Universality and Shannon entropy of codon usage

L. Frappat,[1,*] C. Minichini,[2,†] A. Sciarrino,[2,3,‡] and P. Sorba[4,§]

[1]*Laboratoire d'Annecy-le-Vieux de Physique Théorique LAPTH, CNRS, UMR 5108 associée à l'Université de Savoie, Boîte Postale 110, F-74941 Annecy-le-Vieux Cedex, France*

[2]*Dipartimento di Scienze Fisiche, Università di Napoli "Federico II," Naples, Italy*

[3]*INFN, Sezione di Napoli, Complesso Universitario di Monte S. Angelo, Via Cintia, I-80126 Naples, Italy*

[4]*CERN, Theory Division, CH-1211 Geneva 23, Switzerland*

(Received 28 April 2003; published 24 December 2003)

The distribution functions of codon usage probabilities, computed over all the available GenBank data for 40 eukaryotic biological species and five chloroplasts, are best fitted by the sum of a constant, an exponential, and a linear function in the rank of usage. For mitochondria the analysis is not conclusive. These functions are characterized by parameters that strongly depend on the total guanine and cytosine (*GC*) content of the coding regions of biological species. It is predicted that the codon usage is the same in all exonic genes with the same *GC* content. The Shannon entropy for codons, also strongly dependent on the exonic *GC* content, is computed.

DOI: 10.1103/PhysRevE.68.061910 PACS number(s): 87.10.+e, 02.10.−v

I. INTRODUCTION

In the recent past, some interest has been shown in applying methods of statistical linguistics and information theory for the analysis of DNA sequences,[1] in particular, in investigating whether the frequency distribution of nucleotides or sequences of nucleotides follows Zipf's law [1], and using the Shannon entropy to identify the redundancy or the bias of a nucleotide sequence. Let us recall that, at the end of the 1940s, Zipf remarked that, in natural languages and in many other domains, the distribution function follows an inverse power law, which can be described, denoting by rank $n=1$ the most used word, by $n=2$ the next one, and so on, and with $a>0$, by

$$f_n=\frac{f_1}{n^{\alpha}}. \qquad (1)$$

In [2,3], it was claimed that noncoding sequences of DNA are more similar to natural languages than coding ones, and the Shannon entropy has been used to quantify the redundancy of words. This work raised a debate in the literature (see [4]). In particular, in [5] it was shown that the oligonucleotide frequencies in DNA, in both coding and noncoding sequences, follow a Yule and not a Zipf distribution. Let us recall that the Yule distribution with parameters $a,b,c>0$ is given by [6]

$$f_n=cn^{-a}b^n. \qquad (2)$$

Note that Zipf's law is observed from n ranked random samples of χ^2 distributed variables, as shown in [7]. In a recent work [8] it was argued that Zipf's law is well adapted to represent the abundance of expressed genes, with an exponent $a\approx 1$. However, in [9] the analyzed distributions of gene expressions are well fitted by a family of Pareto distributions.

Indeed, in the literature many have claimed that Zipf's laws are not really power laws. As the main point of our paper is not the analysis of the validity of this law, we will no longer pursue the debate, and we refer the interested reader to the web site on Zipf's law (http://linkage.rockefeller.edu/wli/zipf/), where a large literature (updated to 2001) on the applications of this law in different domains can be found.

Recently, an analysis of the rank distribution for codons, performed in many genes for several biological species, led the authors of [10] to fit experimental data with an exponential function. In particular, by considering separately different coding DNA sequences, they studied the relation between the parameter in the exponential, the frequency of rank 1, and the length of the sequence for different genes. From this very short overview, it follows that the determination of the kind of law followed by the codon rank distribution is extremely interesting in investigations of the nature of the evolutionary process, which has acted upon the codon distribution, i.e., the eventual presence of a bias.

In the last few years, the number of available data for coding sequences has considerably increased, but apparently no analysis using the whole set of data has been performed. Here we present the results of such a study. The main aim of this paper is to show the existence of a universal, i.e., biological species independent, distribution law for codons for the eukaryotic code. As a result of our investigation, we point out that the rank of codon usage probabilities follows a universal law, the frequency function of the rank-*n* codon showing up as a sum of an exponential part and a linear part. Such a universal behavior suggests the presence of general biases, one of which is identified with the total *exonic GC*

[*] Email address: frappat@lapp.in2p3.fr

[†] Email address: minichini@na.infn.it

[‡] Email address: sciarrino@na.infn.it

[§] On leave of absence from LAPTH, Annecy-le-Vieux Cedex, France. Email address: sorba@lapp.in2p3.fr; sorba@cern.ch

[1] DNA is constituted of four bases, adenine (*A*), cytosine (*C*), guanine (*G*), and thymine (*T*), this last one being replaced by uracile (*U*) in messenger RNA. A codon is defined as an ordered sequence of three bases. Coding sequences in DNA are characterized by their constituent codons.

1063-651X/2003/68(6)/061910(12)/$20.00

TABLE I. Values of the best-fit parameters, Eq. (4), for the sample of biological species. Types: vrt vertebrates (6), inv=invertebrates (3), pln=plants (4), fng=fungi (2), bct=bacteria (25).

Type	Species	GC content (%)	α	η	$10^4\beta$	χ^2
vrt	*Homo sapiens*	52.58	0.0214	0.073	1.65	0.0126
pln	*Arabidopsis thaliana*	44.55	0.0185	0.056	1.68	0.0051
inv	*Drosophila melanogaster*	54.03	0.0247	0.081	1.67	0.0089
inv	*Caenorhabditis elegans*	42.79	0.0216	0.064	1.79	0.0063
vrt	*Mus musculus*	52.38	0.0208	0.071	1.57	0.0112
fng	*Saccharomyces cervisiae*	39.69	0.0246	0.069	1.91	0.0127
bct	*Escherichia coli*	50.52	0.0233	0.065	1.91	0.0112
vrt	*Rattus norvegicus*	52.87	0.0222	0.073	1.63	0.0083
pln	*Oryza sativa japonica*	55.84	0.0179	0.073	1.63	0.0211
fng	*Schizosaccharomyces pombe*	39.80	0.0255	0.068	1.98	0.0036
bct	*Bacillus subtilis*	44.32	0.0259	0.084	1.71	0.0241
bct	*Pseudomonas aeruginosa*	65.70	0.0538	0.107	2.76	0.0191
bct	*Mesorhizobium loti*	63.05	0.0416	0.093	2.44	0.0093
bct	*Streptomyces coelicolor* A3	72.41	0.0567	0.098	3.14	0.0456
bct	*Sinorhizobium meliloti*	62.71	0.0359	0.076	2.54	0.0067
bct	*Nostoc* sp. PCC7120	42.36	0.0288	0.098	1.63	0.0140
pln	*Oryza sativa*	54.63	0.0173	0.062	1.59	0.0135
bct	*Agrobacterium tumefaciens* str. C58	59.74	0.0308	0.067	2.43	0.0100
bct	*Ralstonia solanacearum*	67.57	0.0543	0.105	2.87	0.0149
bct	*Yersinia pestis*	48.97	0.0179	0.040	2.17	0.0066
bct	*Methanosarcina acetivorans* str. C24	45.17	0.0228	0.068	1.81	0.0214
bct	*Vibrio cholerae*	47.35	0.0203	0.052	2.02	0.0100
bct	*Escherichia coli* K12	51.83	0.0250	0.065	2.05	0.0117
bct	*Mycobacterium tuberculosis* CDC1551	65.77	0.0401	0.094	2.35	0.0105
bct	*Mycobacterium tuberculosis* H87Rv	65.90	0.0414	0.097	2.29	0.0109
bct	*Bacillus halodurans*	44.32	0.0263	0.100	1.27	0.0233
bct	*Clostridium acetobutylicum*	31.59	0.0434	0.087	2.76	
bct	*Caulobacter crescentus* CB15	67.68	0.0570	0.113	2.86	0.0087
vrt	*Gallus gallus*	52.11	0.0239	0.095	1.17	0.0129
bct	*Synechocystis* sp. PCC6803	48.56	0.0260	0.083	1.49	0.0140
bct	*Sulfolobulus solfataricus*	36.47	0.0290	0.066	2.26	0.0099
bct	*Mycobacterium leprae*	59.90	0.0252	0.071	1.80	0.0065
bct	*Brucella melitensis*	58.25	0.0294	0.067	2.25	0.0121
bct	*Deinococcus radiodurans*	67.24	0.0481	0.098	2.76	0.0113
vrt	*Xenopus laevis*	47.33	0.0193	0.084	0.92	0.0268
bct	*Listeria monocytogenens*	38.39	0.0437	0.136	1.64	0.0267
pln	*Neurospora crassa*	56.17	0.0241	0.086	1.31	0.0166
bct	*Clostridium perfrigens*	29.47	0.0510	0.092	3.11	
inv	*Leishmania major*	63.36	0.0294	0.069	2.21	0.0050
vrt	*Bos taurus*	53.05	0.0240	0.089	1.27	0.0126

content. Indeed, the values of the parameters appearing in the fitting expression are plotted versus the total percentage of exonic *GC* content of the biological species and are reasonably well fitted by a parabola. Finally, from the expression obtained, we derive the theoretical prediction that the usage probability for *rank-ordered* codons is the same in any gene region having the same exonic *GC* content for any biological species.

We compute the Shannon entropy [11] for amino acids and find that its behavior as a function of the exonic *GC* content is also a parabola, whose apex is around the value 0.50 of the *GC* content.

