ELSEVIER

Gene 238 (1999) 3–14

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

The correlation of protein hydropathy with the base composition of coding sequences

Giuseppe D'Onofrio [a], Kamel Jabbari [b], Hector Musto [a,b,c], Giorgio Bernardi [a,b,*]

[a] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, 80120 Napoli, Italy*
[b] *Laboratoire de Génétique Moleculaire, Institut Jacques Monod, 2 Place Jussieu, Paris 75005, France*
[c] *Laboratorio de Organización y Evolución del Genoma, Sección Bioquímica, Facultad de Ciencias, Iguá 4225, Montevideo 11400, Uruguay*

Accepted 15 June 1999; Received by T. Gojobori

Abstract

The "universal correlation" (D'Onofrio, G., Bernardi, G., 1992. A universal compositional correlation amomg codon positions. Gene 110, 81–88.) that holds between $\langle GC_3 \rangle$ and $\langle GC_1 \rangle$ or $\langle GC_2 \rangle$ ($\langle GC \rangle$ values are the average values of the coding sequences of each genome analyzed) at both the inter- and intra-genomic level, was re-analyzed on a vastly larger dataset. The results showed a slight, but significant, difference in the $\langle GC_3 \rangle$ vs. $\langle GC_1 \rangle$ correlations exhibited by prokaryotes and eukaryotes. This finding prompted an analysis of the correlation between $\langle GC_3 \rangle$ and the amino acid frequencies in the encoded proteins, which has shown that positive correlations exist between $\langle GC_3 \rangle$ values of coding sequences and the hydropathy of the corresponding proteins. These correlations are due to the fact that hydrophobic and amphypathic amino acids increase, whereas hydrophilic amino acids decrease with increasing $\langle GC_3 \rangle$ values. Hydropathy values of prokaryotic proteins are systematically higher than those of eukaryotes, but the slopes of the regression lines are identical. The lower hydrophobicity of eukaryotic proteins is due to differences in the amino acid composition. In particular, the twofold higher cysteine (and disulfide bond) level of eukaryotic proteins compared to prokaryotic proteins most probably compensates for their lower hydrophobicity. This supports the viewpoint that hydrophobicity plays a structural and functional role as far as protein stability is concerned.

Keywords: Genetic code; Genomes; Isochores

1. Introduction

In vertebrate genomes, linear relationships were found between the levels of GC (the molar fraction of guanine + cytosine) or GC_3 (the GC levels of third codon positions) of the coding sequences and the GC levels of the isochores embedding them (Bernardi et al., 1985). Moreover, a correlation was reported between GC_3 and GC of coding sequences, which is the same for genes from a number of genomes ranging from bacterial to human (Bernardi and Bernardi, 1985). This was the first suggestion for a general linear relationship between GC_3 and GC_{1+2} (the GC levels of first + second codon positions). In addition, points from different compositional compartments (isochores) of compositionally heterogeneous genomes, such as the genomes of warm-blooded vertebrates, fall on the line of the intergenomic correlations of homogeneous genomes, like bacterial genomes, showing that the same correlation exists not only intergenomically, but also intragenomically.

Further work (Bernardi and Bernardi, 1986) showed that: (1) $\langle GC_1 \rangle$, $\langle GC_2 \rangle$ and $\langle GC_3 \rangle$ values ($\langle GC \rangle$ are values pooled from individual prokaryotic and eukaryotic genomes or genome compartments) are positively correlated with the GC levels of the corresponding genomes, a result also reported by Muto and Osawa (1987) for a small sample of bacterial genomes; (2) the slopes of the compositional correlations between individual codon positions and coding sequences were very close for all classes of organisms; and (3) the frequencies of amino acids change with increasing GC of coding sequences, a point originally made by Sueoka (1961) for bacteria and also reported by Jukes and Bhushan (1986) for bacteria and mitochondria.

Further investigations showed that the same correlation holds between $\langle GC_3 \rangle$ and $\langle GC_{1+2} \rangle$ for human

* Corresponding author. Tel.: +39-081-7641360; fax: +39-081-7641255.
E-mail address: bernardi@alpha.szn.it (G. Bernardi)

0378-1119/99/$ – see front matter
PII: S0378-1119(99)00257-7

genes (Aïssani et al., 1991; D'Onofrio et al., 1991) and for genes from cold-blooded vertebrates, lower eukaryotes, viruses and bacteria (Bernardi and Bernardi, 1991). Finally, investigations by D'Onofrio and Bernardi (1992) led to the definition of a universal correlation among codon positions both inter- and intragenomically.

The present paper has re-analysed the universal correlation on a vastly larger sample of coding sequences and revealed that, in the high GC range of the $\langle GC_3 \rangle$ vs. $\langle GC_1 \rangle$ correlation, there are differences between prokaryotes and eukaryotes. In order to understand the meaning of such differences, we have explored (1) the correlations of $\langle GC_3 \rangle$ versus the frequencies of individual amino acids and (2) the correlation that exists between the base composition of the coding sequences and the hydropathy of the corresponding proteins (Gu et al., 1998; D'Onofrio et al., 1999; these results were presented in June 1998 at the Meeting 'Molecular Strategies in Biological Evolution' of the New York Academy of Sciences). We reach the conclusion that the correlations can be understood in terms of natural selection playing a role in the process of amino acid substitutions. This is in sharp contrast with the neutralist interpretation of a similar analysis reported by other authors (Gu et al., 1998).

2. Materials and methods

The present analysis included coding sequences from 127 prokaryotic and 70 eukaryotic genomes. Data concerning the organisms and their coding sequences (CDS), as well as $\langle GC_1 \rangle$ $\langle GC_2 \rangle$ and $\langle GC_3 \rangle$, are available upon request. A total of 110 000 CDS (about 55 000 each from prokaryotes and eukaryotes) were investigated. CDS of prokaryotes and lower eukaryotes were collected from the Codon Usage Database, release 100 (Nakamura et al., 1997; the database is available via the Internet at http://www.dna.affrc.go.jp/~nakamura/CUTG.html). Only organisms represented by more than 50 CDS were used in the present analysis. Completely sequenced genomes, comprising 16 prokaryotes and *S. cerevisiae*, were also analyzed. Coding sequences of higher eukaryotes, i.e. vertebrates and plants, were extracted from HOVERGEN (Duret et al., 1994) and from GenBank release 105, respectively. CDS were retrieved by ACNUC (Gouy et al., 1985), and the determination of both base compositions and the corresponding amino acid frequencies was carried out using the program ANALSEQ (Gautier and Jacobzone, 1989).

Coding sequences of the compositionally heterogenous genomes, warm-blooded vertebrates and monocots were partitioned into three groups according to GC_3 level: low (L): 0–47%, midrange (M): 47–74%, and high (H): 74–100%. While these GC_3 boundaries correspond to those of genes as distributed in the isochore families of the human genome (Zoubak et al., 1996), in the present paper they were applied to all compartmentalized genomes in order to compare compositionally homogeneous data.

The correlations of $\langle GC_3 \rangle$ versus $\langle GC_1 \rangle$ or $\langle GC_2 \rangle$ were investigated using orthogonal regressions (Jolicoeur, 1990), and a linear regression was used in the plot $\langle GC_3 \rangle$ versus amino acid frequencies or hydropathy of proteins encoded by genes located in each genome or genome compartment.

Hydropathic values of proteins were calculated according to Kyte and Doolittle's hydropathy scale (Kyte and Doolittle, 1982).

3. Results

3.1. Compositional properties of coding sequences

The average compositional properties of genes pooled from prokaryotic and eukaryotic genomes are summarized in Table 1. The $\langle GC_1 \rangle$, $\langle GC_2 \rangle$ and $\langle GC_3 \rangle$ values are very close in prokaryotes and eukaryotes, but the standard deviations are higher in the former. In both prokaryotic and eukaryotic genomes, the range of $\langle GC_3 \rangle$ values was much wider (9.7–93.6% and 13.5–93.0%, respectively) than those of $\langle GC_1 \rangle$ (33.1–73.8% and 38.7–63.8%, respectively) and $\langle GC_2 \rangle$ (28.1–53.0% and 29.9–49.2%, respectively). This ranking was expected because of the different constraints operating on the three codon positions. The upper limit of $\langle GC_1 \rangle$ values of eukaryotes was remarkably lower than that of prokaryotes, showing a difference of 10%.

