Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 369 (2006) 699–713
PHYSICA A
www.elsevier.com/locate/physa

Conspiracy in bacterial genomes

L. Frappat[a,*], A. Sciarrino[b]

[a]*Laboratoire d'Annecy-le-Vieux de Physique Théorique LAPTH, CNRS, UMR 5108 BP 110, F-74941 Annecy-le-Vieux Cedex, France*
[b]*Dipartimento di Scienze Fisiche, Università di Napoli "Federico II" and I.N.F.N., Sezione di Napoli, Complesso Universitario di Monte S. Angelo, Via Cintia, I-80126 Napoli, Italy*

Received 20 July 2005; received in revised form 3 February 2006
Available online 20 March 2006

Abstract

The rank ordered distribution of the codon usage frequencies for 109 eubacteria and 14 archaea is best fitted by a three-parameter function which is the sum of a constant, an exponential and a linear term in the rank n. The parameters depend (two parabolically) from the total GC content. The rank ordered distribution of the amino acids is fitted by a straight line. The Shannon entropy computed over all the codons is well fitted by a parabola in the GC content, while the partial entropies computed over subsets of the codons show peculiar different behavior. Moreover, the sum of the codon usage frequencies over particular sets, e.g. with C and A (respectively G and U) as ith nucleotide, show a clear linear dependence from the GC content, exhibiting a conspiracy effect.

Keywords: Bacterial genomes; Codon distribution; GC content

1. Introduction

The genetic code is degenerate, referring to the fact that almost all the amino acids are encoded by 'synonymous' codons, and this degeneracy is primarily found in the third position of the codon. Some codons are used much more frequently than others to encode a particular amino acid, the pattern of codon usage varying between species. In the last few years, the number of available data for coding sequences has considerably increased [1], allowing for analysis to look for regularities, correlations and general features over the whole exonic region. Recently, from an analysis of the rank distribution for codons, in RNA coding sequences[1] performed in many genes for several biological species, the existence of a universal, i.e. biological species independent, distribution law for codons for the eukaryotic code has been remarked [2]. Indeed, it was pointed out that the rank of codon usage probabilities follows a universal law, the frequency function of the rank ordered codons being very nicely fitted by a sum of an exponential, a linear part and a constant. Such a

*Corresponding author.
E-mail addresses: frappat@lapp.in2p3.fr (L. Frappat), sciarrino@na.infn.it (A. Sciarrino).

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

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2006.02.008

universal behaviour suggested the presence of general biases, one of which was identified with the total *exonic* GC content (denoted throughout the paper by y_{GC}, $0 \leqslant y_{\mathrm{GC}} \leqslant 1$), which is well known to play a strong role in the evolutionary process. In fact, the values of the parameters appearing in the fitting expression, plotted vs. the total exonic GC content of the biological species, were reasonably well fitted by a parabolic function of the GC content. It is worthwhile to recall that the determination of the kind of law the codon rank distribution follows is extremely interesting in the investigations of the nature of the evolutionary process, which has acted upon the codon distribution, i.e. the eventual presence of a bias.

Possible origins of codon bias have been recognised either as the result of natural selection or as the result of mutational pressure acting on the whole genome, also known as neutral evolution theory [3,4]. Experimental evidence has showed now that both effects should occur (see e.g. Ref. [5] for the influence of the biased mutation pressure in the perspective of the neutral theory of molecular evolution, Ref. [6] for exhibiting the role of external selective forces in the case of thermophilic bacteria, Ref. [7] for an analysis of genome-wide codon bias of bacterial species, Ref. [8] for a discussion on the relative role of mutational bias and translational selection for codon usage in many genomes, including bacteria ones, Ref. [9] for a model explaining trends in codon and amino-acid usage vs. GC content). Quantitative model of directional mutational pressure has been proposed to measure to some extent the relative role of selection constraints and neutrality [10,11]. Many efforts have also been made to study codon bias among genes of a given genome, in particular, the influence of the gene expression level [12–14], of the tRNA abundance Ref. [15,16], of translational accuracy [17], see also Ref. [9] and references therein.

It seemed to us interesting to continue our previous analysis on a 'trend across species' basis by focusing on the bacteria, one main interest of the present study arising from the wide variation of the total GC content ranging from 25% to 75 %, whereas the GC variation inside a bacterial genome is much smaller. Our results is that the rank ordered distribution of the codon usage frequencies for bacteria is best fitted by a three-parameter function, which is the sum of a constant, an exponential and a linear term in the rank n. Two of the three parameters depend parabolically on the total exonic GC content. As the sum, over suitable sets below defined, of the codon usage frequencies is well fitted by a straight line in the GC content, we say that *conspiracy* effect has to be present. Moreover, we conjecture the existence of an averaged discrete symmetry in the codon usage frequencies, reasonably confirmed by the data. We also calculate the rank ordered distribution of the 20 amino acids, which is satisfactorily fitted by a straight line in y_{GC}.

We compute the Shannon entropy (as defined in Ref. [18]) and find that its behaviour in function of the exonic GC content is a parabola, whose apex is around the value 0.50 of the GC content, which is expected for the behaviour of the Shannon entropy for two variables. Moreover, the Shannon entropies for the codons, whose orders in rank are, respectively, in the ranges 1–15, 16–25 and 26–61, have peculiar features, which we comment below.

The study of this paper was based on a sample of 109 bacteria and 14 archaea, with a codon statistics larger than 300 000. The data were taken from Codon Usage Tabulated from GenBank [1] (see also http://www.kazusa.or.jp/codon/), release 138 for eubacteria and mainly release 144 for archaea.

2. Codon usage probabilities distribution

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

$$P(XZN) = \lim_{n_{tot} \to \infty} \frac{n_{XZN}}{N_{tot}}, \tag{1}$$

where n_{XZN} is the number of times the codon XZN has been used in all considered processes, for a given biological species, and N_{tot} is the total number of codons used in the same processes. It follows that our analysis and predictions hold for biological species with sufficiently large statistics of codons. For each biological species, codons are ordered following the 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 XZN in the nth position. Of course the same codon occupies in general two different positions in the rank distribution function for two different species. By plotting $f(n)$ vs. the rank we confirm that the data are best fitted by the kind of function we found in Ref. [2], i.e. the sum of

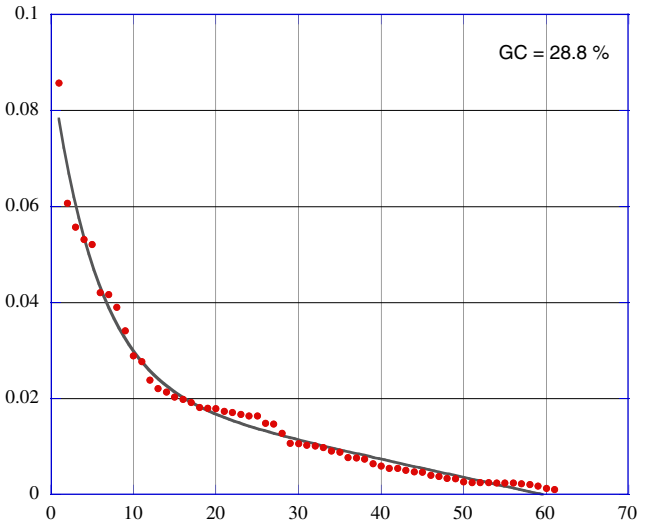

Fig. 1. Codon rank distribution $f(n)$ for *B. burgdorferi*.