II. CODON USAGE PROBABILITY DISTRIBUTION

Let us define the usage probability for the codon *XZN* $(X,Z,N\in\{A,C,G,U\})$ as

$$P(XZN)=\lim_{n_{\mathrm{tot}}\to\infty}\frac{n_{XZN}}{N_{\mathrm{tot}}}, \tag{3}$$

where n_{XZN} is the number of times the codon *XZN* has been used in the analyzed biosynthesis process for a given biological species, and N_{tot} is the total number of codons used in all processes considered. It follows that our analysis and predic-

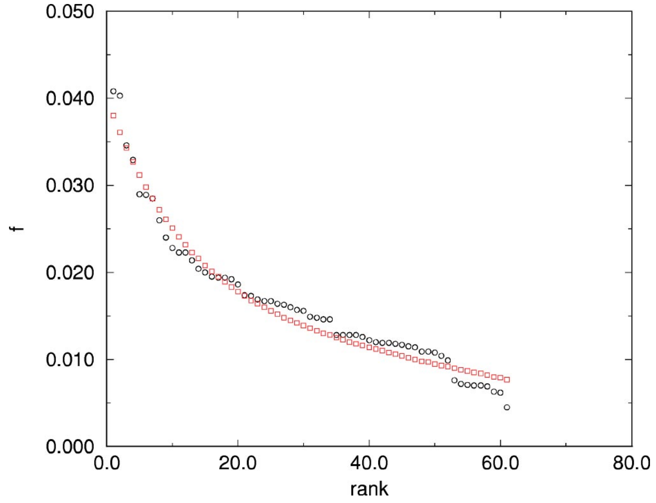


FIG. 1. Rank distribution of the codon usage probabilities for *Homo sapiens*. Circles are experimental values, squares are fitted values.

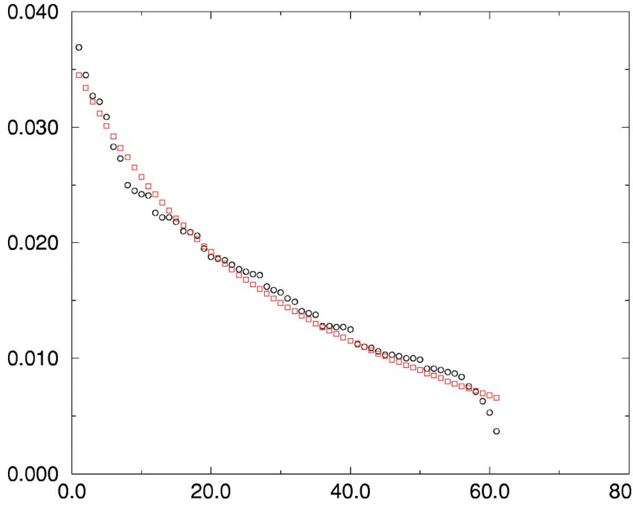


FIG. 3. Rank distribution of the codon usage probabilities for *Arabidopsis thaliana*. Circles are experimental values, squares are fitted values.

tions hold for biological species with sufficiently large statistics of codons. For each biological species, codons are ordered following decreasing order of the values of their usage probabilities, i.e., codon number 1 corresponds to the highest value, codon number 2 is the next highest, and so on. We denote by $f(n)$ the probability $P(XZN)$ of finding that XZN is in the nth position. Of course the same codon occupies in general two different positions in the rank distribution function for two different species. We plot $f(n)$ versus the rank and we determine that the data are well fitted by the sum of an exponential function, a linear function in the rank, and a constant, i.e.,

$$f(n) = \alpha e^{-\eta n} - \beta n + \gamma, \tag{4}$$

where $0.0187 \leqslant \alpha \leqslant 0.0570$, $0.050 \leqslant \eta \leqslant 0.136$, $0.82 \times 10^{-4} \leqslant \beta \leqslant 3.63 \times 10^{-4}$, and $\gamma = 0.016$ are constant depending on the biological species. These four constants have to satisfy the normalization condition

$$\sum_{n} f(n) = 1. \tag{5}$$

In Table I we list the 40 biological species (six vertebrates, four plants, three invertebrates, two fungi and 25 bacteria) with a sample of codons of sizes between 800 000 and 20 000 000 in decreasing order (data from GenBank release 129.0 [12]) whose codon usage has been fitted, specifying for each biological species the value of the parameters computed by a best-fit procedure and the corresponding χ^2. Here and in the following, the χ^2 coefficient is defined by

$$\chi^2 = \sum_{i} \frac{[y_i - y(x_i)]^2}{y(x_i)}, \tag{6}$$

where x_i are the experimental abscissae, y_i the experimental values, and $y(x_i)$ the fitted ones. In some cases, $y(x_i)$ takes

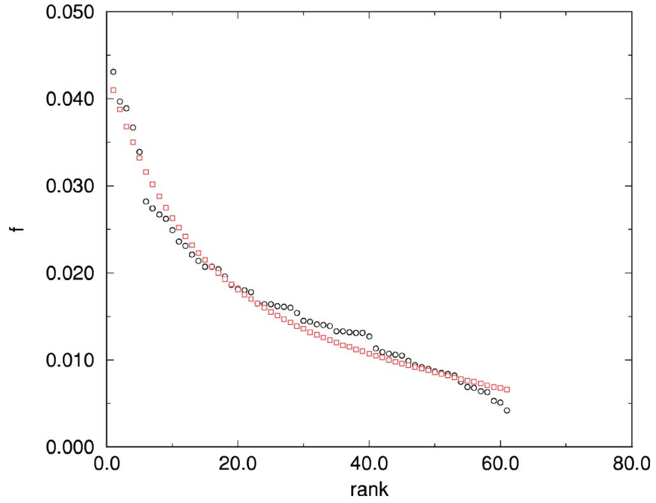


FIG. 2. Rank distribution of the codon usage probabilities for *Drosophila melanogaster*. Circles are experimental values, squares are fitted values.

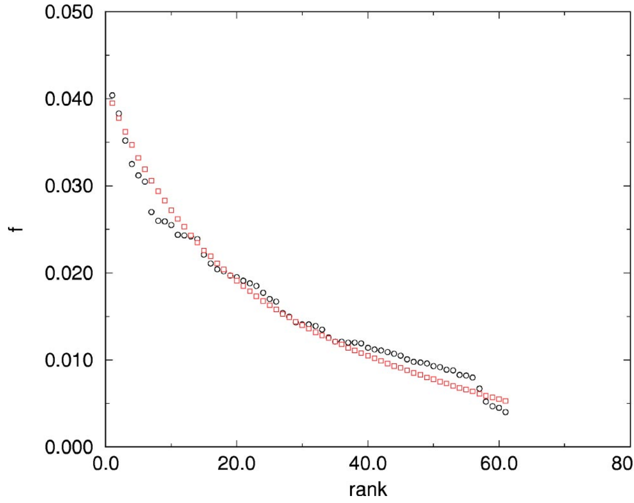


FIG. 4. Rank distribution of the codon usage probabilities for *Escherichia coli*. Circles are experimental values, squares are fitted values.

TABLE II. Type of codons used for the observed rank distribution $f(n)$.