In non-partitioned, compositionally heterogeneous genomes (Table 1), the highest average GC values were observed for rice and maize, a feature shared by other *Gramineae* (Carels et al., 1998). Among warm-blooded vertebrates, mouse GC_3 values show the lowest standard deviation compared to other mammals and chicken. The latter being characterized by a narrower compositional distribution compared to the former, in agreement with the differences found between the 'general pattern' of mammals and the 'special pattern' of murids (Salinas et al., 1986; Mouchiroud et al., 1988; Mouchiroud and Bernardi, 1993; Sabeur et al., 1993).

3.2. Intergenomic compositional correlations

Fig. 1 shows the orthogonal regression lines of $\langle GC_3 \rangle$ vs. $\langle GC_1 \rangle$ and $\langle GC_2 \rangle$, for prokaryotes and eukaryotes. High correlation coefficients were found in $\langle GC_3 \rangle$ vs. $\langle GC_2 \rangle$ plots for both prokaryotes and eukaryotes. The slopes and intercepts of the orthogonal regressions were slightly higher in eukaryotes compared

Table 1
Compositional properties of coding sequences[a]

	CDS	$\langle GC_1 \rangle$	SD	$\langle GC_2 \rangle$	SD	$\langle GC_3 \rangle$	SD
Prokaryotes	56.161	56.60	8.93	40.58	5.40	54.83	23.72
Eukaryotes	54.508	52.48	4.50	40.57	3.18	48.23	16.20
Chicken	956	55.64	6.60	41.81	7.30	64.05	17.06
Man	6.681	55.92	7.14	42.59	7.38	61.07	15.85
Mouse	3.902	55.48	6.69	42.71	7.15	60.84	11.81
Calf	852	55.51	6.92	41.56	7.13	64.02	14.87
Maize	608	58.54	6.78	43.97	7.51	68.68	19.40
Rice	570	57.39	6.00	44.39	7.51	69.44	19.94

[a] CDS denotes the number of coding sequences; $\langle GC_1 \rangle$, $\langle GC_2 \rangle$ and $\langle GC_3 \rangle$ denote the average GC levels of the first, second and third codon position, respectively; SD is the standard deviation.

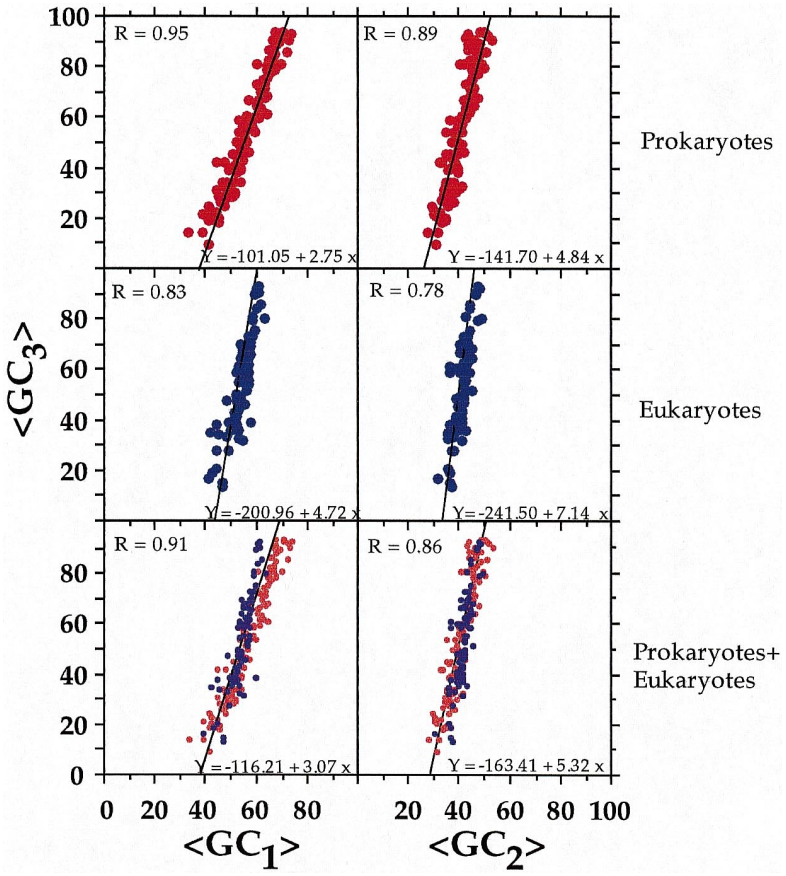

Fig. 1. Intergenomic compositional correlations. $\langle GC_3 \rangle$ values of genes averaged by genome or genome compartments (in the case of heterogeneous genomes; see Section 2) are plotted against the corresponding $\langle GC_1 \rangle$ and $\langle GC_2 \rangle$ values. Plots for prokaryotes (red dots), eukaryotes (blue dots) and prokaryotes + eukaryotes are shown, along with the equations of orthogonal regression lines and correlation coefficients.

to prokaryotes, but a standard test (Jolicoeur, 1990) showed that the differences were not significant. The correlations between $\langle GC_3 \rangle$ and $\langle GC_1 \rangle$ also showed high coefficients for all prokaryotes and eukaryotes, but the slopes were different for the two groups. Table 2 summarizes these results. A more detailed analysis of the correlation $\langle GC_3 \rangle$ vs. $\langle GC_1 \rangle$ showed that up to a $\langle GC_3 \rangle$ level of 65%, the two slopes were not significantly different. At higher $\langle GC_3 \rangle$ levels, the distributions of points from prokaryotes and eukaryotes diverged (Fig. 1), and the difference between the two regression lines became statistically significant ($p < 0.005$).

Fig. 1 also shows the correlation obtained when prokaryotes and eukaryotes are pooled together. Clearly, on a first approximation, a universal correlation still exists between $\langle GC_1 \rangle$ and both $\langle GC_2 \rangle$ and $\langle GC_3 \rangle$. In fact, the equation of the regression line of $\langle GC_3 \rangle$ vs. $\langle GC_{1+2} \rangle$ is not significantly different from that previously published using a small number of genes (D'Onofrio and Bernardi, 1992).

Table 2
Inter- and intra-genomic compositional correlations[a]

	$\langle GC_3 \rangle$ vs. S	$\langle GC_1 \rangle$ α	$\langle GC_3 \rangle$ vs. S	$\langle GC_2 \rangle$ α
Prokaryotes	2.75	70.05	4.84	78.33
Eukaryotes	4.72	78.90	7.14	82.33
Chicken	4.78	78.19	5.63	79.94
Man	3.91	75.66	6.00	80.55
Mouse	3.26	73.00	5.24	79.61
Calf	4.32	76.96	6.91	81.76
Maize	6.68	81.49	6.84	81.68
Rice	6.45	81.05	6.25	80.91

[a] S and α are the slopes and the angles of the orthogonal regression lines, respectively.

3.3. Intragenomic compositional correlations

In the case of compositionally heterogenous genomes for which large samples are available (human, calf, mouse, chicken among warm-blooded vertebrates and maize and rice among *Gramineae*), the intragenomic correlations were investigated and are summarized in Table 2. Some differences are seen that may be due to the differences in the gene samples used.

3.4. Amino acid frequencies

The differences in slopes observed in the $\langle GC_3 \rangle$ vs. $\langle GC_1 \rangle$ correlations of prokaryotes and eukaryotes suggested the existence of differences in amino acid frequencies between the two sets of organisms. This prompted us to test which amino acids showed statistically different frequencies. The mean values, standard errors and statistical significance of differences for the whole $\langle GC_3 \rangle$ range, for the $<65\%$ range and for the $>65\%$ range, are presented in Table 3 (see also Fig. 2).