an exponential function, a linear function and a constant:

$$f(n) = \alpha\, \mathrm{e}^{-\eta n} - \beta n + \gamma. \tag{2}$$

In the following analysis we shall not consider the three stop codons, whose function is very peculiar. Therefore, the four parameters α, η, β and γ are constrained by the normalisation condition

$$\xi = \frac{\alpha\, \mathrm{e}^{-\eta}}{1 - \mathrm{e}^{-\eta}} - 1891\beta + 61\gamma, \tag{3}$$

where $\xi < 1$ is a number that is computed for any biological species from the codon usage frequency for the 61 encoding codons (note that the result is almost unchanged if the data are normalised to one summing over the 64 coding codons).[2]

In the following, the parameters for the different fits have been computed using a best-fit procedure, the curve fit being based on the Levenberg–Marquardt algorithm [19]. The χ^2 coefficient is defined by

$$\chi^2 = \sum_i \frac{(y_i - y'_i)^2}{y'_i} \tag{4}$$

and the Pearson's R coefficient by

$$R = \frac{\sum_i (y_i - \overline{y})(y'_i - \overline{y}')}{\sqrt{\sum_i (y_i - \overline{y})^2}\sqrt{\sum_i (y'_i - \overline{y}')^2}}, \tag{5}$$

where y_i are the actual values, y'_i are the calculated ones, $\overline{y}$ and $\overline{y}'$ are the means of the actual values and of the calculated ones, respectively. Recall that R^2 represents the ratio of the explained variance on the total variance. Fits can be considered as satisfactory for values of $R^2 \gtrsim 0.8$.

In Figs. 1–4, we report the plot of $f(n)$ as a function of n for a few bacteria: *Borrelia burgdorferi* (GC = 28.8%), *Bacillus subtilis* (GC = 44.4%), *Escherichia coli* (GC = 51.8%), and *Ralstonia solanacearum* (GC = 67.5%). One remarks that the codon rank distributions show mainly a not far from uniform or slowly

[2]The procedure to fit $f(n)$ in the present paper is slightly different from the one followed in Ref. [2]. In that paper the value of the parameter γ was fixed to the value corresponding to a uniform distribution, i.e. $\gamma = 1/61 \simeq 0.0164$. Therefore, we were left with only two free parameters, while in the present paper we use three free parameters. While the general feature of the fit are unchanged, the present fits are more accurate with a χ^2 lower of about two orders of magnitude.

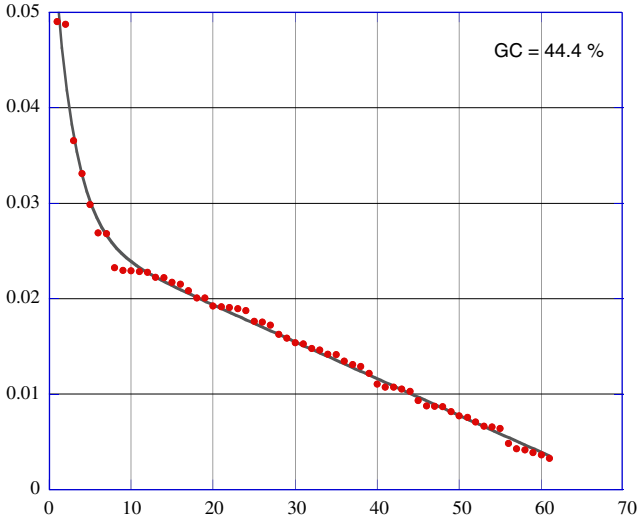

Fig. 2. Codon rank distribution $f(n)$ for *B. subtilis*.

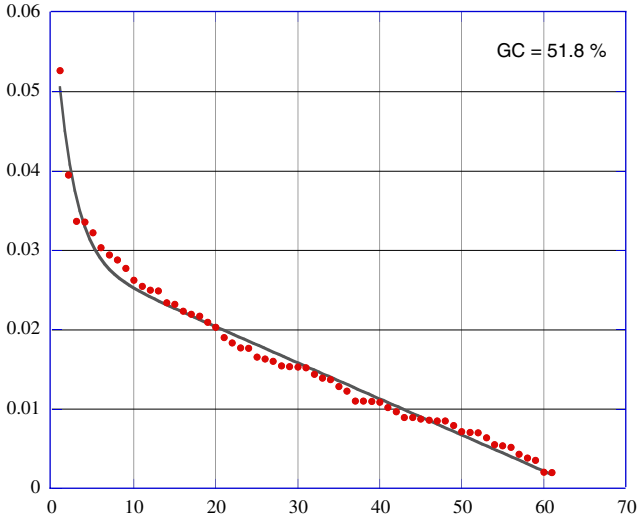

Fig. 3. Codon rank distribution $f(n)$ for *E. coli*.

decreasing behaviour for most of the codons, and a peak of a small number of highly 'overrepresented' codons (codon bias contribution).[3]

Note that the fits of the rank ordered distribution $f(n)$ by a Yule law or Zipf law are unsatisfactory. Indeed, as emphasised in Ref. [20], when the majority of points resides in the tail of the distribution, it is necessary to fit the whole range of data. From the previous discussion, we expect the parameters to depend on the total GC content of the genes region (here the total exonic GC content). We have investigated this dependence and we report in Figs. 5–8 the plots of the parameters α, β, γ and η vs. the total exonic GC content. In terms of the total exonic GC content y_{GC} of the biological species, one finds that the values of α and γ are well fitted by

[3]Statistical tables specifying for each biological species the codon statistics, the total GC content, the values of the parameters α, β, η, computed by a best-fit procedure, the estimated errors $\Delta\alpha$, $\Delta\beta$, $\Delta\eta$ on these parameters, the corresponding estimators of goodness of the fit χ^2 and R^2 can be found on the archives at the following address: http://xxx.lanl.gov/abs/q-bio.GN/0507030v1.

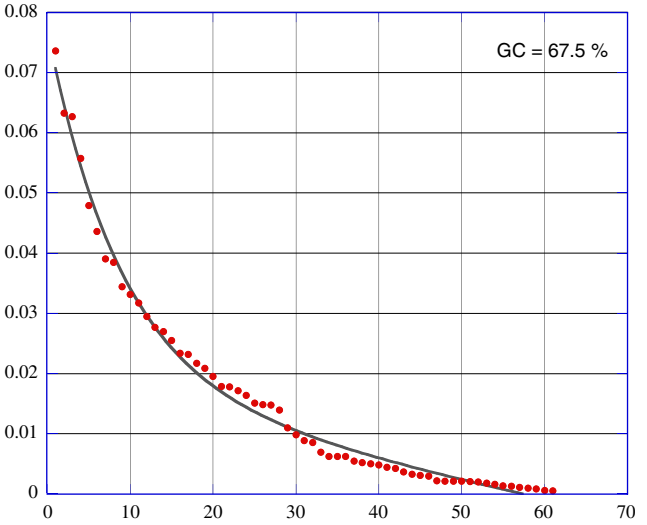

Fig. 4. Codon rank distribution $f(n)$ for *R. solanacearum*.

