7	0.2	0.0
r0-r1a	64.6	0.0	0.0
h0-p2a	0.0	84.9	0.0
h0a-p1	0.0	97.2	0.0
h1-p2	0.0	88.4	0.0
h2a-p2	0.0	63.5	0.0
p0-r1	0.0	77.5	0.0
p1-p1a	0.2	78.6	0.0
p1a-r1	0.0	71.4	0.2
p2-r2a	0.0	83.8	0.0
r0a-r1	0.0	62.2	0.0
r1-r2	0.0	60.5	0.0
h1-p0a	0.0	0.2	65.7
h1a-r0	0.0	0.0	75.5
h2-p1	0.0	0.0	68.1
h2-r0a	0.0	0.0	74.0
h2a-p0	0.2	0.2	77.1
p2-r0a	0.0	0.0	85.4

h0a-p1 at lag 0
involves bases 34 and 3'4'

$`0729`

	0	1
0	51.2	48.8
1	51.2	48.8

$`2309`

	0	1
0	55.4	44.6
1	55.4	44.6

$`0556`

	0	1
0	51.8	48.2
1	51.8	48.2

More random than random? (3): an example on gene 0729

`h0a-p1` at lag 0 – involves bases 34 and 3'4' Original sequence

```
        0     1   Sum
  0  51.2  48.8 100.0
  1  51.2  48.8 100.0
```

```
X-squared   p.value
        0         1
```

Randomly permuted sequence

```
        0     1   Sum
  0  60.7  39.3 100.0
  1  36.4  63.6 100.0
```

```
X-squared   p.value
 1.53e+02  3.09e-35
```

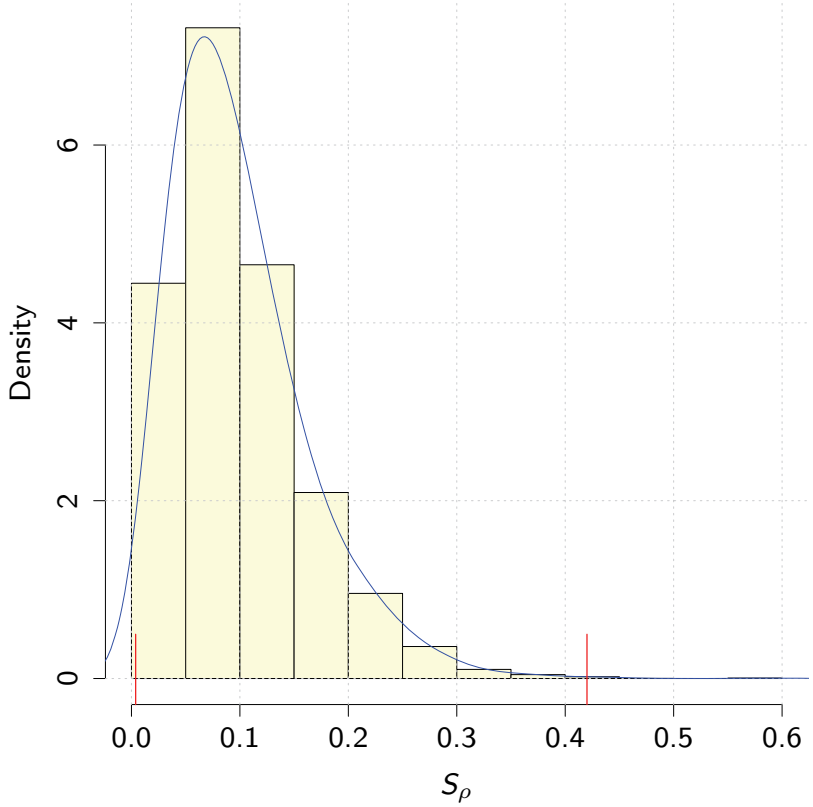
Random synonymous sequence with the same codon usage

```
        0     1   Sum
  0  56.5  43.5 100.0
  1  48.5  51.5 100.0
```

```
X-squared   p.value
 1.58e+01  6.99e-05
```

More random than random? Discussion

Distribution of S_ρ under H_0:



- Right tail rejection implies correlation $\rightarrow$ local structure
- Left tail rejection implies the existence of a global optimization structure
- At positions 34 and 3'4'we have that at the same time the parity class:
  - is **maximally** correlated with Rumer's class
  - is **minimally** correlated with the hidden class.
- Signals with low correlation play an important role in Communication Theory.
- The notions of **resiliency** and **correlation immunity** might be relevant here.

References – The model and its extensions

D. L. Gonzalez.

Can the genetic code be mathematically described?

Medical Science Monitor, 10(4):11–17, 2004.

D. L. Gonzalez.

Error detection and correction codes.

In M. Barbieri and J. Hoffmeyer, editors, *The Codes of Life: The Rules of Macroevolution*, volume 1 of *Biosemiotics*, chapter 17, pages 379–394. Springer Netherlands, 2008.

D. L. Gonzalez.

The mathematical structure of the genetic code.

In M. Barbieri and J. Hoffmeyer, editors, *The Codes of Life: The Rules of Macroevolution*, volume 1 of *Biosemiotics*, chapter 8, pages 111–152. Springer Netherlands, 2008.

D. L. Gonzalez, S. Giannerini, and R. Rosa.

On the origin of the mitochondrial genetic code: towards a unified mathematical framework for the management of genetic information.

Nature Precedings.

References – Dichotomic classes

D. L. Gonzalez, S. Giannerini, and R. Rosa.

Detecting structure in parity binary sequences.

IEEE Engineering in Medicine and Biology Magazine, 25:69–81, 2006.

D. L. Gonzalez, S. Giannerini, and R. Rosa.

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

Physical Review E, 78(5):051918, 2008.

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(3–4):143–157, 2009.

S. Giannerini, D. L. Gonzalez, and R. Rosa.

DNA, frame synchronization and dichotomic classes: a quasicrystal framework.

Phylosophical Transactions of the Royal Society, Series A, Vol. 370, Number 1969, 2987–3006, 2012.

E. Properzi, S. Giannerini, D. L. Gonzalez, and R. Rosa.

Genome characterization through dichotomic classes: an analysis of the whole chromosome 1 of A. Thaliana

Mathematical Biosciences and Engineering, Vol. 10, Number 1, 199–219, 2013.

E. Fimmel, A. Danielli, L. Strüngmann: On dichotomic classes and bijections of the genetic code.

J. Theor. Biol., Vol. 336, 221–230, 2013.

Press coverage

A newspaper article based on our research has been selected by the Atomium Culture consortium (http://atomiumculture.eu) and has been published on the following European newspapers:

- Italy: Il Sole 24 Ore

  `http://www.atomium-culture.ilsole24ore.com/?p=10`

- Spain: El Pais

  `http://www.elpais.com/articulo/sociedad/Counting/on/the/Tree/of/Life/elpepusoc/20110726elpepusoc_13/Tes`

- Germany: Frankfurter Allgemeine Zeitung

  `http://www.faz.net/artikel/C31277/mehrdeutige-nummern-in-der-dna-das-leben-kann-zaehlen-30331235.html`

- Austria: Der Standard

  `http://derstandard.at/1285199352166/Counting-on-the-Tree-of-Life`

- Ireland: Irish Times

  `http://195.7.33.36/newspaper/atomium/2010/2010121335.html`

- Poland: Rzeczpospolita

  `http://www.rp.pl/artykul/567922.html`