	Rank																			
Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Homo sapiens	*GAG*	*CUG*	*CAG*	*AAG*	*GAA*	*GUG*	*GCC*	*GAC*	*AAA*	*GGC*	*AUG*	*GAU*	*AUC*	*UUC*	*CCC*	*AAC*	*CUC*	*AGC*	*ACC*	*GCU*
Mus musculus	*CUG*	*GAG*	*AAG*	*CAG*	*GUG*	*GAC*	*GAA*	*GCC*	*AUC*	*AUG*	*GGC*	*UUC*	*GAU*	*AAA*	*AAC*	*CUC*	*GCU*	*AGC*	*ACC*	*CCC*
Rattus norvegicus	*CUG*	*GAG*	*AAG*	*CAG*	*GUG*	*GAC*	*GCC*	*GAA*	*AUC*	*UUC*	*AUG*	*AAC*	*GGC*	*CUC*	*GAU*	*AAA*	*ACC*	*AGC*	*GCU*	*CCC*
Gallus gallus	*GAG*	*CUG*	*AAG*	*CAG*	*GAA*	*GUG*	*AAA*	*GAC*	*GCC*	*GAU*	*AAC*	*AUC*	*AUG*	*GGC*	*AGC*	*UUC*	*GCU*	*CCC*	*UAC*	*ACC*
Xenopus laevis	*GAA*	*GAG*	*AAA*	*AAG*	*CAG*	*GAU*	*CUG*	*AUG*	*GAC*	*AAU*	*AAC*	*GGA*	*GUG*	*GCU*	*CCA*	*AUU*	*GCA*	*UUU*	*UGU*	*AGA*
Bos taurus	*CUG*	*GAG*	*AAG*	*CAG*	*GUG*	*GCC*	*GAC*	*GAA*	*AUC*	*GGC*	*UUC*	*AAC*	*AUG*	*AAA*	*ACC*	*GAU*	*CUC*	*CCC*	*UAC*	*AGC*
Arabidopsis thaliana	*GAU*	*GAA*	*AAG*	*GAG*	*AAA*	*GCU*	*GUU*	*UCU*	*AUG*	*CUU*	*GGA*	*AAU*	*GGU*	*UUU*	*AUU*	*UUG*	*AAC*	*UUC*	*CAA*	*AGA*
Oryza sativa japonica	*GAG*	*GCC*	*GGC*	*AAG*	*GAC*	*GCG*	*CUC*	*GUG*	*GAU*	*AUG*	*UUC*	*CUG*	*GAA*	*CAG*	*GUC*	*AUC*	*CCG*	*GCU*	*AAC*	*CGC*
Oryza sativa	*GAG*	*AAG*	*GCC*	*GGC*	*GAC*	*GAU*	*CUC*	*AUG*	*GUG*	*GCG*	*UUC*	*CAG*	*GAA*	*AUC*	*GCU*	*AAC*	*GUC*	*CUG*	*GCA*	*GGG*
Neurospora crassa	*GAG*	*AAG*	*GCC*	*GAC*	*GGC*	*AAC*	*CUC*	*AUC*	*CAG*	*GUC*	*ACC*	*GAU*	*CCC*	*UUC*	*AUG*	*GAA*	*GCU*	*UCC*	*GGU*	*UAC*
Drosophila melanogaster	*GAG*	*AAG*	*CUG*	*CAG*	*GCC*	*GUG*	*GAU*	*GGC*	*AAC*	*GAC*	*AUG*	*AUC*	*UUC*	*ACC*	*GAA*	*AAU*	*AGC*	*UCC*	*UAC*	*CGC*
Caenorhabditis elegans	*GAA*	*AAA*	*GAU*	*AUU*	*GGA*	*AAU*	*CAA*	*AUG*	*AAG*	*CCA*	*UUU*	*UUC*	*GAG*	*GUU*	*GCU*	*CUU*	*UCA*	*UUG*	*ACA*	*GCA*
Leishmania major	*GCG*	*GAG*	*GCC*	*CUG*	*GUG*	*GGC*	*GAC*	*CAG*	*CGC*	*AAG*	*CCG*	*AGC*	*CUC*	*ACG*	*AUG*	*AAC*	*UCG*	*CAC*	*GCA*	*UAC*
Sacch. cerevisiae	*GAA*	*AAA*	*GAU*	*AAU*	*AAG*	*AUU*	*CAA*	*UUG*	*UUA*	*UUU*	*AAC*	*GGU*	*UCU*	*GUU*	*AGA*	*GCU*	*AUG*	*GAC*	*ACU*	*GAG*
Schizosacch. pombe	*GAA*	*AAA*	*GAU*	*AUU*	*AAU*	*UUU*	*UCU*	*GCU*	*GUU*	*CAA*	*UUA*	*CUU*	*AAG*	*UUG*	*ACU*	*UAU*	*CCU*	*GGU*	*GAG*	*AUG*
Escherichia coli	*CUG*	*GAA*	*AAA*	*GAU*	*GCG*	*AUU*	*CAG*	*GGC*	*AUG*	*GGU*	*GUG*	*GCC*	*AUC*	*UUU*	*ACC*	*AAC*	*GCA*	*CCG*	*AAU*	*CGU*
Bacillus subtilis	*AAA*	*GAA*	*AUU*	*GAU*	*UUU*	*AUC*	*AUG*	*GGC*	*GAG*	*CUG*	*CUU*	*UAU*	*AAU*	*ACA*	*GGA*	*GCA*	*AAG*	*GCG*	*CAA*	*UUA*
Pseudom. aeruginosa	*CUG*	*GCC*	*GGC*	*CGC*	*GAC*	*GCG*	*AUC*	*GAG*	*CAG*	*GUG*	*UUC*	*ACC*	*CCG*	*GUC*	*CUC*	*AAG*	*AGC*	*GAA*	*AAC*	*AUG*
Mesorhizobium loti	*GGC*	*GCC*	*CUG*	*AUC*	*GCG*	*GUC*	*GAC*	*CGC*	*UUC*	*GAG*	*CCG*	*AAG*	*CUC*	*GUG*	*ACC*	*CAG*	*AUG*	*GAA*	*UCG*	*GAU*
Streptom. coelicolon A3	*GCC*	*GGC*	*CUG*	*GAC*	*GCG*	*GAG*	*GUC*	*ACC*	*CGC*	*CUC*	*GUG*	*CCG*	*CGG*	*AUC*	*UUC*	*CCC*	*CAG*	*CAC*	*UCC*	*AAG*
Sinorhizobium meliloti	*GGC*	*GCC*	*GCG*	*AUC*	*GUC*	*CUC*	*CUG*	*GAC*	*CGC*	*GAG*	*UUC*	*CCG*	*AAG*	*GAA*	*AUG*	*CAG*	*GUG*	*ACC*	*ACG*	*UCG*
Nostoc, sp. PCC7120	*GAA*	*AUU*	*CAA*	*UUA*	*AAA*	*GAU*	*AAU*	*UUU*	*GCU*	*GGU*	*GCA*	*UUG*	*GUU*	*ACU*	*UAU*	*GUA*	*ACA*	*AUC*	*GCC*	*AUG*
Agrobact. tumefaciens	*GCC*	*GGC*	*CUG*	*AUC*	*GCG*	*GAA*	*CGC*	*GUC*	*UUC*	*GAU*	*GAC*	*AAG*	*CUC*	*CCG*	*AUG*	*GUG*	*CAG*	*GAG*	*ACC*	*ACG*
Ralstonia solanacearum	*CUG*	*GCC*	*GGC*	*GCG*	*CGC*	*GUG*	*AUC*	*GAC*	*CCG*	*CAG*	*GAG*	*UUC*	*ACC*	*GUC*	*AAG*	*ACG*	*AUG*	*AAC*	*UCG*	*CUC*
Yersinia pestis	*CUG*	*GAU*	*GAA*	*AAA*	*AUU*	*GCC*	*AUG*	*GGU*	*CAG*	*AAU*	*GCG*	*GGC*	*CAA*	*AUC*	*UUG*	*GUG*	*UUU*	*ACC*	*UUA*	*GAG*
Methanosarc. acetivorans	*GAA*	*AAA*	*CUU*	*GAU*	*GGA*	*AUU*	*GCA*	*AUC*	*GAG*	*UUU*	*CUG*	*GAC*	*AUG*	*AAU*	*AAG*	*AAC*	*AUA*	*GUU*	*UAU*	*UUC*
Vibrio cholerae	*GAA*	*GAU*	*AAA*	*CAA*	*AUU*	*GCG*	*GUG*	*UUU*	*CUG*	*GGU*	*AUG*	*AUC*	*GAG*	*GGC*	*UUG*	*AAU*	*GCC*	*UUA*	*GCU*	*ACC*
Escherichia coli K12	*CUG*	*GAA*	*GCG*	*AAA*	*GAU*	*AUU*	*GGC*	*CAG*	*AUG*	*GUG*	*GCC*	*AUC*	*GGU*	*ACC*	*CCG*	*UUU*	*CGC*	*AAC*	*CGU*	*GCA*
Mycobact. tuber. CDC1551	*GCC*	*CUG*	*GGC*	*GCG*	*GAC*	*GUG*	*ACC*	*AUC*	*GUC*	*CCG*	*GAG*	*CGC*	*CGG*	*CAG*	*UUC*	*UCG*	*AAC*	*GGG*	*GGU*	*AUG*
Mycobact. tuber. H37Rv	*GCC*	*GGC*	*CUG*	*GCG*	*GAC*	*GUG*	*ACC*	*AUC*	*GUC*	*CCG*	*GAG*	*CGC*	*CGG*	*UUC*	*CAG*	*AAC*	*UCG*	*GGG*	*GGU*	*AUG*
Bacillus. halodurans	*GAA*	*AUU*	*AAA*	*GAU*	*UUU*	*GAG*	*CAA*	*UUA*	*AUG*	*AUC*	*UAU*	*CUU*	*GGA*	*GUU*	*AAG*	*ACG*	*AAU*	*GCA*	*GUG*	*GCG*
Clostridium acetobutylicum	*AAA*	*AUA*	*AAU*	*GAA*	*GAU*	*UUU*	*UUA*	*AUU*	*UAU*	*GGA*	*AAG*	*GUU*	*GUA*	*CUU*	*GCA*	*AUG*	*AGA*	*GCU*	*ACA*	*GGU*
Caulobacter crescentus CB15	*GCC*	*CUG*	*GGC*	*GCG*	*GAC*	*CGC*	*AUC*	*GUC*	*GAG*	*ACC*	*AAG*	*UUC*	*GUG*	*CCG*	*CAG*	*AUG*	*UCG*	*AAC*	*CCC*	*CUC*
Synechocystis sp. PCC6803	*GAA*	*AUU*	*GCC*	*CAA*	*GAU*	*UUG*	*AAA*	*UUU*	*GUG*	*ACC*	*UUA*	*CCC*	*AAU*	*GGC*	*CAG*	*CUG*	*GCU*	*GGU*	*AUG*	*GAC*
Sulfolobus solfataricus	*AUA*	*UUA*	*AAA*	*GAA*	*AAG*	*GAU*	*AUU*	*AAU*	*UAU*	*GAG*	*GUA*	*GUU*	*UUU*	*GGA*	*AGA*	*GCU*	*GGU*	*AUG*	*ACU*	*GCA*
Mycobacterium leprae	*GCC*	*CUG*	*GUG*	*GAC*	*GCG*	*GGC*	*AUC*	*GUC*	*ACC*	*GAG*	*CCG*	*UUG*	*GGU*	*CGC*	*CAG*	*GAU*	*GAA*	*GCU*	*UUC*	*CGG*
Brucella melitensis	*GGC*	*GCC*	*CUG*	*GAA*	*CGC*	*GCG*	*AUC*	*GAU*	*AAG*	*GUG*	*CCG*	*UUC*	*AUG*	*CAG*	*CUU*	*GAC*	*GUC*	*ACC*	*CUC*	*GAG*
Deinococcus radiodurans	*CUG*	*GCC*	*GGC*	*GUG*	*GCG*	*GAC*	*CGC*	*ACC*	*CAG*	*CUC*	*GAG*	*CCC*	*GAA*	*CCG*	*AGC*	*GUC*	*AUC*	*UUC*	*GGG*	*CGG*
Listeria monocytogenes	*AAA*	*GAA*	*AUU*	*GAU*	*UUA*	*AAU*	*UUU*	*CAA*	*GCA*	*GUU*	*AUG*	*ACA*	*GGU*	*UAU*	*GCU*	*GUA*	*CUU*	*GGA*	*AUC*	*CCA*
Clostridium perfringens	*AAA*	*GAA*	*UUA*	*AUA*	*AAU*	*GAU*	*GGA*	*UUU*	*GUU*	*UAU*	*AUU*	*GCU*	*AGA*	*AAG*	*GUA*	*AUG*	*ACU*	*UCA*	*GCA*	*ACA*