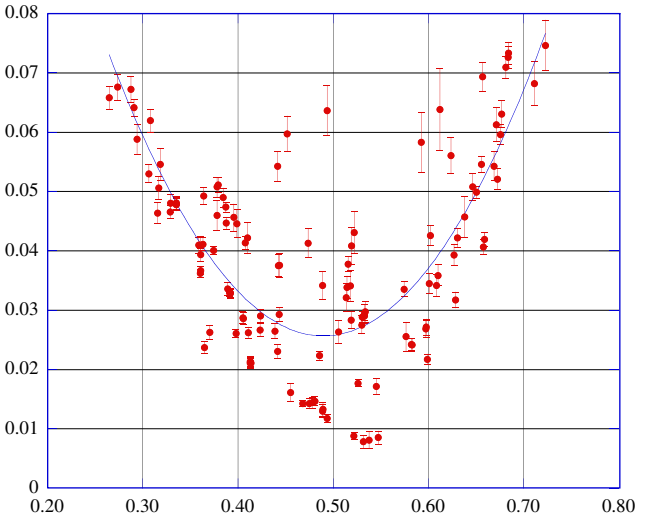

Fig. 5. Parameter α vs. GC content.

polynomial functions:

$$\alpha = \alpha_0 + \alpha_1 y_{GC} + \alpha_2 y_{GC}^2, \tag{6}$$

where

$$\alpha_0 = 0.250 \pm 0.017, \quad \alpha_1 = -0.919 \pm 0.071, \quad \alpha_2 = 0.939 \pm 0.072, \tag{7}$$

the goodness of the fit being given by

$$\chi^2 = 0.013 \quad \text{and} \quad R = 0.768, \tag{8}$$

and

$$\gamma = \gamma_0 + \gamma_1 y_{GC} + \gamma_2 y_{GC}^2, \tag{9}$$

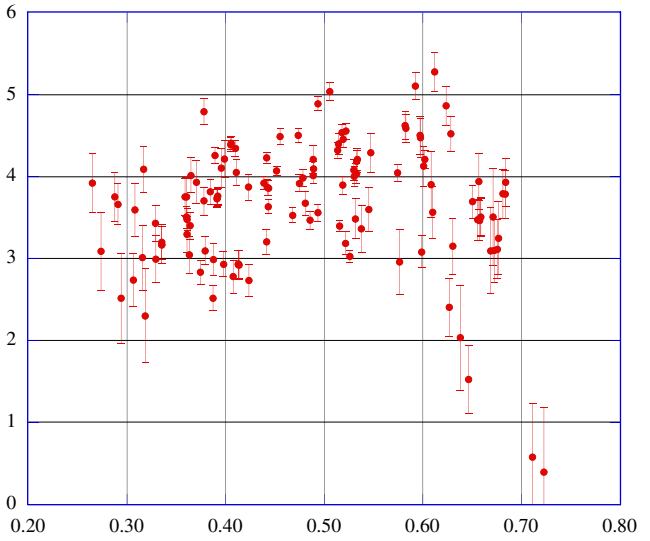

Fig. 6. Parameter $10^4\beta$ vs. GC content.

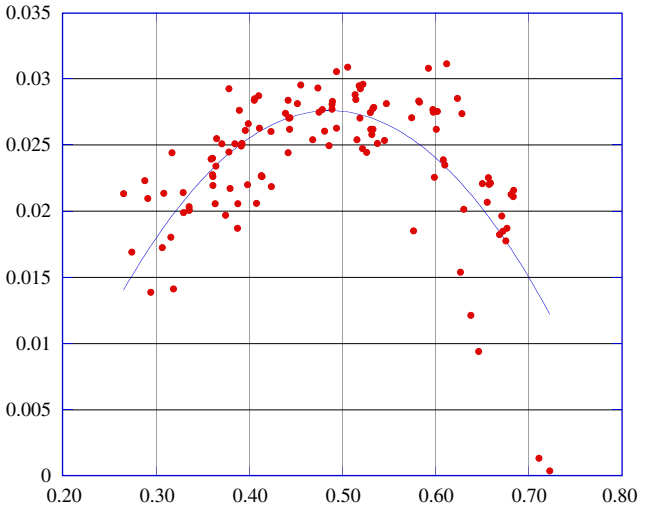

Fig. 7. Parameter γ vs. GC content.

where

$$\gamma_0 = -0.0376 \pm 0.0057, \quad \gamma_1 = 0.268 \pm 0.024, \quad \gamma_2 = -0.275 \pm 0.024, \tag{10}$$

the goodness of the fit being given by

$$\chi^2 = 0.0015 \quad \text{and} \quad R = 0.728. \tag{11}$$

The η parameter shows a 'λ-like' behaviour in terms of y_{GC}, centred on the value $y_{GC} = 0.50$, while the values of the β parameter are mainly in the range $3 \times 10^{-4} \leqslant \beta \leqslant 5 \times 10^{-4}$. Note that the errors on β become large when y_{GC} is far from the mean value 0.50.

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

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

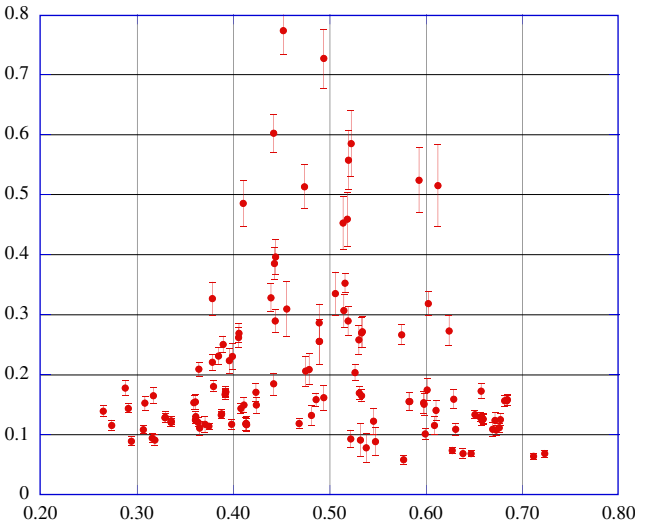

Fig. 8. Parameter η vs. GC content.

where the sum is over the set J 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.