When considering the whole $\langle GC_3 \rangle$ range, seven amino acids, namely Arg, Leu, Ala, Gly, Val, Asp and Ile, were significantly higher in prokaryotes, whereas six amino acids, Ser, Pro, Lys, His, Cys and Met, were lower.

In the $<65\%$ range, seven amino acids were found to be very significantly different, with p values higher than 5‰, namely Leu and Ile, which were higher in prokaryotes, and Ser, Pro, His, Tyr, Cys which were lower in prokaryotes. Among these amino acids, Leu, His and Pro are the only amino acids having C in first codon positions. Among the amino acids having G in first codon position, Ala and Val were at the borderline level ($p<5\%$).

In the $>65\%$ range, there were 13 statistically different amino acids, Arg, Leu, Ala, Gly, Val and Asp, which were higher in prokaryotes and Ser, Lys, Asn, Gln, Tyr, Cys and Phe, which were lower in prokaryotes. Among them, only five, Leu, Val, Gln, His, and Asp, have G or C in the first codon positions.

3.5. Correlations between GC_3 and amino acid frequencies

In order to understand better the differences in amino acid frequencies between prokaryotes and eukaryotes, the correlations between $\langle GC_3 \rangle$ and amino acid frequencies were analysed. The slopes (S), correlation coefficients (R), and p values of the regression lines of each amino acid are listed in Table 4 (see also Fig. 3). As in D'Onofrio et al. (1991), amino acids were divided into an AT, a GC and an intermediate class, according to the occurrence of only A/T, only G/C, or A/T and G/C in their first and second codon positions (see Fig. 4).

Even if the correlation coefficients were generally higher in prokaryotes (mainly in the GC class, where the first and second codon positions only comprise G or C; see Table 4), highly significant p values were found in both data sets, and amino acids showing no correlation in the prokaryotes also showed no correlation in eukaryotes. The exceptions were the amino acids Leu and Ser, which showed significant correlations in prokaryotes, but not in eukaryotes, and Cys, Glu and Met, which showed the opposite. All these amino acids, except Met, belong to the intermediate class, where both G/C and A/T are present in the first and second codon positions; therefore, the correlations of their frequencies with $\langle GC_3 \rangle$ levels were not expected.

Although the correlations between amino acids and $\langle GC_3 \rangle$ levels were similar for prokaryotes and eukaryotes and high positive or negative correlation coefficients were found for amino acids that belong to the GC or AT class of codons, all amino acids, except for Pro, showed higher slopes in prokaryotes. Nine examples of different correlations between amino acid frequencies and GC_3 levels in prokaryotes and eukaryotes are reported in Fig. 3, which shows that: (1) the regression lines of Arg in prokaryotes and eukaryotes intersect each other at 45% GC_3; (2) Lys shows a similar behaviour with an opposite trend; (3) other amino acids show positive convergent (proline) or divergent trends (alanine, valine and leucine), or negative convergent (isoleucine) or divergent trends (serine); (4) cysteine showed an essentially parallel behavior, but with eukaryotic values twice as large as prokaryotic values.

Completely sequenced genomes were also analyzed. Remarkably, all the trends already observed were not only confirmed but even enhanced, either at the nucleotide level or at the amino acid level. Furthermore, the completely sequenced genome of *S. cerevisiae* showed a lower hydropathy and a higher content of cysteine compared to the mean values of the other 16 completely sequenced genomes of prokaryotes.

3.6. Quartets and duets

Fig. 4 shows the Grantham (1980) representation of the genetic code. The original representation was slightly

Table 3
Frequencies of amino acids encoded by prokaryotic and eukaryotic genes[a]

$0 < GC_3 < 100\%$						$GC_3 < 65\%$						$GC_3 > 65\%$					
	Prokaryotes		Eukaryotes				Prokaryotes		Eukaryotes				Prokaryotes		Eukaryotes		
	Mean	SE	Mean	SE	p[b]		Mean	SE	Mean	SE	p[b]		Mean	SE	Mean	SE	p[b]
Ala	9.29	2.34	7.73	1.61	****	Ala	7.85	1.51	7.25	1.27	*	Ala	11.67	1.26	9.15	1.71	****
Arg	5.49	1.58	5.05	0.79	**	Arg	4.52	1.03	4.88	0.75	*	Arg	7.08	0.86	5.55	0.71	****
Asn	4.23	1.60	4.43	1.09	ns	Asn	5.08	1.38	4.65	1.11	ns	Asn	2.83	0.69	3.79	0.72	****
Asp	5.54	0.67	5.26	0.53	****	Asp	5.41	0.45	5.3	0.58	ns	Asp	5.76	0.89	5.15	0.34	****
Cys	0.94	0.36	1.89	0.53	****	Cys	0.92	0.36	1.86	0.55	****	Cys	0.98	0.37	1.97	0.44	****
Gln	3.78	0.89	4.18	1.74	ns	Gln	3.98	0.95	4.27	1.99	ns	Gln	3.46	0.69	3.92	0.52	***
Glu	6.32	1.08	6.44	0.92	ns	Glu	6.45	0.87	6.61	0.91	ns	Glu	6.11	1.33	5.95	0.79	ns
Gly	7.56	1.16	7.17	1.11	*	Gly	6.92	0.91	6.93	0.96	ns	Gly	8.63	0.63	7.89	1.24	*
His	2.02	0.38	2.24	0.25	****	His	1.90	0.38	2.21	0.25	****	His	2.22	0.28	2.30	1.26	ns
Ile	6.09	1.65	5.23	0.94	****	Ile	7.01	1.30	5.49	0.85	****	Ile	4.57	0.82	4.44	0.75	ns
Leu	9.68	0.97	8.74	0.77	****	Leu	9.57	0.97	8.74	0.71	****	Leu	9.85	0.95	8.74	0.93	***
Lys	5.48	2.25	6.26	1.22	***	Lys	6.67	1.97	6.53	1.20	ns	Lys	3.52	0.89	5.44	0.84	****
Met	2.25	0.41	2.36	0.31	*	Met	2.27	0.39	2.35	0.30	ns	Met	2.23	0.44	2.39	0.37	ns
Phe	3.83	0.65	3.92	0.43	ns	Phe	4.15	0.57	4.01	0.40	ns	Phe	3.32	0.43	3.65	0.39	**
Pro	4.34	0.96	5.22	0.83	***	Pro	3.79	0.69	5.13	0.89	****	Pro	5.23	0.61	5.48	0.54	ns
Ser	6.19	0.93	7.50	0.80	***	Ser	6.55	0.79	7.59	0.79	****	Ser	5.60	0.85	7.25	0.80	****
Thr	5.60	0.72	5.61	0.54	ns	Thr	5.67	0.75	5.56	0.44	ns	Thr	5.49	0.67	5.79	0.73	ns
Trp	1.17	0.33	1.15	0.22	ns	Trp	1.06	0.20	1.1	0.21	ns	Trp	1.35	0.25	1.30	0.21	ns
Tyr	3.10	0.81	3.06	0.41	ns	Tyr	3.50	0.71	3.09	0.42	****	Tyr	2.43	0.42	2.95	0.38	****
Val	7.12	0.86	6.57	0.61	****	Val	6.76	0.74	6.45	0.60	*	Val	7.72	0.69	6.91	0.52	****

[a] The statistical test used was the unpaired t-test.
[b] $*p<5\%$; $**p<1\%$; $***p<0.5\%$; $****p<0.1\%$.