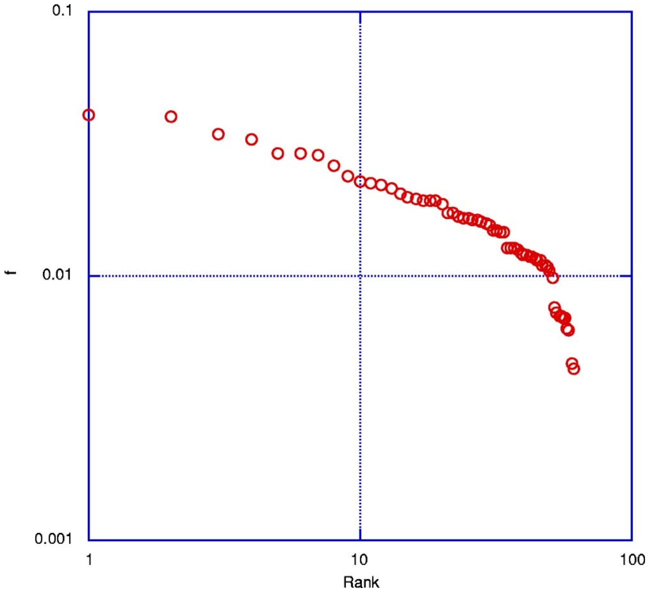


FIG. 5. Log-log ranked distribution of the codon usage probabilities for *Homo sapiens*.

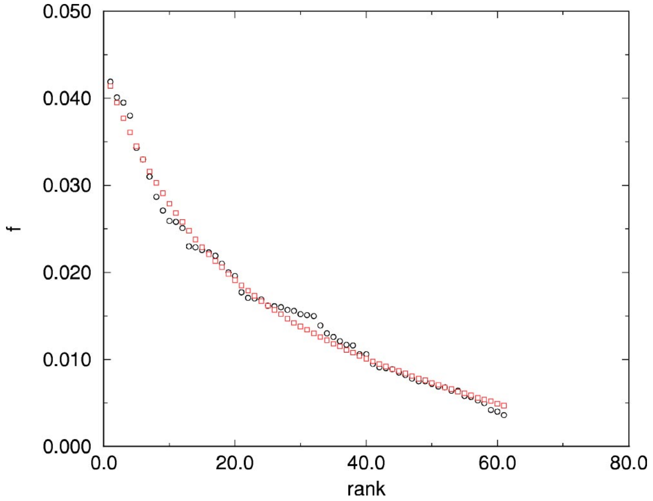


FIG. 6. Rank distribution of the codon usage probabilities for chloroplast *Arabidopsis thaliana*. Circles are experimental values, squares are fitted values.

vanishing or negative values for a few points and hence the χ^2 is not reported. In Figs. 1–4, we report the plots of $f(n)$ as a function of n for a few biological species (*Homo sapiens, Drosophila melanogaster, Arabidopsis thaliana*, and *Escherichia coli*). The plot has been cut to $n=61$ to take into account the fact that in standard code there are three Stop codons (to end the biosynthesis process), whose function is very peculiar. For the same reason, the χ^2 has been computed by taking into account the 61 coding codons only. In Table II, we report the type of the 20 most used codons of the observed rank distribution $f(n)$. The goodness of fit can be estimated by $P(n/2,\chi^2/2)$, where $P(a,x)$ is the incomplete Gamma function and n is the number of degrees of freedom. $P(n/2,\chi^2/2)$ is the probability that the observed χ^2 for a correct model should be less than the calculated χ^2. In the present case, $P(n/2,\chi^2/2)$ is less than 10^{-5} for each species. In Fig. 5, for *Homo sapiens*, we draw the log-log ranked plot, which obviously does not show a linear trend taking into account all the points, as would be the case for a Zipf's law behavior. Indeed, as emphasized in [5], when the majority of points reside in the tail of the distribution, it is necessary to fit the whole range of data.

A similar study, for a sample of 20 vertebrates with codon statistics larger than 100 000, reveals that, for almost all biological species, the four most used codons are *GAG, CUG, AAG*, and *CAG*. All these codons have a *G* nucleotide in the third position and three of them encode doublets. An analysis performed on the chloroplast codon usage for a sample of five plants gives the same result for the rank distribution $f(n)$; see Table III and Fig. 6 (Chloroplast *Arabidopsis thaliana*). We also report, in Table IV, the values of the parameters and the χ^2 for a sample of nine mitochondria with codon statistics larger than 15 000. The fits for *Homo sapiens* and *Arabidopsis thaliana* are presented in Figs. 7 and 8. We point out, however, that for mitochondria the codon usage frequency distribution for several species (e.g., *Arabidopsis thaliana* or *Drosophila melanogaster*) is ill fitted by Eq. (4). This may be an indication that mitochondria do not follow the universal law (4). Note that the mitochondrial codes have a few differences from the eukaryotic code and vary slightly between species; see, e.g., [13]. In these cases, the χ^2 has been computed over the corresponding coding codons. The value of the constant γ is approximately equal to $1/61=0.0164$ or $1/64=0.0156$, i.e., the value of the codon usage probability in the case of a uniform and unbiased codon distribution. Therefore the other two terms in Eq. (4) can be viewed as the effect of the bias mechanism. The appearance of the linear term is more intriguing. Let us remark that in [10], where an exponential function is used to fit the rank of usage in genes (not the rank of usage probability), the linear

TABLE III. Values of the best-fit parameters, Eq. (4), for the sample of chloroplasts.

Species	*GC* content (%)	α	η	$10^4\beta$	χ^2
Arabidopsis thaliana	38.37	0.0254	0.067	1.95	0.0030
Chaetosphaeridium globosum	30.29	0.0515	0.110	2.59	0.0174
Chlorella vulgaris	34.63	0.0513	0.114	2.04	0.0093
Cyanidium caldarium	33.31	0.0379	0.092	2.24	0.0103
Guillardia theta	33.20	0.0452	0.103	2.20	0.0089

TABLE IV. Values of the best-fit parameters, Eq. (4), for the sample of mitochondria.

Type	Species	GC content (%)	α	η	$10^4\beta$	χ^2
vrt	*Homo sapiens*	44.99	0.0414	0.099	2.31	0.0207
pln	*Arabidopsis thaliana*	44.18	0.0136	0.049	1.39	0.0589
vrt	*Mus musculus*	37.23	0.0455	0.104	2.44	0.0226
fng	*Saccharomyces cerevisiae*	24.17	0.0879	0.198	2.66	0.0611
inv	*Physarum polycephalum*	25.69	0.0624	0.128	2.70	0.0262
pln	*Pylaiella littoralis*	37.06	0.0336	0.108	1.72	0.0112
pln	*Neurospora crassa*	33.20	0.0388	0.101	2.14	0.0225
vrt	*Bos taurus*	39.73	0.0422	0.106	2.25	0.0430
vrt	*Sus scrofa*	40.52	0.0497	0.112	2.51	0.0372

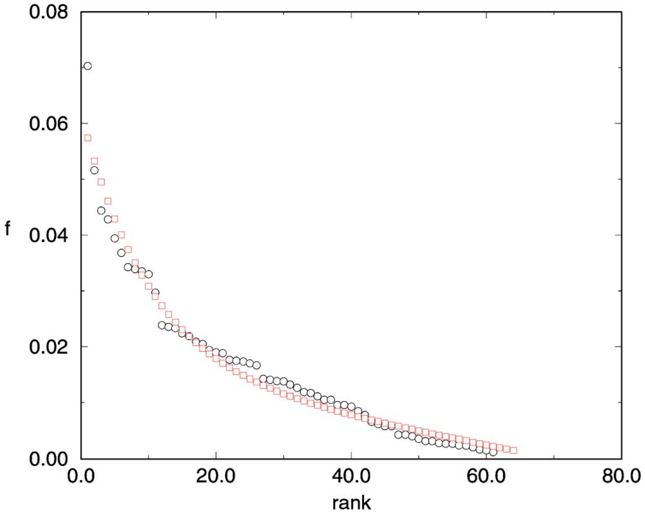


FIG. 7. Rank distribution of the codon usage probabilities for mitochondrial *Homo sapiens*. Circles are experimental values, squares are fitted values.