3. 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 bacterial 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,

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

In terms of the total exonic GC content, the parameters B and F for eubacteria show a parabolic dependence (Figs. 9 and 10): $B = B_0 + B_1 y_{\mathrm{GC}} + B_2 y_{\mathrm{GC}}^2$ where

$$B_0 = 0.00915 \pm 0.00034, \quad B_1 = -0.0232 \pm 0.0014, \quad B_2 = -0.0258 \pm 0.0014, \tag{14}$$

the goodness of the fit being given by

$$\chi^2 = 4.4 \times 10^{-6} \quad \text{and} \quad R = 0.911, \tag{15}$$

and $F = F_0 + F_1\, y_{\mathrm{GC}} + F_2\, y_{\mathrm{GC}}^2$ where

$$F_0 = 0.145 \pm 0.004, \quad F_1 = -0.243 \pm 0.015, \quad F_2 = 0.270 \pm 0.015, \tag{16}$$

the goodness of the fit being given by

$$\chi^2 = 48 \times 10^{-5} \quad \text{and} \quad R = 0.911. \tag{17}$$

One remarks that the most frequent amino acid is always above the line. This can be easily understood in the light of Eq. (2). Indeed, the most frequent amino acids gets, in general, a contribution of the exponential term of (2) with a low value of n.

Of course, the frequency of an amino acid is correlated to frequencies of its encoding codons given by (2). If the ranks of the encoding codons were completely random, we do not expect that their sum should take equally spaced values, as is the case in a regression line. Therefore, we can infer, for the biological species whose amino-acid frequencies are very well fitted by a line, the existence of some functional constraints (conspiracy effect) on the codon usage. An analysis of the influence of the total GC content on the amino-acids composition of the protein of 59 bacteria has been reported in Ref. [21], where references to previous works on

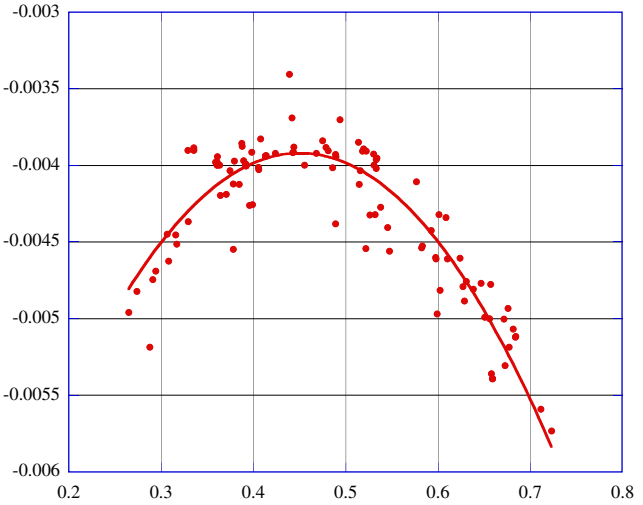

Fig. 9. Parameter B vs. GC content (eubacteria).

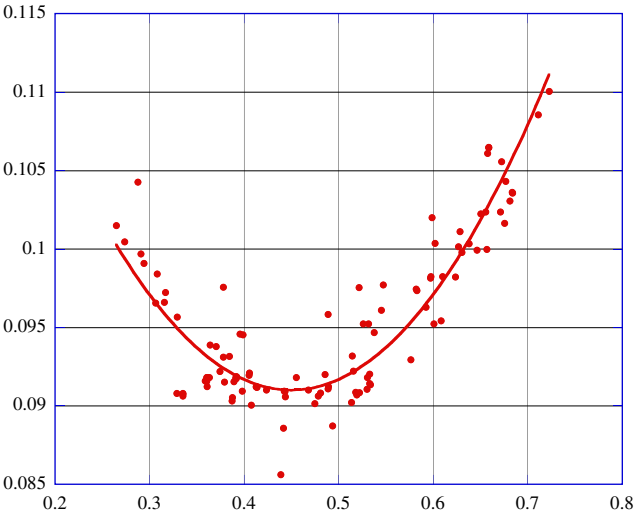

Fig. 10. Parameter F vs. GC content (eubacteria).

the subject can be found, see also Refs. [10,11]. The figures, reported in Ref. [21], qualitatively agree with the linear behavior given by Eq. (13) and show sensible quantitative differences with the expected frequencies, computed on the basis of the neutral model, where the frequency of an amino acid is the sum of the synonymous codon frequencies, the frequency of a codon being computed as the product of the probability of the three nucleotides (mean field model). This analysis already hints in the direction of the presence of functional constraints.

However, the behaviour predicted by (2) fits the experimental data very well, while the shape of the distribution of amino acids seems more sensible to the biological species. In fact, one can remark on many plots of the amino-acid distributions the existence of one or two plateaux, which obviously indicate an equal probabilities of use for some amino acids. Presently, we have neither any arguments to explain the uniform distribution of amino acids from the ranked distribution of the corresponding codons nor we know of any explanation of this pattern of distribution.

4. The Shannon entropy

We compute the Shannon entropy,

$$S = -\sum_{n=1}^{n=61} f(n)\log_2 f(n), \tag{18}$$

related to the codon rank distribution and plot it vs. the total exonic GC content y_{GC} for the sample of 109 eubacteria and 14 archaea, see Fig. 11. The contribution of the stop codons is negligible and represents less than 0.5% of the total entropy. The Shannon entropy is very well fitted by a parabola

$$S = s_0 + s_1 y_{GC} + s_2 y_{GC}^2, \tag{19}$$

where

$$s_0 = 2.670 \pm 0.084, \quad s_1 = 12.36 \pm 0.35, \quad s_2 = -12.73 \pm 0.36, \tag{20}$$

the goodness of the fit being given by

$$\chi^2 = 0.329 \quad \text{and} \quad R^2 = 0.916. \tag{21}$$

Note that the parabola has its apex for $y_{GC} \approx 0.50$, which is expected for the behaviour of the Shannon entropy for two variables (here GC and AU). We have computed the partial Shannon entropies for the codons whose orders in rank are, respectively, in the ranges 1–15, 16–25 and 26–61 and we report them in Fig. 11 vs. the GC content. In any of the three sets, we have put the codons whose contributions to the entropy, with respect the GC content, have similar behaviour. Indeed, the codons of the first set are primarily influenced by the exponential term in the rank distribution (2), leading for the partial entropy S_{1-15} to a parabolic behaviour with positive curvature and minimum at 50% GC content. For the codons of the intermediate set, the exponential term and the last two terms are of the same order of magnitude, hence the partial entropy S_{16-25} is almost uniform with respect to the GC content. For the codons of the last set, the exponential term is completely negligeable, and since $f(n)$ is very small, $-f(n)\log_2 f(n) \simeq f(n)/\ln 2$. The trend of the partial entropy S_{26-61} is thus essentially given by the behaviour of the γ parameter, hence the parabolic shape with negative curvature and maximum at 50% GC content.