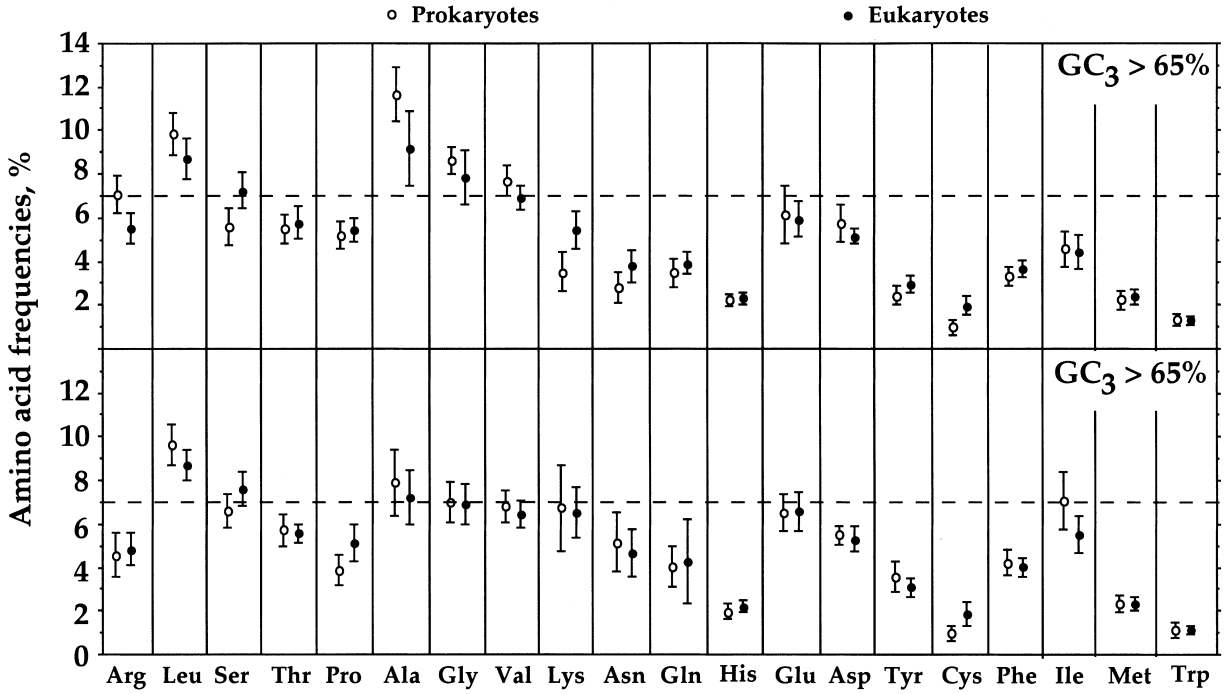

Fig. 2. Mean and standard deviation of amino acid frequencies in prokaryotes (open symbols) and eukaryotes (closed symbols).

modified such that: (1) codons rather than anticodons are displayed; (2) a distinction is made among nucleotides in third codon positions of quartets, duets and odd number codons; and (3) hydropathy values for amino acids (Kyte and Doolittle, 1982) are shown. The most hydrophobic amino acids are in red, the most hydrophilic in blue, and the intermediates in green boxes. We consider this modified Grantham representation of the

Table 4
Correlations between amino acid frequencies and $\langle GC_3 \rangle$ levels

	Prokaryotes				Eukaryotes			
	aa	*S*	*R*	*p*	aa	*S*	*R*	*p*
AT class	Phe	−0.02	0.67	0.0001	Phe	−0.01	0.44	0.0001
	Ile	−0.06	0.87	0.0001	Ile	−0.04	0.82	0.0001
	Met	0.00	0.00	ns	Met	0.004	0.23	0.032
	Tyr	−0.02	0.72	0.0001	Tyr	−0.01	0.31	0.004
	Asn	−0.06	0.86	0.0001	Asn	−0.04	0.61	0.0001
	Lys	−0.08	0.83	0.0001	Lys	−0.04	0.54	0.0001
Intermediate class	Leu	0.01	0.25	0.005	Leu	−0.002	0.05	ns
	Val	0.02	0.67	0.0001	Val	0.01	0.41	0.0001
	Trp	0.01	0.54	0.0001	Trp	0.01	0.40	0.0001
	Thr	−0.01	0.15	ns	Thr	−0.004	0.07	ns
	Ser	−0.02	0.61	0.0001	Ser	−0.01	0.15	ns
	Cys	0.002	0.12	ns	Cys	0.01	0.32	0.003
	Gln	−0.01	0.19	ns	Gln	−0.01	0.09	ns
	His	0.01	0.53	0.0001	His	0.004	0.30	0.01
	Asp	0.01	0.22	0.02	Asp	−0.004	0.13	ns
	Glu	−0.01	0.12	ns	Glu	−0.01	0.27	0.01
GC class	Arg	0.06	0.90	0.0001	Arg	0.02	0.56	0.0001
	Ala	0.09	0.89	0.0001	Ala	0.06	0.63	0.0001
	Gly	0.04	0.85	0.0001	Gly	0.03	0.50	0.0001
	Pro	0.04	0.86	0.0001	Pro	0.02	0.38	0.0003

genetic code to be the best available, especially for the purpose of the present paper.

Correlations of $\langle GC_3 \rangle$ versus quartets and duets were found to be very similar in prokaryotes and eukaryotes (Fig. 5). Correlation coefficients were very high when all quartets and duets (including those from sextets), were taken into account, and slightly lower when the latter were excluded, this finding being explained by the very good negative correlation within the quartets and duets of sextets coding for Arg and Leu.

3.7. Correlations between GC_3 and hydropathy

The frequencies of hydrophilic, hydrophobic and amphipatic amino acids (as indicated by red, green and blue boxes, respectively, in Fig. 4) are correlated very significantly with $\langle GC_3 \rangle$ ($p < 10^{-4}$), in both prokaryotes and eukaryotes (Fig. 6). More precisely, hydrophilic amino acids are correlated negatively, whereas hydrophobic and amphipathic amino acids are correlated positively. All correlations showed a higher correlation coefficient in prokaryotes compared to eukaryotes. In eukaryotes, the slopes were lower for hydrophobic and hydrophilic amino acids and higher for amphipathic amino acids, compared to prokaryotes.

When the $\langle GC_3 \rangle$ levels were plotted against the hydropathic mean values of proteins, using the hydrophobicity scale of Kyte and Doolittle (1982), two practically parallel regression lines (Fig. 7) were found to be statistically significant for prokaryotes and eukaryotes ($p < 0.001$ and $p < 0.007$, respectively). A statistically significant correlation between $\langle GC_3 \rangle$ and hydropathic mean values ($p < 0.004$) was also observed using the GSE hydrophobicity scale (Engelman et al., 1986). An exhaustive analysis performed on hydrophobicity scales reviewed by Cornette et al. (1987) showed that this was also true using 30 out of 40 scales, whereas the remaining scales showed no statistically significant correlation.

The increase in hydropathy was not the only property of amino acids that changed with increasing $\langle GC_3 \rangle$ levels. Indeed, the frequencies of positively (Arg, Lys and His) and negatively (Asp and Glu) charged amino acids were also affected. Positively charged amino acids showed no correlation vs. $\langle GC_3 \rangle$, in eukaryotes, and a negative and significant correlation in prokaryotes. On the contrary, negatively charged amino acids showed no correlation vs. $\langle GC_3 \rangle$ in prokaryotes and a negative correlation in eukaryotes. This different behaviour of charged amino acids is unclear, but it has been speculated that this is due to a different intracellular pH, that at least in mammals is slightly acid (Karlin et al., 1992)

4. Discussion

4.1. Correlations among codon positions

The present paper revisited the 'universal correlation' (D'Onofrio and Bernardi, 1992) existing between $\langle GC_1 \rangle$ or $\langle GC_2 \rangle$ versus $\langle GC_3 \rangle$ values, as averaged over all coding sequences from individual prokaryotes and eukaryotes. Increasing the number of CDS analyzed from 2300 to 110 000 (about 55 000 each for prokaryotes and eukaryotes) did not lead to any significant change