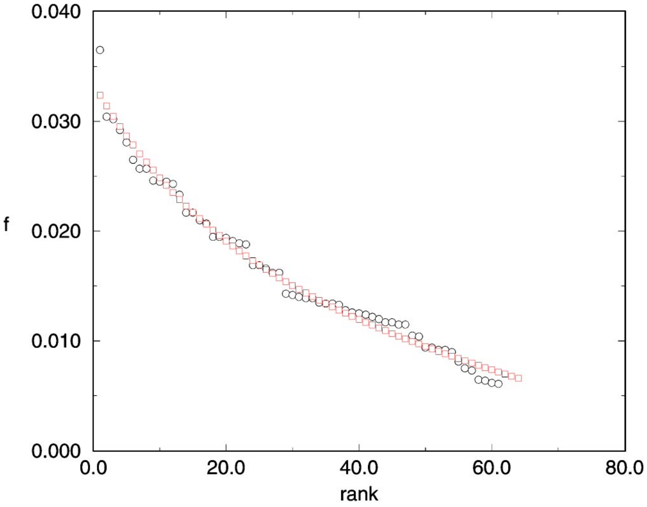


FIG. 8. Rank distribution of the codon usage probabilities for mitochondrial *Arabidopsis thaliana*. Circles are experimental values, squares are fitted values.

term was observed, as its contribution becomes noticeable for approximately $n \geqslant 20$. Owing to the analysis of genes (with at most a few hundred codons), the fits in that paper end before this value of the rank. It is believed that the main causes of codon usage bias are translational efficiency, selection pressure, and spontaneous mutations. From the smallness of the parameter β in Eq. (4), it is tempting to identify it as a consequence of the mutation effect and the first term in Eq. (4) as the effect of selection pressure, i.e., the interaction with the environment.

Since it is well known that the GC content plays a strong role in the evolutionary process, we expect the parameters to depend on the total GC content of the gene region (here the total exonic GC content) that is indeed correlated with the evolution of the system (see [14] and references therein). We have investigated this dependence and report, in Fig. 9, the fits of α and β to the total exonic GC content Y_{GC} of the biological species. One finds that the values of α and β are well fitted by polynomial functions (with $0 \leqslant Y_{GC} \leqslant 100\%$):

$$\alpha = 0.21145 - 0.00776 Y_{GC} + 7.92 \times 10^{-5} Y_{GC}^2, \quad \chi^2 = 0.0262, \tag{7}$$

$$10^2\beta = 0.10096 - 0.00345 Y_{GC} + 3.50 \times 10^{-5} Y_{GC}^2, \quad \chi^2 = 0.0170. \tag{8}$$

The two parameters α and β appear to be correlated. Indeed the plot representing β as a function of α is satisfactorily fitted by a regression line (see Fig. 10):

$$10^2\beta = 0.00851 + 0.375\alpha, \quad \chi^2 = 0.0218. \tag{9}$$

The value of the η parameter is largely uncorrelated with the total exonic GC content. Let us recall, however, that η is a function of α and β due to the normalization condition of Eq. (5). Indeed we have[2] (assuming $e^{-65\eta} \approx 0$)

$$1 = \frac{\alpha e^{-\eta}}{1 - e^{-\eta}} + 2080\beta + 64\gamma. \tag{10}$$

[2] Note that the result is almost unchanged if the data are normalized on the 61 coding codons.

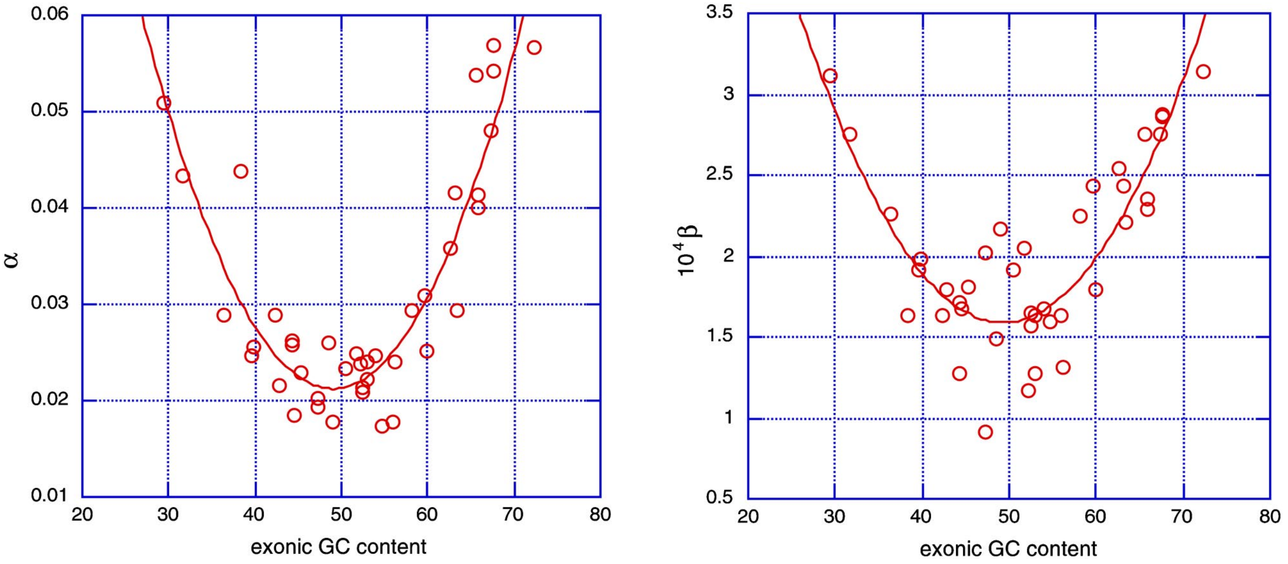


FIG. 9. Fits for the α and β parameters.

Using the fits for α and β, we can write the probability distribution function for any biological species, whose total GC content in percent in the exonic regions is Y_{GC}, as

$$f(n)=(\alpha_0+\alpha_1 Y_{GC}+\alpha_2 Y_{GC}^2)e^{-\eta n}-n(\beta_0+\beta_1 Y_{GC} + \beta_2 Y_{GC}^2)+\gamma, \tag{11}$$

where η is obtained by solving Eq. (10). Of course we are not able to predict which codon occupies the nth rank. Finally, let us remark that the total exonic GC content Y_{GC} has to satisfy the consistency condition

$$Y_{GC}=\frac{1}{3}\sum_{i\in I} d_i f(i), \tag{12}$$

where the sum is over the set I of integers to which the 56 codons containing G and/or C nucleotides belong and d_i is the multiplicity of these nucleotides inside the ith codon.

III. AMINO-ACID RANK DISTRIBUTION

It is natural to wonder if some kind of universality is also present in the rank distribution of amino acids. From the available data for codon usage, we can immediately compute (using the eukaryotic code) the frequency of appearance of any amino acid $F(n)$ $(1\leqslant n\leqslant 20)$ in the whole set of coding sequences. The calculated values as a function of the rank are satisfactorily fitted by a straight line

$$F(n)=F_0-Bn. \tag{13}$$

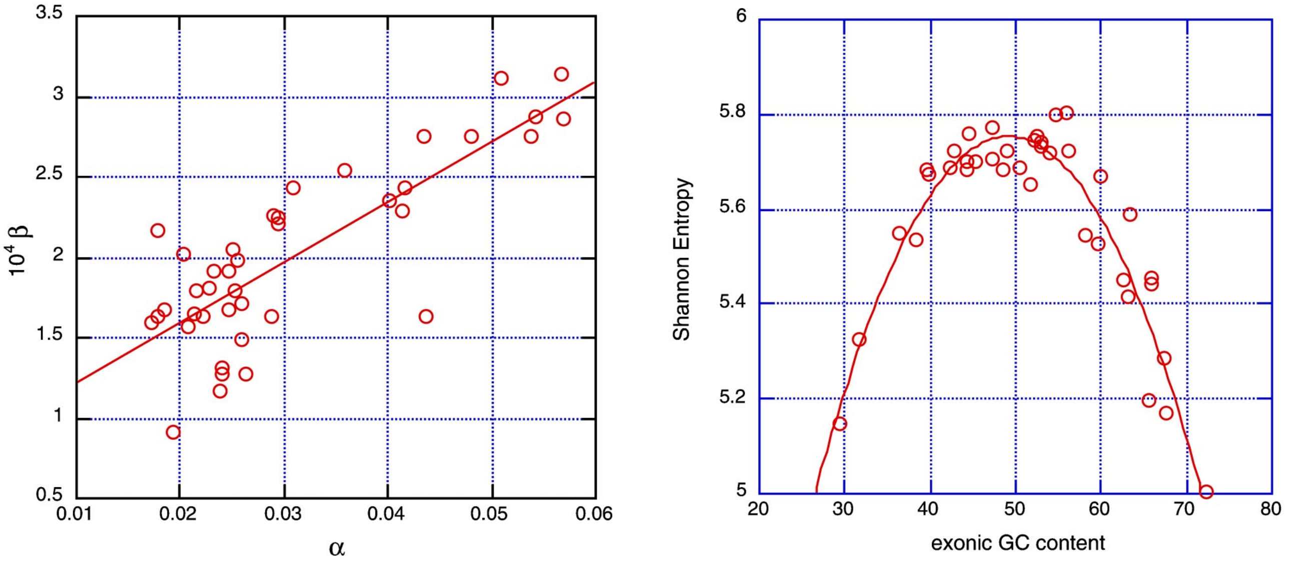


FIG. 10. Fits for the α and β parameters and for the Shannon entropy.

TABLE V. Values of the best-fit parameters for amino acids.