The fact that the Shannon entropy is a parabola shows obviously that the codon distribution is not uniform: the uniform distribution corresponds to the maximal entropy $S = \log_2 N$, independent of the GC content,

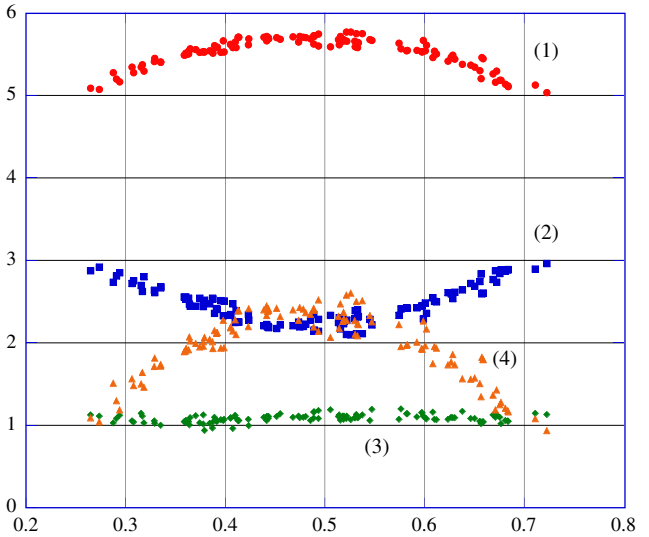

Fig. 11. Shannon entropy and partial entropies vs. GC content: (1) total entropy, (2) partial entropy rank 1–15, (3) partial entropy rank 16–25, (4) partial entropy rank 26–61.

Table 1
Regression coefficients for $\sum P(X_i) = a\, y_{\mathrm{GC}} + b$ (eubacteria)

$\sum P(X_i)$	a	b	R (Pearson)
$\sum P(\mathrm{C}NN')$	0.424	0.004	0.949
$\sum P(\mathrm{A}NN')$	−0.452	0.490	0.960
$\sum P(\mathrm{G}NN')$	0.283	0.209	0.928
$\sum P(\mathrm{U}NN')$	−0.255	0.297	0.955
$\sum P(N\mathrm{C}N')$	0.239	0.108	0.925
$\sum P(N\mathrm{A}N')$	−0.338	0.467	0.959
$\sum P(N\mathrm{G}N')$	0.215	0.065	0.955
$\sum P(N\mathrm{U}N')$	−0.117	0.358	0.898
$\sum P(NN'\mathrm{C})$	1.067	−0.257	0.984
$\sum P(NN'\mathrm{A})$	−0.928	0.676	0.985
$\sum P(NN'\mathrm{G})$	0.770	−0.130	0.984
$\sum P(NN'\mathrm{U})$	−0.909	0.711	0.977

where N is the number of considered codons (here $N = 61$). The concavity of the entropy is also well understood in this context in terms of deviations to the uniform distribution when y_{GC} is far from the mean value 50%. On the contrary, the partial entropy S_{1-15}, which is associated to the 'overrepresented' codons, exhibits a convex shape. This feature can also be understood by realizing that in the case of y_{GC} far from 50%, the overrepresented codons distribution would be closer to the expected value for a nonbiased distribution than in the case of $y_{\mathrm{GC}} \approx 50\%$.

5. The mystery of the straight lines

In recent papers [22,23], it has ben remarked, for 129 eubacteria and for 19 archaea, that the codon-specific nucleotide frequency P^i_X, i.e. the sum of the codon usage for any given nucleotide X in any fixed ith position, that is, for example

$$P^1_X = \sum_{Y,Z} P(XYZ), \tag{22}$$

is fitted, vs. the GC content, by a straight line, with coefficients depending on the position and on the nature of the nucleotide and slightly different for eubacteria and archaebacteria (see Table 1 for eubacteria and Table 2 for archaea for the values of the slope and the axis intercept in the case of our sample of bacteria).[4] There are 12 codon-specific nucleotide frequencies, which are constrained by three normalisation conditions

$$P^i_{\mathrm{C}} + P^i_{\mathrm{U}} + P^i_{\mathrm{G}} + P^i_{\mathrm{A}} = 1, \quad i = 1, 2, 3. \tag{23}$$

The fact that the sum of the codon position-specific nucleotide frequencies is a linear function of GC is a mysterious feature, which becomes more mysterious in the light of the results of Section 2, where we have remarked that the parameters appearing in the rank ordered frequency (2) are parabolic functions of GC. A *conspiracy* has to be present between the $f(n)$, see Eq. (2), for any bacteria such that the sum of $f(n)$ over the 16-dimensional sets Q, depending on the considered bacteria and the corresponding codon position-specific frequencies, produces a linear function of GC, that is

$$\sum_{n \in Q} f(n) = a\, y_{\mathrm{GC}} + b. \tag{24}$$

Things become still more mysterious taking into accounts what we have remarked, guided by the mathematical structure of the *crystal basis model* of the genetic code [24]. Let us recall that in this model the

[4]Note that the difference of behaviour between eubacteria and archaea could stem from the fact that eubacteria (at least in our sample) are mainly mesophilic while archaea are essentially thermophilic.

Table 2
Regression coefficients for $\sum P(X_i) = a\, y_{\mathrm{GC}} + b$ (archaea)

$\sum P(X_i)$	a	b	R (Pearson)
$\sum P(\mathrm{C}NN')$	0.414	−0.028	0.951
$\sum P(\mathrm{A}NN')$	−0.557	0.569	0.963
$\sum P(\mathrm{G}NN')$	0.393	0.173	0.917
$\sum P(\mathrm{U}NN')$	−0.249	0.286	0.901
$\sum P(N\mathrm{C}N')$	0.269	0.079	0.891
$\sum P(N\mathrm{A}N')$	−0.285	0.440	0.841
$\sum P(N\mathrm{G}N')$	0.190	0.082	0.859
$\sum P(N\mathrm{U}N')$	−0.174	0.398	0.845
$\sum P(NN'\mathrm{C})$	1.047	−0.241	0.982
$\sum P(NN'\mathrm{A})$	−0.923	0.687	0.988
$\sum P(NN'\mathrm{G})$	0.687	−0.066	0.943
$\sum P(NN'\mathrm{U})$	−0.812	0.619	0.983

four nucleotides are assigned to the 4-dim fundamental irreducible representation $(\frac{1}{2},\frac{1}{2})$ of $\mathscr{U}_{q\to 0}(so(4)) \equiv \mathscr{U}_{q\to 0}(sl(2)\oplus sl(2)) = \mathscr{U}_{q\to 0}(sl_H(2)\oplus sl_V(2))$ (the lower labels H and V denote the two commuting $sl(2)$), with the state assignment (in the following the first number denotes the value of the label J specifying the irreducible representation (irrep.) of $sl(2)$ and the second one the value of the label J_3 denoting the state in the irrep., $J_3 = J, J-1, \ldots, -J$, $2J \in \mathbb{Z}_+$)

$$\begin{aligned}
&\mathrm{C} \equiv (\mathbf{1/2},\mathbf{1/2})_H, (\mathbf{1/2},\mathbf{1/2})_V,\\
&\mathrm{U} \equiv (\mathbf{1/2},\mathbf{-1/2})_H, (\mathbf{1/2},\mathbf{1/2})_V,\\
&\mathrm{G} \equiv (\mathbf{1/2},\mathbf{1/2})_H, (\mathbf{1/2},\mathbf{-1/2})_V,\\
&\mathrm{A} \equiv (\mathbf{1/2},\mathbf{-1/2})_H, (\mathbf{1/2},\mathbf{-1/2})_V,
\end{aligned} \tag{25}$$

which, in matrix notation, can be written as

$$\begin{pmatrix} \mathrm{C} & \mathrm{U} \\ \mathrm{G} & \mathrm{A} \end{pmatrix}.$$