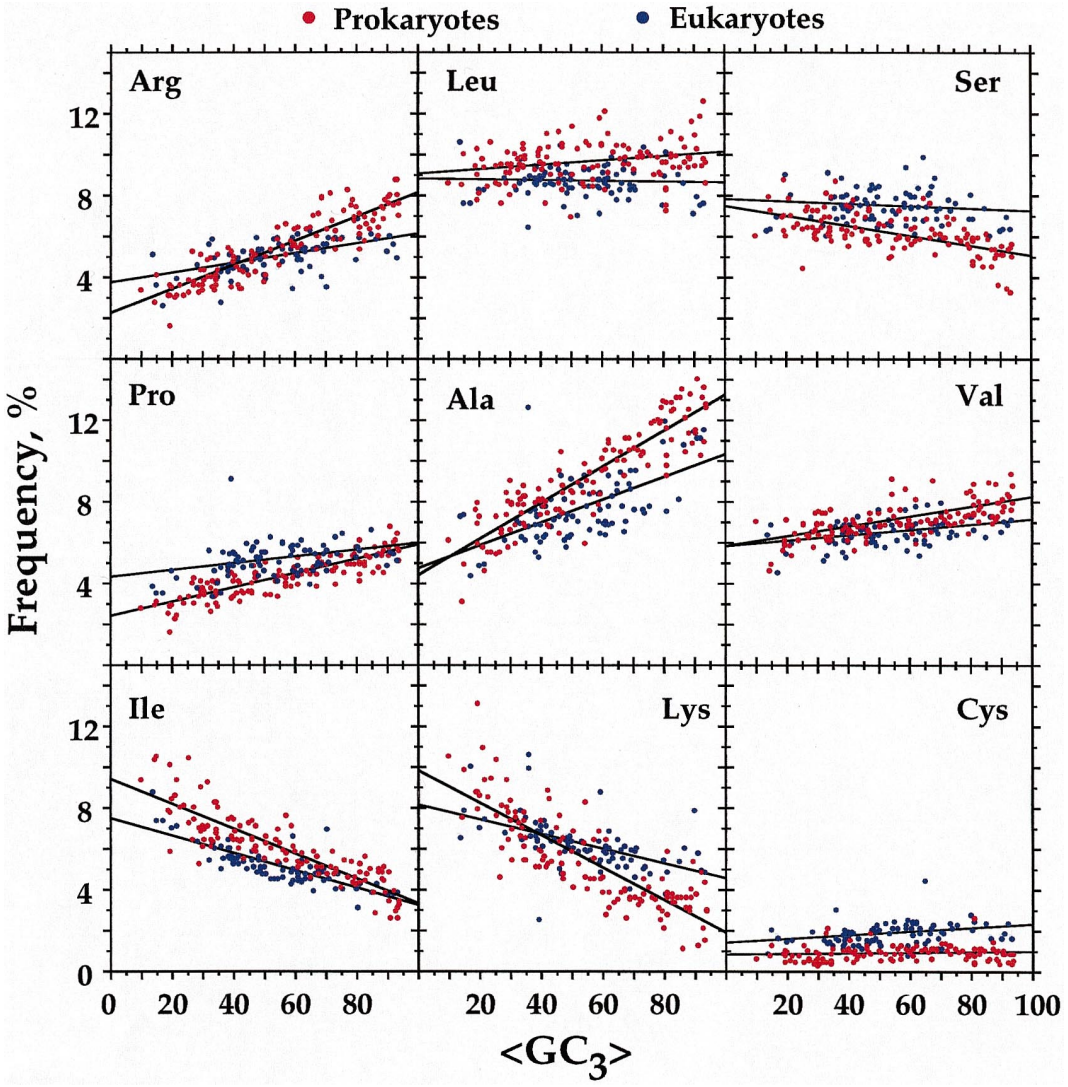

Fig. 3. Frequencies of nine amino acids from prokaryotes and eukaryotes plotted against $\langle GC_3 \rangle$.

in the $\langle GC_3 \rangle$ vs. $\langle GC_2 \rangle$ correlation previously published. This is a very strong indication that no significant changes should be expected in the $\langle GC_3 \rangle$ vs. $\langle GC_2 \rangle$ correlation upon further increases in the data set. In contrast, the $\langle GC_3 \rangle$ vs. $\langle GC_1 \rangle$ correlation showed a small yet significant difference between prokaryotes and eukaryotes for $\langle GC_3 \rangle$ values higher than 65%.

4.2. Correlation between $\langle GC_3 \rangle$ and amino acid frequencies

Since the most obvious explanation for the difference found was a difference in the frequencies of amino acids in prokaryotes and eukaryotes, these frequencies were tested. This led us to an analysis that revealed very interesting correlations between the GC levels (especially for $\langle GC_3 \rangle$) of the CDS and the amino acid frequencies of the corresponding genes. As far as the correlation between $\langle GC_3 \rangle$ and amino acid frequency is concerned, only some amino acids did not show this correlation. These comprised all the amino acids of the intermediate class (whether they have C/G in first codon position or not) and Met, the least represented among the non-intermediary amino acids. The other amino acids showed the same trends, already described. Remarkably, these correlations between $\langle GC_3 \rangle$ and amino acid frequencies were different for prokaryotes and eukaryotes.

Indeed, even in the $\langle GC_3 \rangle$ range $< 65\%$, where both the correlation of $\langle GC_3 \rangle$ vs. $\langle GC_1 \rangle$ and that of $\langle GC_3 \rangle$ vs. $\langle GC_2 \rangle$ show no difference between eukaryotes and prokaryotes, as many as seven amino acids have significantly different frequencies. In the $> 65\%$ range, whereas the $\langle GC_3 \rangle$ vs. $\langle GC_2 \rangle$ regression line is still the same in eukaryotes and prokaryotes, the correlation $\langle GC_3 \rangle$ vs. $\langle GC_1 \rangle$ starts to diverge, and the number of amino acids showing significant differences rise up to 13. It should be mentioned that differences in amino acid frequencies between eukaryotes and prokaryotes