Species	10^3B	F_0	χ^2
Homo sapiens	3.8	0.089	0.0072
Arabidopsis thaliana	3.8	0.090	0.0068
Drosophila melanogaster	3.5	0.087	0.0125
Caenorhabditis elegans	3.3	0.084	0.0124
Mus musculus	3.7	0.088	0.0087
Saccharomyces cerevisiae	3.9	0.090	0.0121
Escherichia coli	4.0	0.091	0.0115
Rattus norvegicus	3.7	0.088	0.0084
Oryza sativa japonica	4.1	0.093	0.0057
Schizosaccharomyces pombe	3.8	0.089	0.0162
Bacillus subtilis	4.0	0.091	0.0104
Pseudomonas aeruginosa	4.9	0.101	0.0493
Mesorhizobium loti	4.7	0.100	0.0215
Streptomyces coelicolor A3	5.6	0.109	0.0624
Sinorhizobium meliloti	4.7	0.100	0.0188
Nostoc sp. PCC7120	4.0	0.092	0.0174
Oryza sativa	3.9	0.091	0.0028
Agrobacterium tumefaciens str. C38	4.6	0.098	0.0144
Ralstonia solanacearum	4.7	0.101	0.0351
Yersinia pestis	4.0	0.092	0.0135
Methanosarcina acetivorans str. C2A	4.1	0.092	0.0063
Vibrio cholerae	3.9	0.091	0.0148
Escherichia coli K12	4.0	0.091	0.0154
Mycobacterium tuberculosis CDC1551	5.2	0.105	0.01121
Mycobacterium tuberculosis H37Rv	5.3	0.106	-
Bacillus halodurans	4.0	0.091	0.0100
Clostridium acetobutylicum	4.6	0.097	0.0076
Caulobacter crescentus CB15	5.1	0.104	0.0524
Gallus gallus	3.6	0.088	0.0040
Synechocystis sp. PCC6803	4.1	0.093	0.0168
Sulfolobus solfataricus	4.4	0.096	0.0143
Mycobacterium leprae	4.9	0.101	0.0401
Brucella melitensis	4.5	0.097	0.0142
Deinococcus radiodurans	5.2	0.105	0.0679
Xenopus laevis	3.5	0.086	0.0084
Listeria monocytogenes	4.2	0.093	0.0088
Neurospora crassa	4.0	0.091	0.0042
Clostridium perfringens	4.6	0.098	0.0035
Leishmania major	4.7	0.099	0.0367
Bos taurus	3.6	0.087	0.0082

The parameters F_0 and B and the corresponding χ^2 for the fits are reported in Table V. It is interesting to recall that the linear trend was noted, from the analysis of a small number of proteins, in 1955 by Gamow and Ycas [15]. A better fit can be obtained in general by using a third-degree polynomial; however, the range of the four parameters for this fit is larger than the range of the two-parameter fit. For a few biological species, we give below the parameters for the two fits (see also Fig. 11). The plots of the linear fits for a few biological species are given in Fig. 12. Note that the 21st point is just the contribution of the Stop codons, which of course has not been taken into account for the fits. One can remark that the most frequent amino acid is always above the line. This can be easily understood in the light of Eq. (4). Indeed, the most frequent amino acids get, in general, a contribution of the exponential term of Eq. (4) with a low value of n.

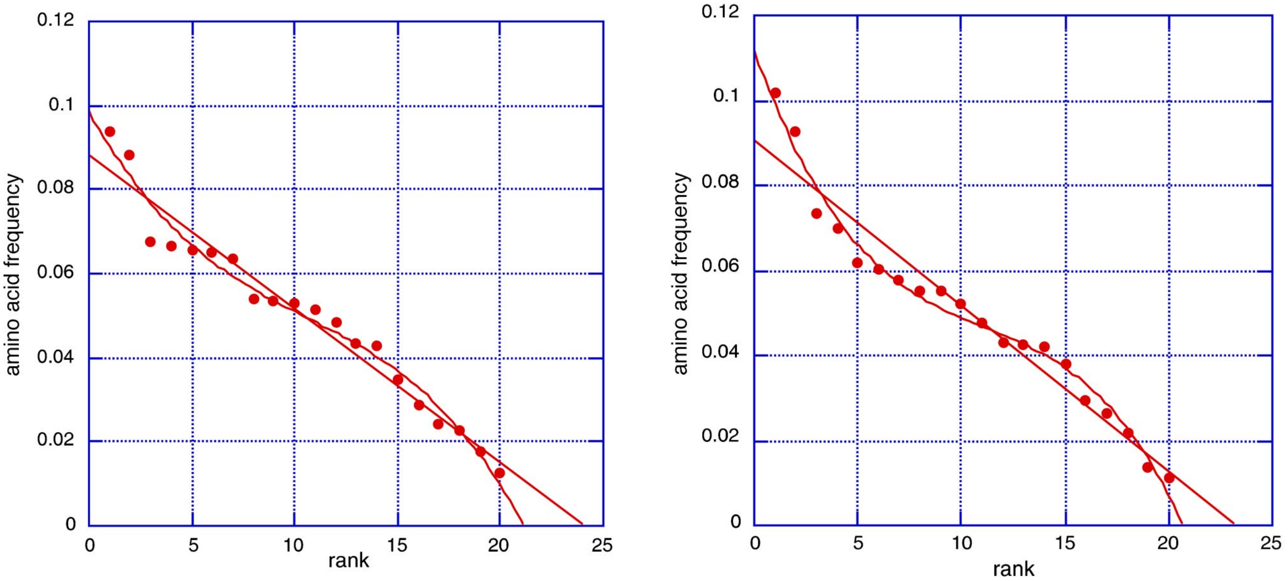
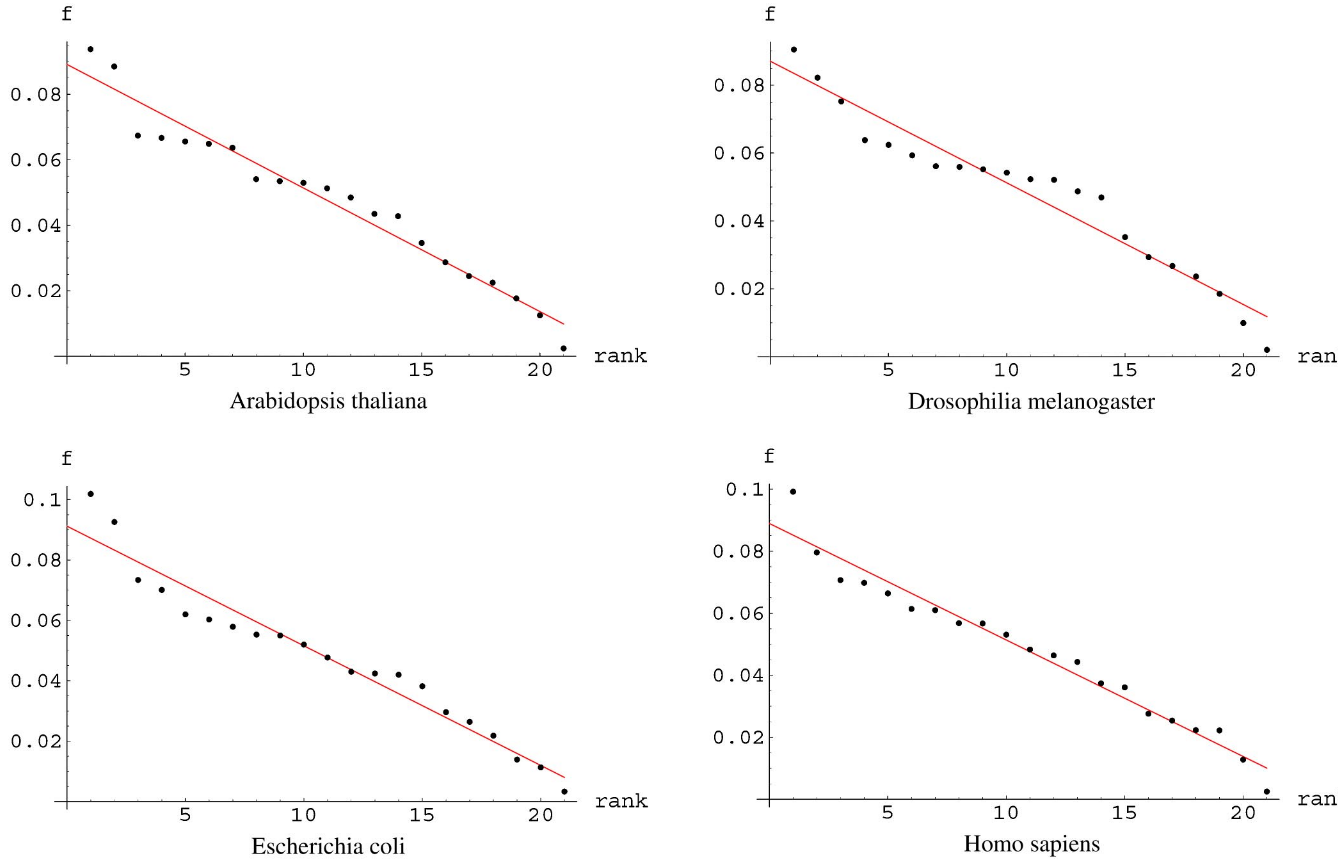


FIG. 11. Amino-acid frequency: linear vs cubic fits.



Arabidopsis thaliana

Drosophilia melanogaster

Escherichia coli

Homo sapiens

FIG. 12. Amino-acid rank distributions.