The codons are the composite states of the three-fold tensor product of the fundamental irreducible representation, that is

$$\begin{pmatrix} \mathrm{C} & \mathrm{U} \\ \mathrm{G} & \mathrm{A} \end{pmatrix}^{\otimes 3},$$

which can be written in matrix form as

$$\begin{pmatrix}
\mathrm{CCC} & \mathrm{CCU} & \mathrm{CUC} & \mathrm{CUU} & \mathrm{UCC} & \mathrm{UCU} & \mathrm{UUC} & \mathrm{UUU} \\
\mathrm{CCG} & \mathrm{CCA} & \mathrm{CUG} & \mathrm{CUA} & \mathrm{UCG} & \mathrm{UCA} & \mathrm{UUG} & \mathrm{UUA} \\
\mathrm{CGC} & \mathrm{CGU} & \mathrm{CAC} & \mathrm{CAU} & \mathrm{UGC} & \mathrm{UGU} & \mathrm{UAC} & \mathrm{UAU} \\
\mathrm{CGG} & \mathrm{CGA} & \mathrm{CAG} & \mathrm{CAA} & \mathrm{UGG} & \mathrm{UGA} & \mathrm{UAG} & \mathrm{UAA} \\
\mathrm{GCC} & \mathrm{GCU} & \mathrm{GUC} & \mathrm{GUU} & \mathrm{ACC} & \mathrm{ACU} & \mathrm{AUC} & \mathrm{AUU} \\
\mathrm{GCG} & \mathrm{GCA} & \mathrm{GUG} & \mathrm{GUA} & \mathrm{ACG} & \mathrm{ACA} & \mathrm{AUG} & \mathrm{AUA} \\
\mathrm{GGC} & \mathrm{GGU} & \mathrm{GAC} & \mathrm{GAU} & \mathrm{AGC} & \mathrm{AGU} & \mathrm{AAC} & \mathrm{AAU} \\
\mathrm{GGG} & \mathrm{GGA} & \mathrm{GAG} & \mathrm{GAA} & \mathrm{AGG} & \mathrm{AGA} & \mathrm{AAG} & \mathrm{AAA}
\end{pmatrix}. \tag{26}$$

Eq. (23) can be rewritten, for instance for $i = 1$, as

$$P^1_{\mathrm{C}} + P^1_{\mathrm{A}} - 1/2 = -P^1_{\mathrm{U}} - P^1_{\mathrm{G}} + 1/2 \tag{27}$$

and, looking at the codon matrix, we remark that the l.h.s., respectively the r.h.s., is, up to $\mp 1/2$, the sum of the two 4×4 matrices obtained by *mirror inversion* with respect to the centre of the matrix, that is the exchange C $\to$ A, G $\to$ U.[5] So we make the following conjecture. The sum of codon usage frequencies over a set of r codons ($4\leqslant r\leqslant 16$), placed on the same side of the diagonals of the matrix, plus the mirror set is described by a straight line in y_{GC} with a very small slope coefficient a ($|a|<0.1$) and a constant term b equal to $r/32$. In other words, the sum of codon usage frequencies over a set of r codons, placed on the same side of the diagonals of the matrix with a plus the numerical factor $-r/16$, is equal, up a factor $r/32$, to the opposite of the sum of the codon usage frequencies over the set of r mirror codons. The mirror symmetry with respect to the secondary diagonal (resp. principal diagonal) is equivalent, in terms of $\mathscr{U}_{q\to 0}(sl_H(2)\oplus sl_V(2))$ to the change of the sign of the third component of $J_{3,H}$ and $J_{3,V}$ (resp. of the sign of the third component of $J_{3,V}$). Denoting $P^i(XYZ)$ (resp. $P^i(\overline{XYZ})$) the codon usage frequency of the ith codon XYZ (resp. of the mirror codon $\overline{XYZ}$), we make the ansatz that the following equation holds:

$$\sum_{k\in I}\left(P^k(XYZ)+P^k(\overline{XYZ})\right)=a\,y_{\mathrm{GC}}+\frac{r}{32}, \tag{28}$$

where I is r-dimensional set of codons placed on the same side of the principal diagonal or secondary diagonal of the codon matrix . The mirror codon $\overline{XYZ}$ with respect to the secondary diagonal (resp. to the principal diagonal) has opposite values of $J_{3,H}$ and $J_{3,V}$ of the XYZ codon (respectively, opposite values of $J_{3,V}$). One can see from Table 3 that indeed the values of the axis intercept b are very close to the conjectured values $r/32$ ($\frac{4}{32}=0.125$, $\frac{8}{32}=0.250$, $\frac{12}{32}=0.375$). So our conjecture seems supported by the experimental data and suggests the existence of an *averaged symmetry* C $\to$ A, G $\to$ U or C $\to$ G, U $\to$ A for the codon usage frequencies.[6] Of course when $2r$ is equal to the number of the codons in the considered set, the above conjecture is trivially the normalisation condition.

An analysis over our sample of 109 eubacteria, randomly drawing 4, 8, 12 codons in the 16-dim set of codons with a specified nucleotide in a fixed position, shows that the straight line behaviour is surprisingly present already for 4 codons (see Table 3), where we present few examples which confirm the conjecture. In this table, N denotes the number of random drawings of n different codons belonging to a given set I (e.g. in the first line of Table 3, $I=\{CNN'\}$, $N,N'=\mathrm{A,C,G,U}$). For each drawing lot of codons in the set I, we expect the sum $\sum_{XYZ\in I}P(XYZ)+P(\overline{XYZ})$ to behave linearly in terms of the GC content y_{GC}, with slope a and axis intercept b. We observe that the distribution of the coefficients a and b is peaked around mean values $\overline{a}$ and $\overline{b}$ with standard deviations σ_a and σ_b, which are reported in Table 3. Two particular subsets have also been considered, $I_p=\{\mathrm{UNN',CUN,CCC,CCU,CCA,CAC,CAU,CAA,AUN,ACU,AAU}\}$ and $I_s=\{\mathrm{CNN',GCN,GGC,GGG,GGU,GUC,GUG,GUU,UCN,UUC,UGC}\}$, corresponding respectively to codons above the principal and secondary diagonals of the codon matrix (26). We have summed the experimental codon usage of these codons and of the n mirror codons with respect to the secondary diagonal. As it can be read from Tables 1 and 3, no real meaningful difference appears between $n<16$ and $n=16$ for the first six sets I of Table 3, which correspond to fix the C or G codons (hence the A or U codons) in first, second and third position, respectively. Note that the number of possible different configurations is 1946 for $n=4$ or 12 and 12 870 for $n=8$. Surprisingly, this behaviour is more evident for codons with fixed nucleotides in *third* position, that is, for codons encoding different amino acids, strongly suggesting strong constraints between amino acids.

Finally, let us remark that such a behaviour does not show up any longer if one chooses for I sets which violates the diagonal symmetry (for example, the first four lines or the first four columns of the codon matrix (26), or the whole set of codons) and considers random drawings in the same conditions as above.