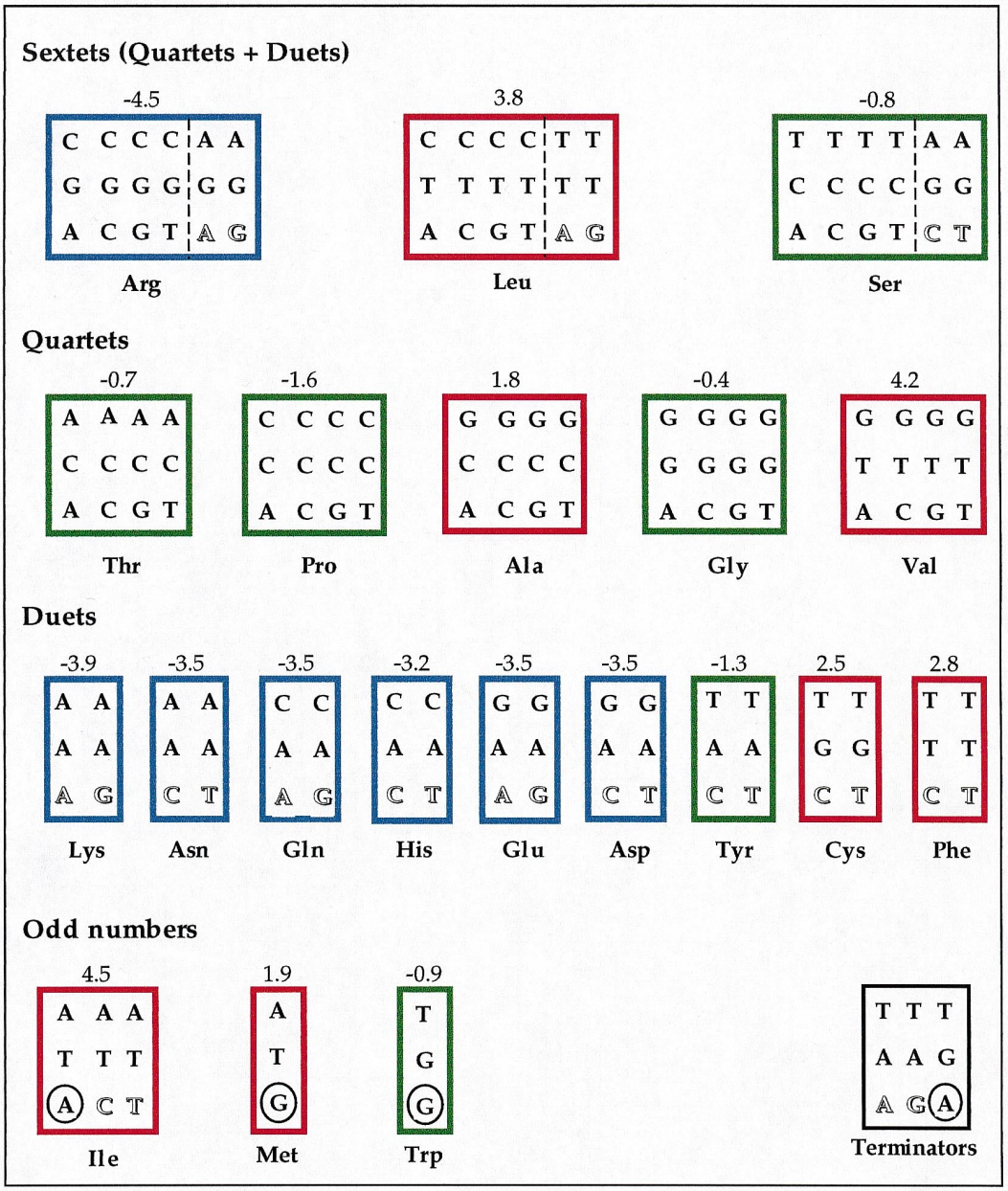

Fig. 4. Grantham (1980) representation of the genetic code, slightly modified such that (1) codons rather than anticodons are shown, (2) a distinction is made among nucleotides in third codon positions of quartets, duets and odd number codons, and (3) hydropathy values for amino acids (Kyte and Doolittle, 1982) are shown. The most hydrophobic amino acids are in red boxes, the most hydrophilic in blue boxes, and the intermediate class in green boxes.

were also reported by Karlin (1992) using a more general approach analyzing four prokaryotes and two eukaryotes.

The result that amino acids show different frequencies at comparable $\langle GC_3 \rangle$ levels indicates that the CDS base composition cannot predict amino acid frequencies, even if they are correlated (Sueoka, 1961; D'Onofrio et al., 1991; Lobry, 1977). In agreement with this conclusion, Lobry (1977) reported that among "12 amino acids whose frequencies are expected to be affected by G+C content […], 10 amino acids do not follow the predicted trend with the expected magnitude" and that the frequencies of the remaining amino acids are not expected to be influenced by base composition.

4.3. Correlation between $\langle GC_3 \rangle$ and protein hydropathy

As far as the hydropathy of the proteins is concerned, the present results show that in both eukaryotes and prokaryotes: (1) the frequencies of hydrophobic and amphipathic amino acids increase, whereas those of hydrophilic amino acids decrease with increasing

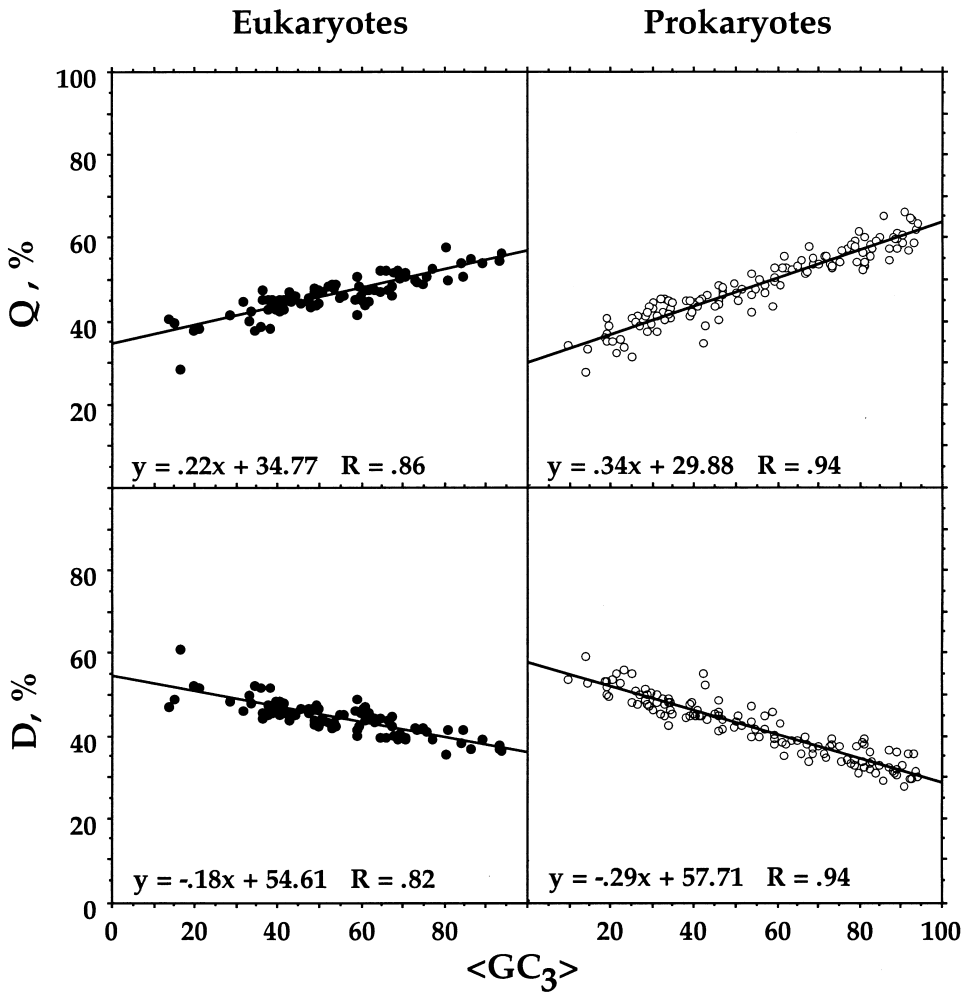

Fig. 5. Plot of frequencies of quartet (Q%) and duet (D%) codons against $\langle GC_3 \rangle$.

$\langle GC_3 \rangle$; (2) the frequency correlations of quartet and duet codons with $\langle GC_3 \rangle$ showed positive and negative correlations, respectively; and (3) the slopes of the regression line of hydropathy vs. $\langle GC_3 \rangle$ plots were very similar in eukaryotes and prokaryotes; intercepts were, however, different, with hydropathy values being systematically higher in prokaryotes.

The first finding clearly indicates that $\langle GC_3 \rangle$ changes are accompanied by amino acid changes whose functional meaning will be discussed below. The second finding is a direct consequence of the first. Indeed, with the single exception of the arginine quartet, all quartet codons correspond to hydrophobic and amphipathic amino acids, whereas seven out of 12 duets correspond to hydrophilic amino acids, two to 'intermediate' codons and three to low-frequency amino acids. Interestingly, if one sums up the hydropathic values [using the Kyte and Doolittle (1982) scale] of quartets and duets, as represented in Fig. 1, the mean values are +3.3 and −16.7, respectively, stressing the large difference between quartets and duets.