Species	Linear/cubic fits	χ^2
Homo sapiens	lin. $f=0.087-0.0036n$	0.0072
	cub. $f=0.099-0.0088n+57\times10^{-5}n^2-1.7\times10^{-5}n^3$	0.0055
Arabidopsis thaliana	lin. $f=0.088-0.0036n$	0.0068
	cub. $f=0.099-0.0090n+62\times10^{-5}n^2-1.95\times10^{-5}n^3$	0.0049
Drosophila melanogaster	lin. $f=0.087-0.0036n$	0.0125
	cub. $f=0.097-0.0096n+76\times10^{-5}n^2-2.5\times10^{-5}n^3$	0.0042
Escherichia coli	lin. $f=0.090-0.0039n$	0.0115
	cub. $f=0.112-0.0136n+105\times10^{-5}n^2-3.1\times10^{-5}n^3$	0.0067

Of course, the frequency of an amino acid is given by the sum of the frequencies of its encoding codons given by Eq. (4). If the ranks of the encoding codons were completely random, we would not expect their sum to take equally spaced values, as is the case in a regression line. Therefore, we can infer, for the biological species whose amino-acid frequency is very well fitted by a line, the existence of some functional constraints on the codon usage.

We report in Table VI the distribution of the amino acids for the different biological species. There is no clear correlation between the rank of the codons and the rank of the encoded amino acids. As has been previously remarked, in many species three of the four most used codons encode for doublets which are generally less used than the five quartets and the three sextets.[3] The statement is illustrated by Fig. 13, where we plot, for *Homo sapiens*, the frequencies of the codons according to the rank of the encoded amino acids, indicating for each amino acid the rank of the corresponding codons. In the legend, for each amino acid, "codon 1" means the most used codon, "codon 2" the next most used codon, and so on.

However, the behavior predicted by Eq. (4) fits the experimental data very well, while the shape of the distribution of amino acids seems more sensible for biological species. In fact, one can remark in many plots of the amino-acid distributions (see, e.g., Fig. 12) the existence of one or two plateaus, which obviously indicate equal probabilities of use for some amino acids. Presently, we do not have any argument to explain the uniform distribution of amino acids from the ranked distribution of the corresponding codons.

IV. CONSEQUENCES OF PROBABILITY DISTRIBUTION

We now derive a few consequences of Eq. (4). In the following, we denote by y the *local* exonic *GC* content (i.e., for coding sequences of genes) for a given biological species. Let us assume that the exonic *GC* content of a biological species is essentially comprised in the interval $y_1-y_0=\Delta$ (e.g., for *Homo sapiens* $y_0=35\%$ and $y_1=70\%$). We can write

$$f(n)=\frac{1}{\Delta}\int_{y_0}^{y_1} f(y,n)\,dy. \tag{14}$$

Since the left-hand side of the above equation has the form given by Eq. (4) for any n and for any biological species, if we do not want to invoke some "fine-tuning" in the integrand function $f(y,n)$, we have to assume that

$$f(y,n)=a(y)e^{-\eta n}-b(y)n+\gamma \tag{15}$$

with the condition

$$\alpha=\frac{1}{\Delta}\int_{y_0}^{y_1} a(y)\,dy, \quad \beta=\frac{1}{\Delta}\int_{y_0}^{y_1} b(y)\,dy. \tag{16}$$

As a consequence, we predict that the codon usage probability is the same for any codon in any exonic genic region with the same *GC* content. The form of the $a(y)$ and $b(y)$ functions is yet undetermined. For *Homo sapiens*, we remark that the total exonic *GC* content Y_{GC} is, in a very good approximation, equal to the mean value of the interval $[y_0,y_1]$. Therefore, inserting Eqs. (14) and (15) into Eq. (12), we derive the result that the functions $a(y)$ and $b(y)$ have to be *linear* functions of y. This theoretical derivation is in accordance with the conclusions of Zeeberg [16] obtained by an analysis of 7357 genes. On a quantitative level, using the numerical linear fits of Zeeberg, we find a very good agreement with our calculations. Note that this result is not in contradiction with Eq. (11), since the previous analysis is valid for the fixed value of the exonic *GC* content for *Homo sapiens*. For bacteria, the range of variation Δ of the local exonic *GC* content is very small. Therefore we expect the functions $a(y)$ and $b(y)$ to have the same shape as the functions α and β given in Eqs. (7) and (8). Hence the functions α and β depend on the biological species.

We compute the Shannon entropy, given by

$$S=-\sum_n f(n)\log_2 f(n), \tag{17}$$

[3] Here and elsewhere, the words doublet, quartet, sextet, etc., refer to the group of (synonymous) codons coding for the same amino acid.

TABLE VI. Type of amino acids of the observed rank distribution.

	Rank																			
Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Homo sapiens	Leu	Ser	Ala	Glu	Gly	Val	Pro	Lys	Arg	Thr	Asp	Gln	Ile	Phe	Asn	Tyr	His	Met	Cys	Trp
Mus musculus	Leu	Ser	Ala	Gly	Glu	Val	Pro	Lys	Arg	Thr	Asp	Ile	Gln	Phe	Asn	Tyr	His	Cys	Met	Trp
Rattus norvegicus	Lue	Ser	Ala	Glu	Gly	Val	Pro	Lys	Thr	Arg	Asp	Ile	Gln	Phe	Asn	Tyr	His	Met	Cys	Trp
Gallus gallus	Leu	Ser	Glu	Ala	Gly	Lys	Val	Pro	Thr	Arg	Asp	Ile	Gln	Asn	Phe	Tyr	His	Met	Cys	Trp
Xenopus laevis	Leu	Ser	Glu	Lys	Ala	Gly	Val	Pro	Thr	Asp	Arg	Ile	Gln	Asn	Phe	Tyr	Met	His	Cys	Trp
Bos taurus	Leu	Ser	Ala	Gly	Glu	Val	Lys	Pro	Thr	Arg	Asp	Ile	Gln	Phe	Asn	Tyr	Cys	His	Met	Trp
Arabidopsis thaliana	Leu	Ser	Val	Glu	Gly	Ala	Lys	Asp	Arg	Ile	Thr	Pro	Asn	Phe	Gln	Tyr	Met	His	Cys	Trp
Oryza sativa japonica	Ala	Leu	Gly	Ser	Arg	Val	Glu	Pro	Asp	Thr	Lys	Ile	Phe	Gln	Asn	His	Tyr	Met	Cys	Trp
Oryza sativa	Ala	Leu	Gly	Ser	Arg	Val	Glu	Pro	Asp	Lys	Thr	Ile	Phe	Gln	Asn	Tyr	His	Met	Cys	Trp
Neurospora crassa	Ala	Leu	Ser	Gly	Glu	Pro	Arg	Thr	Val	Asp	Lys	Ile	Gln	Asn	Phe	Tyr	His	Met	Trp	Cys
Drosophila melanogaster	Leu	Ser	Ala	Glu	Gly	Val	Lys	Thr	Arg	Pro	Asp	Gln	Ile	Asn	Phe	Tyr	His	Met	Cys	Trp
Caenorhabditis elegans	Leu	Ser	Glu	Lys	Ala	Val	Ile	Thr	Gly	Arg	Asp	Asn	Phe	Pro	Gln	Tyr	Met	His	Cys	Trp
Leishmania major	Ala	Leu	Ser	Arg	Val	Gly	Thr	Pro	Glu	Asp	Gln	Lys	Ile	His	Phe	Asn	Tyr	Met	Cys	Trp
Sacch. cerevisiae	Leu	Ser	Lys	Ile	Glu	Asn	Thr	Asp	Val	Ala	Gly	Arg	Phe	Pro	Gln	Tyr	His	Met	Cys	Trp
Schizosacch. pombe	Leu	Ser	Glu	Lys	Ala	Ile	Val	Thr	Asp	Asn	Gly	Arg	Pro	Phe	Gln	Tyr	His	Met	Cys	Trp
Escherichia coli	Leu	Ala	Gly	Val	Ser	Ile	Glu	Thr	Arg	Asp	Lys	Gln	Pro	Asn	Phe	Tyr	Met	His	Trp	Cys
Bacillus subtilis	Leu	Ala	Ile	Glu	Lys	Gly	Val	Ser	Thr	Asp	Phe	Arg	Asn	Gln	Pro	Tyr	Met	His	Trp	Cys
Pseudom. aeruginosa	Leu	Ala	Gly	Arg	Val	Glu	Ser	Asp	Pro	Gln	Thr	Ile	Phe	Lys	Asn	Tyr	His	Met	Trp	Cys
Mesorhizobium loti	Ala	Leu	Gly	Val	Arg	Ser	Asp	Ile	Glu	Thr	Pro	Phe	Lys	Gln	Asn	Met	Tyr	His	Trp	Cys
Streptom. coelicolor A3	Ala	Leu	Gly	Val	Arg	Pro	Thr	Asp	Glu	Ser	Ile	Phe	Gln	His	Lys	Tyr	Asn	Met	Trp	Cys
Sinorhizobium meliloti	Ala	Leu	Gly	Val	Arg	Glu	Ser	Ile	Asp	Thr	Pro	Phe	Lys	Gln	Asn	Met	Tyr	His	Trp	Cys
Nostoo sp. PCC7120	Leu	Ala	Ile	Val	Gly	Ser	Glu	Thr	Gln	Arg	Lys	Asp	Pro	Asn	Phe	Tyr	His	Met	Trp	Cys
Agrobact. tumefaciens	Ala	Leu	Gly	Val	Arg	Ser	Glu	Ile	Asp	Thr	Pro	Phe	Lys	Gln	Asn	Met	Tyr	His	Trp	Cys
Ralstonia solanacearum	Ala	Leu	Gly	Val	Arg	Thr	Asp	Pro	Ser	Glu	Ile	Gln	Phe	Lys	Asn	Tyr	His	Met	Trp	Cys
Yersinia pestis	Leu	Ala	Gly	Val	Ser	Ile	Glu	Thr	Arg	Asp	Gln	Lys	Pro	Asn	Phe	Tyr	Met	His	Trp	Cys
Methanosarc. acetivorans	Leu	Glu	Ile	Gly	Ser	Ala	Val	Lys	Thr	Asp	Arg	Asn	Phe	Pro	Tyr	Gln	Met	His	Cys	Trp
Vibrio cholerae	Leu	Ala	Val	Gly	Ser	Ile	Glu	Thr	Asp	Gln	Lys	Arg	Phe	Asn	Pro	Tyr	Met	His	Trp	Cys
Escherichia coli K12	Leu	Ala	Gly	Val	Ile	Ser	Glu	Arg	Thr	Asp	Gln	Pro	Lys	Asn	Phe	Tyr	Met	His	Trp	Cys
Mycobact. tuber. CDC1551	Ala	Leu	Gly	Val	Arg	Thr	Pro	Asp	Ser	Glu	Ile	Gln	Phe	Asn	His	Tyr	Lys	Met	Trp	Cys
Mycobact. tuber. H37Rv	Ala	Gly	Leu	Val	Arg	Thr	Asp	Pro	Ser	Glu	Ile	Gln	Phe	Asn	His	Tyr	Lys	Met	Trp	Cys
Bacillus halodurans	Leu	Glu	Val	Ala	Gly	Ile	Lys	Ser	Thr	Asp	Arg	Phe	Gln	Pro	Asn	Tyr	Met	His	Trp	Cys
Clostridium acetobutylicum	Ile	Lys	Leu	Ser	Glu	Val	Asn	Gly	Ala	Asp	Thr	Phe	Tyr	Arg	Pro	Met	Gln	His	Cys	Trp
Caulobacter crescentus CB15	Ala	Leu	Gly	Val	Arg	Asp	Pro	Glu	Thr	Ser	Ile	Phe	Lys	Gln	Asn	Met	Tyr	His	Trp	Cys
Synechocystis sp. PCC6803	Leu	Ala	Gly	Val	Ile	Glu	Ser	Gln	Thr	Pro	Arg	Asp	Lys	Asn	Phe	Tyr	Met	His	Trp	Cys
Sulfolobus solfatarcus	Leu	Ile	Lys	Val	Glu	Ser	Gly	Ala	Asn	Tyr	Arg	Thr	Asp	Phe	Pro	Gln	Met	His	Trp	Cys
Mycobacterium leprae	Ala	Leu	Val	Gly	Arg	Thr	Ser	Asp	Pro	Glu	Ile	Gln	Phe	Lys	Asn	His	Tyr	Met	Trp	Cys
Brucella melitensis	Ala	Leu	Gly	Val	Arg	Ile	Glu	Ser	Asp	Thr	Pro	Lys	Phe	Gln	Asn	Met	Tyr	His	Trp	Cys
Deinococcus radiodurans	Ala	Leu	Gly	Val	Arg	Pro	Thr	Glu	Ser	Asp	Gln	Ile	Phe	Lys	Asn	Tyr	His	Met	Trp	Cys
Listeria monocytogenes	Leu	Ile	Ala	Glu	Lys	Val	Gly	Thr	Ser	Asp	Asn	Phe	Arg	Pro	Gln	Tyr	Met	His	Trp	Cys
Clostridium perfringens	Ile	Leu	Lys	Glu	Gly	Val	Asn	Ser	Asp	Ala	Thr	Phe	Tyr	Arg	Pro	Met	Gln	His	Cys	Trp