[5]The mirror inversion in the crystal basis formalism means change of the sign of $J_{3,H}$ and $J_{3,V}$.

[6]These symmetries have already appeared in the literature in a different context. Indeed Rumer [25], as quoted in [26], remarked in the sixties, that the first symmetry exchanges the 32 *strong codons*, which form quartets or the quartet subset of sextets, with the remaining 32 *weak codons*. This symmetry is sometimes referred to as the Rumer symmetry [26]. The second symmetry has been remarked by Konopel'chenko and Rumer [27] as the symmetry which transforms each of the *octets* into itself, an *octet* being a set of eight dinucleotides appearing or not appearing as the first two nucleotides in the quartets.

Table 3
Sum of frequencies for drawing lots of n codons belonging to some subset I of size N.

sum	n	N	$\overline{a}$	σ_a	$\overline{b}$	σ_b
$\sum_{N,N'} P(CNN') + P(\overline{CNN'})$	4	1800	−0.007	0.107	0.123	0.055
idem	8	4000	−0.013	0.12	0.247	0.064
idem	12	1800	−0.02	0.107	0.370	0.055
$\sum_{N,N'} P(NCN') + P(\overline{NCN'})$	4	1800	−0.024	0.113	0.144	0.060
idem	8	4000	−0.048	0.13	0.287	0.07
idem	12	1800	−0.074	0.11	0.432	0.06
$\sum_{N,N'} P(NN'C) + P(\overline{NN'C})$	4	1800	0.034	0.116	0.105	0.065
idem	8	4000	0.067	0.134	0.21	0.074
idem	12	1800	0.10	0.117	0.314	0.065
$\sum_{N,N'} P(GNN') + P(\overline{GNN'})$	4	1800	0.007	0.115	0.126	0.061
idem	8	4000	0.013	0.132	0.253	0.070
idem	12	1800	0.021	0.115	0.379	0.061
$\sum_{N,N'} P(NGN') + P(\overline{NGN'})$	4	1800	0.025	0.108	0.106	0.035
idem	8	4000	0.050	0.124	0.211	0.063
idem	12	1800	0.074	0.108	0.318	0.055
$\sum_{N,N'} P(NN'G) + P(\overline{NN'G})$	4	1800	−0.035	0.103	0.145	0.049
idem	8	4000	−0.074	0.120	0.293	0.057
idem	12	1800	−0.104	0.103	0.436	0.049
$\sum_{N,N',N''} P(NN'N'') + P(\overline{NN'N''})$	16	10000	−0.0024	0.21	0.50	0.112
$\sum_{N,N',N''} P(NN'N'') + P(\overline{NN'N''})$	24	10000	−0.0042	0.245	0.75	0.128
$\sum_{N,N',N''} P(NN'N'')$	32	10000	$8\ 10^{-4}$	0.30	0.50	0.147
$\sum_{N,N',N''\in I_s} P(NN'N'')$	16	10000	0.45	0.20	0.019	0.088
$\sum_{N,N',N''\in I_s} P(\overline{NN'N''})$	16	10000	−0.445	0.189	0.48	0.106
$\sum_{N,N',N''\in I_s} P(NN'N'') + P(\overline{NN'N''})$	16	10000	0.002	0.18	0.50	0.094
$\sum_{N,N',N''\in I_d} P(NN'N'')$	16	10000	−0.232	0.189	0.328	0.098
$\sum_{N,N',N''\in I_d} P(\overline{NN'N''})$	16	10000	0.229	0.226	0.173	0.113
$\sum_{N,N',N''\in I_d} P(NN'N'') + P(\overline{NN'N''})$	16	10000	−0.003	0.18	0.50	0.094

So, we derive the conclusion that the sum of the codon usage frequencies over a suitable set of codons is a linear function of GC, showing a *conspiracy* effect between the $f(n)$ similar to the one of Eq. (24), that is

$$\sum_{n\in Q} f(n) = a\, y_{\mathrm{GC}} + b, \tag{29}$$

where now Q is any $2r$-dimensional set of codons satisfying the above discussed characteristics.

The results above were based on sums of codons of the type $P(XYZ) + P(\overline{XYZ})$, the bar meaning the *mirror* codon obtained by the mirror symmetry C ↔ A, G ↔ U. It is interesting to see what happens if one considers instead sums of codons and their *reverse complementary* codons, i.e. $P(XYZ) + P(\widehat{ZYX})$ where the hat means the complementary rule C ↔ G, A ↔ U (in other words, summing codons on one strand and complementary codons on the other strand). A preliminary analysis shows that the behaviour (28) also holds, but the slope and axis intercept ranges are wider and the mean slope is not close to zero. Such an analysis could be instructive because it might shed some light on the DNA strand asymmetry [28–31]. However, in the present

work, we analyse the coding sequences extracted from the CUTG database, for which one does not know whether the coding sequence comes from a gene on the leading strand or on the lagging strand, and the intergenic (noncoding) sequences were obviously not considered. Hence we are unable to comment on this point for the moment.

6. Discussion and conclusion

The distribution of the experimental codon probabilities for the total exonic region for a large sample of bacteria is well fitted by the law (2). The spectrum of the distribution is universal, but the codon, which occupies a fixed level, depends on the biological species. The universal form of the function $f(n)$ strongly suggests the presence of a bias of very general origin. Indeed, in Ref. [32], a mutation model has been proposed where the intensity of the mutation matrix, essentially in one point mutation, depends from the variation of the labels identifying the states of the irrep. of $U_{q\to 0}(sl(2)\oplus sl(2))$ describing, in the crystal basis [24], the codons. The numerically computed stationary solution of the master equation for the distribution of the 64 codons is nicely fitted by a function of the form Eq. (2). The model depends on the choice of the form of the fitness function and on the values of the arbitrary parameters appearing in the mutation matrix, but unrealistic choice of these values destroys the goodness of the fit. The remark of Ref. [23] on the *mystery of the straight line* in the codon position-specific frequencies as function of the total exonic GC content, raises, on the light of the previously discussed form of the rank ordered distribution, an even more intriguing question: which is the mechanism which ensures that the sum of the frequencies $f(n)$, given by Eq. (2) over a set of codons, whose rank distribution generally depends on the biological species, is a linear function in y_{GC}? From this property, we say that there is a *conspiracy* between the different codons. These effects are more evident when we sum the frequencies of the codons with fixed third nucleotide, which implies constraints on the amino-acids distribution. We have also discussed in detail the structure of the Shannon entropy. It is commonly stated in the literature that the noncoding part of DNA exhibits more correlation than the coding part, which is in contradiction with what one would naively expect as the coding part is more subjected to functional constraints. The results of our work points out the existence of strong correlations in the exonic part, which very likely witness the existence of functional bias, worthwhile of further analysis, and which should be better interpreted and understood.