The third finding also is a consequence of the first, in that hydropathy reflects the contributions of hydrophobic, amphypathic and hydrophilic amino acids. Here, the interesting point is the parallelism of the two correlations of eukaryotes and prokaryotes and the lower values of prokaryotes. Indeed, the lower hydrophobicity values are accompanied by higher values of cysteine. The latter may be interpreted, in agreement with previous proposals (Cedano et al., 1997), to compensate, as far as stability is concerned, for the lower hydrophobicity of eukaryotic proteins. In support of this interpretation, it should be mentioned that the cysteine content is higher in extracellular proteins of eukaryotes to provide the required stability, whereas in bacteria, extracellular proteins may even be cysteine-free (Fahey et al., 1977). Along the same line, nuclear proteins are generally poor in hydrophobic amino acids and rich in charged amino acids (Nakashima and Nishikawa, 1994). An important point of Fig. 2 is that if the difference between eukaryotic and prokaryotic hydropathies is functionally meaningful, so should also be the difference in hydropathies associated with increasing $\langle GC_3 \rangle$. Indeed, the hydropathic increments in prokaryotes, Δp, and in

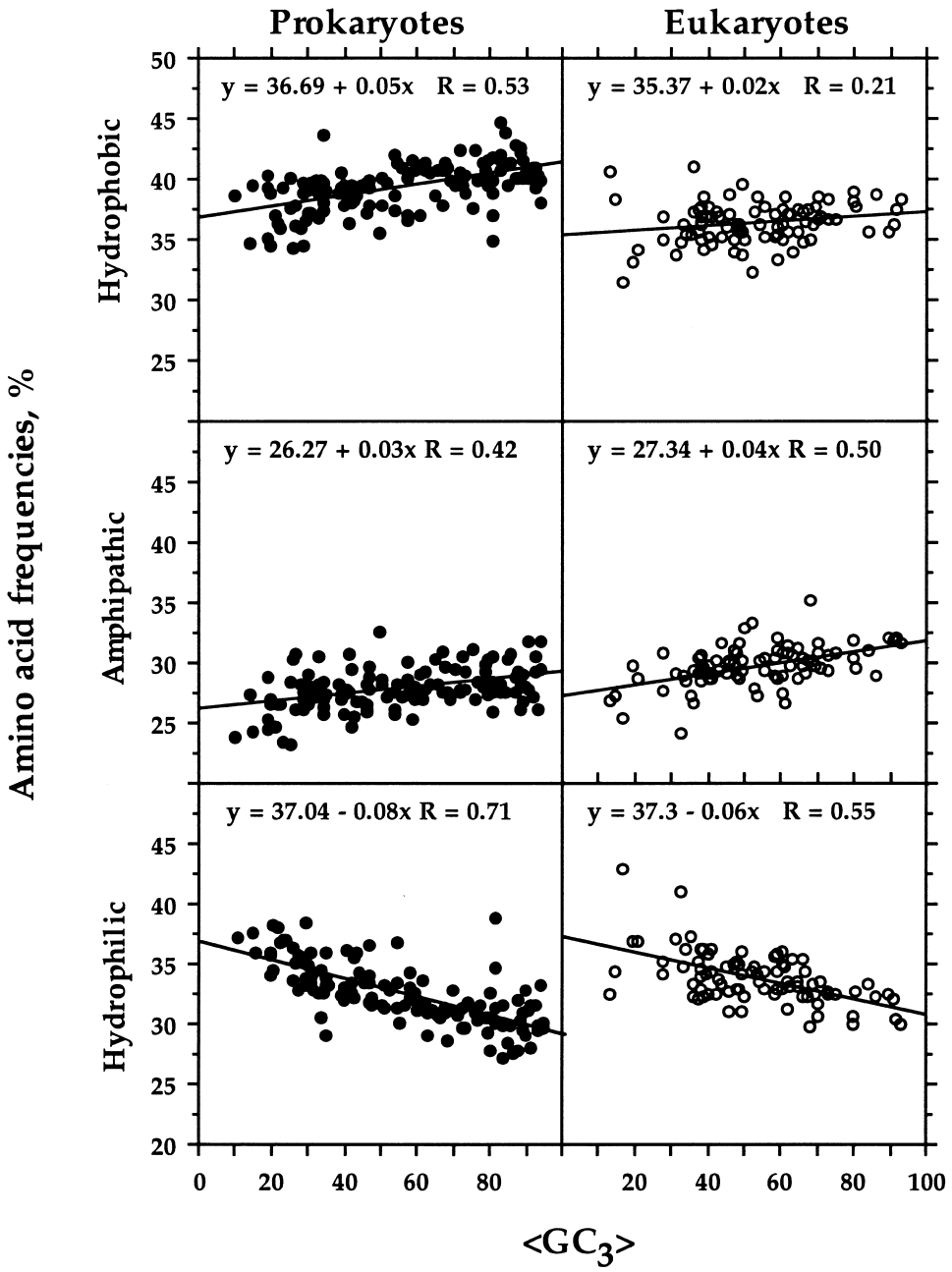

Fig. 6. Frequencies of hydrophilic, hydrophobic and amphipathic amino acids from prokaryotes (closed circles) and eukaryotes (open circles) plotted against $\langle GC_3 \rangle$.

eukaryotes Δe, associated with increasing $\langle GC_3 \rangle$ levels are equal to the difference in hydropathy between prokaryotes and eukaryotes ($\Delta p - e$, i.e. the difference between the two regression lines).

Our conclusion is, therefore, that changes in $\langle GC_3 \rangle$ are accompanied by significant structural and, possibly, functional changes in the encoded proteins. These latter changes drive the former changes. In other words, selection for stabilizing amino acids leads to an increase in GC of coding sequences, itself a factor stabilizing both DNA and RNA, as already suggested (Bernardi and Bernardi, 1986).

Interestingly, a very recent paper dealing with the same problem (Gu et al., 1998) reached an opposite conclusion. These authors have studied the dnaA protein from *E. coli* and 14 other bacterias, plus 10–14 other prokaryotic proteins, and claimed that "both strongly hydrophobic and strongly hydrophilic amino acids tend to change to ambivalent amino acids, suggesting that the majority of these amino acid substitutions are not caused by positive Darwinian selection". We will try to explain first why the findings reported by Gu et al. (1998) are different from those presented here and second why the interpretation is different.

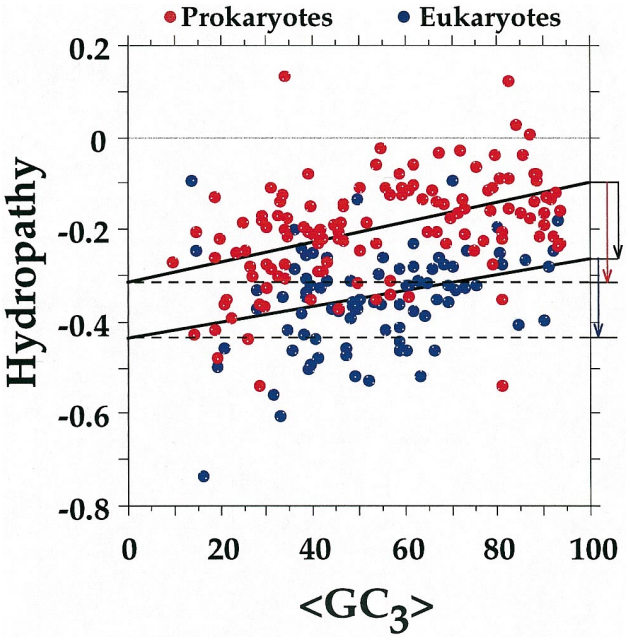

Fig. 7. Hydropathy values of proteins from prokaryotes (red circles) [$y=0.0021x$ —0.30; $R=0.44$] and eukaryotes (blue circles) [$y=0.0017$ —0.44; $R=0.27$] plotted against $\langle GC_3 \rangle$.