for the codons of a biological species and plot it versus the total exonic *GC* content; see Fig. 10. The Shannon entropy is rather well fitted by a parabola:

$$S = 2.2186 + 0.144 Y_{GC} - 0.00146 Y_{GC}^2, \quad \chi^2 = 0.0315. \tag{18}$$

Note that the parabola has its apex for $y \approx 0.50$, which is expected for the behavior of the Shannon entropy for two variables (here *GC* and its complementary *AU*).

The same behavior has been shown by analogous computations made by Zeeberg [16] for *Homo sapiens*. So it seems that the entropy in the gene coding sequences and in the total exonic region as functions of the exonic *GC* content show the same pattern.

In conclusion, the distribution of the experimental codon probabilities for a large total exonic region of several biological species has been very well fitted by the law of Eq. (4). The spectrum of the distribution is universal, but the codon, which occupies a fixed level, depends on the biological species. Indeed, a more detailed analysis shows that, for

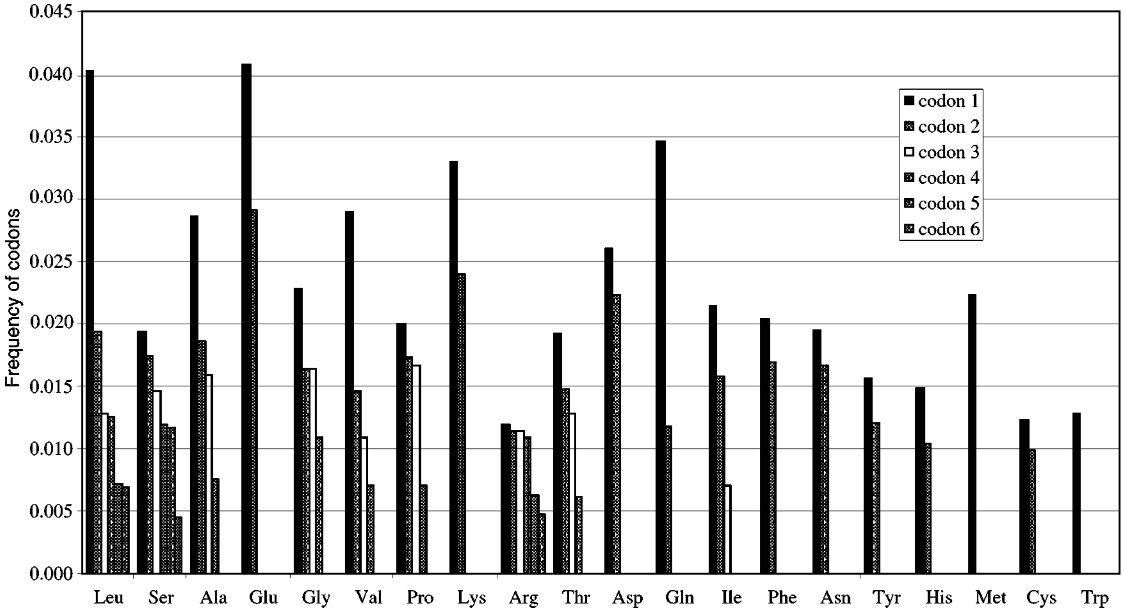


FIG. 13. Amino-acid and codon ranked distributions for *Homo sapiens*.

close biological species, e.g., vertebrates, a fixed codon occupies almost the same position in $f(n)$, while for distant biological species the codons occupy very different positions in the rank distribution. We have also derived that the codon frequency for any gene region is the same for fixed biological species and fixed *GC* content. Entropy analysis has shown that the behavior observed in genes with different *GC* content for the same biological species is very similar to that shown by the total exonic region with different *GC* content for different biological species.

ACKNOWLEDGMENTS

The authors would like to thank A. Arneodo for fruitful discussions. A.S. is indebted to the Universite´ de Savoie for financial support and LAPTH for its kind hospitality.

[1] G. K. Zipf, *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Cambridge, MA, 1949).

[2] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994).

[3] A. Czirok, R. N. Mantegna, S. Havlin, and H. E. Stanley, Phys. Rev. E **52**, 446 (1995).

[4] R. Israeloff, M. Kagalenko, and K. Chan, Phys. Rev. Lett. **76**, 1976 (1996); S. Bonhoeffer, A. V. M. Herz, M. C. Boerlijst, S. Nee, M. A. Nowak, and R. M. May, *ibid.* **76**, 1977 (1996); R. F. Voss, *ibid.* **76**, 1978 (1996); R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, *ibid.* **76**, 1979 (1996).

[5] C. Martindale and A. K. Konopka, Comput. Chem. (Oxford) **20**, 35 (1996).

[6] G. U. Yule, Philos. Trans. R. Soc. London, Ser. B **213**, 21 (1924); *The Statistical Study of Literary Vocabulary* (Cambridge University Press, Cambridge, U.K., 1944).

[7] W. Li and Y. Yang, J. Theor. Biol. **219**, 539 (2002).

[8] C. Furusawa and K. Kaneko, Phys. Rev. Lett. **90**, 088102 (2003).

[9] V. A. Kuznetsov, Signal Process. **83**, 889 (2003).

[10] A. Som, S. Chattapadhyay, J. Chakrabarti, and D. Bandyopadhyay, Phys. Rev. E **63**, 051908 (2001).

[11] C. E. Shannon, Bell Syst. Tech. J. **27**, 623 (1948).

[12] Y. Nakamura, T. Gojobori, and T. Ikemura, Nucleic Acids Res. **28**, 292 (2000).

[13] L. Frappat, A. Sciarrino, and P. Sorba, J. Biol. Phys. **17**, 1 (2001).

[14] R. D. Knight, S. J. Freeland, and L. F. Landweber, Genome Biol. **2**, 1 (2001).

[15] G. Gamow and M. Ycas, Proc. Natl. Acad. Sci. U.S.A. **41**, 1011 (1955).

[16] B. Zeeberg, Genome Res. **12**, 944 (2002).