It has been known for some time that the plots of the first/second codon position GC content vs. the third codon position GC content (or vs. the total GC content) show straight line correlations [5,11,9,33]. The straight line behaviour of the codon position-specific frequencies remarked in Ref. [23] (see also Tables 1 and 2) of course implies these correlations, but is stronger. In Ref. [9], a four-parameter model is developed to explain the 64 codons and the 20 amino-acids usage frequencies in function of the GC content. Considering the sums of the eight codons with the same GC content in position (which correspond to the columns of our codon matrix (26)), a satisfactory agreement between experimental values and theoretical curves is found. The authors conclude that the GC content primarily drives the codon usage rather than the inverse. Our results, concerning the GC dependence of the codon usage, although settled on different grounds, go in the same direction.

As a last remark, let us make some comments on genes. Although the GC content varies much less inside a genome than across genomes for bacteria, it is instructive to see whether the straight line effect also occurs at the gene level. In the case of *E. coli K12* for example, for which the mean GC content is 51.8%, most of the genes correspond to values of GC close to the mean value, but the tails of the P^i_X distributions ($X = \mathrm{C}, \mathrm{G}, \mathrm{U}, \mathrm{A}$ and $i = 1, 2, 3$) show a trend towards straight lines. Surely, a more detailed study deserves attention. Analogous trend for *Homo sapiens* has been observed in Ref. [33].

Finally, we have conjectured the existence of a quasi-symmetry with respect to the principal and secondary diagonal of the codon matrix, written in the form suggested by the crystal basis model of the genetic code [24]. Clearly, these discrete symmetries can be formulated without any reference to the crystal basis, but they appear naturally in this mathematical modellisation. So, as conclusive remark, let us point out that crystal basis model seems to provide the *kinematical variables* useful to describe some properties of the genetic code. As it is well known, the use of appropriate variables in mathematics, in physics, and, very likely, also in biology is an essential step for facing effectively a problem.

Acknowledgements

L. F. would like to thank the INFN of Naples and A. S. would like to thank Université de Savoie both for partial financial supports.

References

[1] Y. Nakamura, T. Gojobori, T. Ikemura, Codon usage tabulated from international DNA sequence databases: status for the year 2000, Nucleic Acids Res. 28 (1) (2000) 292 ⟨http://www.kazusa.or.jp/codon/⟩.
[2] L. Frappat, C. Minichini, A. Sciarrino, P. Sorba, Universality and Shannon entropy of codon usage, Phys. Rev. E 68 (2003) 061910.
[3] M. Kimura, Evolutionary rate at the molecular level, Nature 217 (1968) 624–626;
M. Kimura (Ed.), The Neutral Theory of Molecular Evolution, Cambridge University Press, Cambridge, UK, 1983.
[4] J.L. King, T.H. Jukes, Non-Darwinian evolution, Science 164 (1969) 788–798.
[5] A. Muto, S. Osawa, The guanine and cytosine content of bacterial evolution, Proc. Natl. Acad. Sci. USA 84 (1987) 166–169.
[6] D.J. Lynn, G.A.C. Singer, D.A. Hickey, Synonymous codon usage is subject to selection in thermophilic bacteria, Nucleic Acids Res. 30 (19) (2002) 4272–4277.
[7] S.L. Chen, W. Lee, A.K. Hottes, L. Shapiro, H.H. McAdams, Codon usage between genomes is constrained by genome-wide mutational processes, Proc. Natl. Acad. Sci. USA 101 (10) (2004) 3480–3485.
[8] P.M. Sharp, M. Stenico, J.F. Peden, A.T. Lloyd, Codon usage: mutational bias, translational selection, or both, Biochem. Soc. Trans. 21 (4) (1993) 835–841.
[9] R.D. Knight, S.J. Freeland, L.F. Landweber, A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes, Genome Biol. 2 (2001) ⟨http://genomebiology.com/2001/2/4/research/00101⟩.
[10] N. Sueoka, On the genetic basis of variation and heterogeneity of DNA base composition, Proc. Natl. Acad. Sci. USA 48 (1962) 582–592.
[11] N. Sueoka, Directional mutation pressure and neutral molecular evolution, Proc. Natl. Acad. Sci. USA 85 (8) (1988) 2653–2657.
[12] M. Gouy, C. Gautier, Codon usage in bacteria: correlation with gene expressivity, Nucleic Acids Res. 10 (22) (1982) 7055–7074.
[13] P.M. Sharp, W.H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms, J. Mol. Evol. 24 (1–2) (1986) 28–38.
[14] M. Bulmer, The selection–mutation–drift theory of synonymous codon usage, Genetics 129 (1991) 897–907.
[15] T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms, Mol. Biol. Evol. 2 (1985) 13–34.
[16] P.M. Sharp, W.H. Li, Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons, Nucleic Acids Res. 14 (19) (1986) 7737–7749.
[17] A. Eyre-Walker, Synonymous codon bias is related to selection for translational accuracy?, Mol. Biol. Evol. 13 (1996) 864–872.
[18] C.E. Shannon, A mathematical theory of communication, Bell System Tech. 27 (1948) 379–423 and 623–656.
[19] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, Cambridge, UK, 1992.
[20] C. Martindale, A.K. Konopka, Oligonucleotide frequencies in DNA follow a Yule distribution, Computers Chem. 20 (1996) 35–38.
[21] J.R. Lobry, Influence of genomic G + C content on average amino-acid composition of proteins from 59 bacterial species, Gene 205 (1997) 309–316.
[22] A. Gorban, T. Popova, A. Zinovyev, Codon usage trajectories and 7-cluster structure of 143 complete bacterial genomic sequences, Physica A 353 (2005) 365–387.
[23] A. Gorban, A. Zinovyev, The mystery of two straight lines in bacterial genome statistics, q-bio.GN/0412015.
[24] L. Frappat, A. Sciarrino, P. Sorba, A crystal base for the genetic code, Phys. Lett. A 250 (1998) 214–221.
[25] Y.B. Rumer, Codon systematization in the genetic code, Dokl. Akad. Nauk SSSR 167 (1966) 1393–1394.
[26] V.I. Shcherbak, Rumer's rule and transformation in the context of the co-operative symmetry of the genetic code, J. Theor. Biol. 139 (1989) 271–276.
[27] B.G. Konopel'chenko, Y.B. Rumer, Classification of codons in the genetic code, Dokl. Akad. Nauk SSSR 223 (1975) 471–474.
[28] N. Sueoka, Intrastrand parity rules of DNA base composition and usage biases of synonymous codons, J. Mol. Evol. 40 (3) (1995) 318–325.
[29] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria, Mol. Biol. Evol. 13 (5) (1996) 660–665.
[30] J. Mrazek, S. Karlin, Strand compositional asymmetry in bacterial and large viral genomes, Proc. Natl. Acad. Sci. USA 95 (1998) 3720–3725.
[31] M. Kowalczuk, et al., DNA asymmetry and the replicational mutational pressure, J. Appl. Genet. 42 (4) (2001) 553–577.
[32] C. Minichini, A. Sciarrino, Mutation model for nucleotide sequences based on crystal basis, q-bio.BM/0506010, Biosystems, to appear, 2006.
[33] D.R. Forsdyke, Regions of relative GC% uniformity are recombinational isolators, J. Biol. Systems 12 (3) (2004) 261–271.