Gu et al. (1998) divided amino acids into three classes, according to the GC levels of first and second positions of the corresponding codons [essentially as previously done by Jukes and Bhushan (1986) and D'Onofrio et al. (1991)], and plotted the frequencies of amino acids against genomic GC and found that the GC-rich class increased, the intermediate class remained constant and the GC-poor class decreased. These results could be predicted from the plots of the frequencies of amino acids vs. GC_{1+2} of D'Onofrio et al. (1991). More interestingly, they plotted the frequencies of external (Asp, Glu, Lys, Arg, His, Asn and Glu), internal (Phe, Len, Ile, Met, Val, Tyr and Trp) and ambivalent (Ala, Pro, Gly, Ser, Thr, Cys) amino acids [the classification of amino acids was from Dickerson and Geis (1983)] against genomic GC and found an increase in ambivalent amino acids and a parallel decrease in both external and internal amino acids. Therefore, their conclusion.

The authors correctly note that the classification of external (hydrophilic), internal (hydrophobic) and ambivalent (amphypathic) amino acids that they used "is not unambiguous". Indeed, Ala and Cys [classified as ambivalent by Dickerson and Geis (1983)] are hydrophobic, and Tyr and Trp [classified as internal by Dickerson and Geis (1983)] are 'intermediate' amino acids according not only to the generally accepted classification of Kyte and Doolittle (1982), but also to other authors (Engelman et al., 1986; Rose et al., 1986; Fiser et al., 1996). The problem, in the case of Gu et al. (1998), is, however, not simply the assignment of a given amino acid to a given class, but the fact that their approach considers all amino acids belonging to a given class as being equivalent in hydropathy. This is incorrect, as is widely recognized. If the hydropathy values for each amino acid are used, following the classification of Kyte and Doolittle (1982), it is clear that only hydrophylic amino acids decrease with increasing $\langle GC \rangle$ (or $\langle GC_3 \rangle$), whereas both hydrophobic and amphipathic amino acids increase. Needless to say, this conclusion is in sharp contrast with that of Gu et al. (1998). Indeed, $\langle GC_3 \rangle$ increases in coding sequences are accompanied by increases in the hydropathy of the encoded proteins and, according to a widely accepted viewpoint, by an increase in protein stability, namely by a structurally and functionally meaningful change. The significance of the change is stressed by the need for compensating the lower hydrophobicity of eukaryotic proteins by an increase in another stabilizing factor, namely disulfide bridges.

At this point, it becomes obvious that the driving forces of the nucleotide changes are the changes of the encoded proteins, which reverses the neutralist view (Gu et al., 1998) that the mutational bias is responsible for them. Other arguments along the same line will be provided in future articles (Chiusano et al., in press; Cruveiller et al., in press).

References

Aïssani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C., Bernardi, G., 1991. The compositional properties of human genome. J. Mol. Evol. 32, 493–503.

Bernardi, G., Olofson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.

Bernardi, G., Bernardi, G., 1985. Codon usage and genome composition. J. Mol. Evol. 22, 363–365.

Bernardi, G., Bernardi, G., 1986. Compositionl constraints and genome evolution. J. Mol. Evol. 24, 1–11.

Bernardi, G., Bernardi, G., 1991. Compositional properties of nuclear genes from cold-blooded vertebrates. J. Mol. Evol. 33, 57–67.

Carels, N., Hatey, P., Jabbari, K., Bernardi, G., 1998. Compositional properties of homologous coding sequences from plants. J. Mol. Evol. 46, 45–53.

Cedano, J., Aloy, P., Perez-Pons, J.A., Querol, E., 1997. Relation between amino acid composition and cellular location of proteins. J. Mol. Biol. 266, 594–600.

Chiusano, M.L., D'Onofrio, G., Alvarez-Valin, F., Jabbari, K., Bernardi, G., publication. Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. Gene, in press.

Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L, Berzofsky, J.A., DeLisi, C., 1987. Hydrophobicity scales and computational techniques for detecting amphipatic structures in proteins. J. Mol. Biol. 195, 659–685.

Cruveiller, S., Jabbari, K., D'Onofrio, G., Bernardi, G., 1999. Different hydrophobicities of orthologous proteins from *Xenopus* and man Gene 238, 15–21.

Dickerson, R.E., Geis, I., 1983. Hemoglobins: Structure, Function, Evolution and Pathology. The Benjamin Cummings Publishing Company, Menlo Park, CA.

D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., Bernardi, G., 1991. Correlation between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J. Mol. Evol. 32, 504–510.

D'Onofrio, G., Bernardi, G., 1992. A universal correlation amomg codon positions. Gene 110, 81–88.

D'Onofrio, G., Jabbari, K., Musto, H., Alvarez-Valin, F., Cruveiller, S., Bernardi, G., 1999. Evolutionary genomics of vertebrates and its implications. In: Caporale, L.H., Arber, W. (Eds.), Molecular Strategies in Biological Evolution. Ann. NY Acad. Sci., New York, NY, pp. 1–14.

Duret, L., Mouchiroud, D., Gouy, M., 1994. HOVERGEN: Homologous Vertebrate Genes database. Nucleic Acids Res. 22, 2360–2363.

Engelman, D.M., Steitz, T.A., Goldman, A., 1986. Identifying nonpolar transbilayer helices in amino acids sequences of membrane proteins. Ann. Rev. Biophys. Biochem. 15, 321–353.

Fahey, R.C., Hunt, J.S., Windham, G.C., 1977. On the cysteine and cystine content of proteins. Differences between intracellular and extracellular proteins. J. Mol. Evol. 10, 155–160.

Fiser, A., Simon, I., Barton, G.J., 1996. Conservation of amino acids in multiple alignments: aspartic acid has unexpected conservation. Febs. Lett. 397, 225–229.

Gautier, C., Jacobzone, M., 1989. Publication interne UMR CNRS 5558 Biométrie Génétique et Biologie des Populations. Université Claude Bernard, Lyon I France. http://biom3.univlyon.fr:8080/doclogi/docanals/manuel.html

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., di Paola, G., 1985. ACNUC — a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. CABIOS 1, 167–172.

Grantham, R., 1980. Workings on the genetic code. Trends Biochem. Sci. 5, 327–333.

Gu, X., Hewett-Emmett, D., Li, W.H., 1998. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. Genetica 103, 383–391.

Jolicoeur, P., 1990. Bivariate allometry: interval estimation of slopes of the ordinary and standardized normal major axes and structural relationship. J. Theor. Biol. 144, 275–285.

Jukes, T.H., Bushan, V., 1986. Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. J. Mol. Evol. 24, 39–44.

Karlin, S., Blaisdell, B.E., Bucher, P., 1992. Quantile distributions of amino acid usage in protein classes. Prot. Eng. 5, 729–738.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying hydropathic haracter of a protein. J. Mol. Biol. 157, 105–132.

Lobry, J.R., 1977. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205, 309–316.

Mouchiroud, D., Gautier, C., Bernardi, G., 1988. The compositional distribution of coding sequences and DNA molecules in humans and murids. J. Mol. Evol. 27, 311–320.

Mouchiroud, D., Bernardi, G., 1993. Compositional properties of coding sequences and mammalian phylogeny. J. Mol. Evol. 37, 109–116.

Muto, A., Osawa, S., 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA 84, 166–169.

Nakamura, Y., Gojobori, T., Ikemura, T., 1997. Codon usage tabulated from the international DNA sequence databases. Nucleic Acids Res. 25, 244–245.

Nakashima, H., Nishikawa, K., 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol. 238, 54–61.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H., 1986. Hydrophobicity of amino acid residues in globular proteins. Science 229, 834–838.

Sabeur, G., Macaya, G., Kadi, F., Bernardi, G., 1993. The isochore patterns of mammalian genomes and their phylogenetic implications. J. Mol. Evol. 37, 93–108.

Salinas, J., Zerial, M., Filipski, J., Bernardi, G., 1986. Gene distribution and nucleotide sequence organization in the mouse genome. Eur. J. Biochem. 160, 469–478.

Sueoka, N., 1961. Correlation between base composition of the deoxyribonucleic acid and amino acid and composition of proteins. Proc. Natl. Acad. Sci. USA 47, 1141–1149.

Zoubak, S., Clay, O., Bernardi, G., 1996. The gene distribution of the human genome. Gene 174, 95–102.