Gene, 110 (1992) 81–88

GENE 06202

A universal compositional correlation among codon positions

(Isochores; coding sequences; prokaryotes; eukaryotes)

Giuseppe D'Onofrio * and Giorgio Bernardi

Laboratoire de Génétique Moleculaire, Institut Jacques Monod, 75005 Paris (France)

Received by E. Zuckerkandl: 13 August 1991
Revised/Accepted: 3 September/16 September 1991
Received at publishers: 15 October 1991

SUMMARY

We have investigated the compositional distributions of third codon positions of genes from the 16 prokaryotes and seven eukaryotes for which the largest numbers of coding sequences are available in data banks. In prokaryotes, both narrow and broad distributions were found. In eukaryotes, distributions were very broad (except for *Saccharomyces cerevisiae*) and remarkably different for different genomes. In low-GC genomes, third codon positions were lower in GC than first + second codon positions and trailed towards high GC; the opposite situation was found for high-GC genomes. In all genomes, first codon positions were higher in GC than second codon positions. We then investigated the compositional correlations between third and first + second codon positions in prokaryotic genomes (the 16 mentioned above plus 87 additional ones) and in genome compartments of eukaryotes. A general, common relationship was found, which also holds within the same (heterogeneous) genomes. This universal correlation is due to the fact that the relative effects of compositional constraints on different codon positions are the same, on the average, whatever the genome under consideration.

INTRODUCTION

Seven years ago, the localization of a number of genes in the isochores of vertebrate genomes led to the discovery that linear relationships hold between the GC levels of coding sequences (and of their third codon positions) on the one hand, and those of the isochores (or, more precisely, of the large DNA fragments) in which the genes were embedded, on the other hand (Bernardi et al., 1985; isochores are the long, >300 kb, compositionally homogeneous sequences that make up the vertebrate genomes; isochores belong to a small number of families characterized by different GC levels). This correlation appeared to hold independent of the species to which the genes belong, suggesting a general compositional correlation between coding sequences and the corresponding isochore families of vertebrates.

These findings prompted investigations (Bernardi and Bernardi, 1985), which showed that the GC levels of third codon positions of 302 codings sequences (or, rather, their average values per genome) from 49 genomes ranging from bacteria to man were positively correlated with the GC levels of the corresponding genomes, or with the GC levels of the isochore families (or compositional compartments), in the case of strongly compartmentalized genomes, like those from warm-blooded vertebrates. The slopes of the least-square lines through the points were very close for all classes of living organisms tested (viruses, bacteria, lower eukaryotes, invertebrates, vertebrates). In the case of eukaryotes, the line was, however, shifted to the left relative to those of prokaryotes and viruses, owing to the presence,

Correspondence to: Dr. G. Bernardi, Laboratoire de Génétique Moleculaire, Institut Jacques Monod, 2 Place Jussieu, Paris 75005 (France)
Tel. (33-1)43295824; Fax (33-1)44277977.
* Permanent address: Stazione Zoologica, Villa Comunale, 80121 Naples (Italy) Tel. (39-81)5833111.

Abbreviations: aa, amino acid(s); (%)GC, (% of) guanine + cytosine.

in eukaryotic genomes, of intergenic sequences characterized by lower GC levels. Indeed, when third codon positions from eukaryotic genes were plotted against GC levels of coding sequences (instead of against GC levels of genomes, or of their compositional compartments), lines for all classes of living organisms fell into quasi-coincidence and exhibited a slope of about 2. This was the first indication for the existence of a general linear relationship between the GC levels of third and those of first + second codon positions. Moreover, the correlation appeared to hold not only inter-genomically, but also intra-genomically, as suggested by the fact that points from different compositional compartments of the same genomes fell on the line of the intergenomic correlation.

Subsequent, more detailed work (Bernardi and Bernardi, 1986), using a slightly expanded sample of the coding sequences previously analyzed, showed (*i*) that the GC levels of first, second and third codon positions of averaged coding sequences from individual bacteria and viruses are positively correlated with the GC levels of the corresponding genomes, the slopes increasing from second to first and to third positions (similar results were reported for eleven bacterial genomes by Muto and Osawa, 1987); (*ii*) that the GC levels of first, second and third codon positions of individual genes (or of pooled genes from the same compositional compartments) from vertebrates are positively correlated with the GC levels of the corresponding coding sequences and genome compartments; (*iii*) that the slopes of the compositional correlations between individual codon positions and coding sequences (pooled according to the genome, or the genome compartment) were very close for all classes of living organisms (except for the second codon positions of viruses, a case which could, however, be accounted for). This stressed the universal nature of the relationship, which held both inter- and intra-genomically. It should be noted that plots identical to those of Bernardi and Bernardi (1986) were recently used to fit theoretical curves in an attempt to explain the reasons for the relationship (Wada et al., 1991).

More recently, the correlations between the GC levels of third and first + second codon positions were re-investigated in more detail for the genes from a single, strongly compartmentalized genome, the human genome (Aïssani et al., 1991; D'Onofrio et al., 1991), and for the genes of the weakly compartmentalized genomes from cold-blooded vertebrates (Bernardi and Bernardi, 1991). The linear correlation found for the coding sequences from cold-blooded vertebrates was very close to that found for human coding sequences (Bernardi and Bernardi, 1991), supporting the previous indications for the existence of a general (inter- and intra-genomic) relationship for all vertebrates. This result prompted a re-examination of the results, as found for the genes from the wide range of living organisms previously studied (Bernardi and Bernardi, 1985; 1986). As expected, this analysis revealed a common relationship between the GC levels of third and first + second codon positions for that vast array of organisms (Bernardi and Bernardi, 1991).

The above conclusions were considered to be important enough to deserve a more detailed study, which is presented here, and which concerns the compositional patterns of codon positions from prokaryotes and eukaryotes and the compositional correlation among codon positions from those organisms. The results obtained shed light on the reasons for the existence of the common relationships mentioned above.

RESULTS AND DISCUSSION

(a) Compositional patterns of prokaryotic genomes

The coding sequences analyzed in the present work were extracted from GenBank, Release 67 (June 1991). They comprised sequences from 16 prokaryotes and seven eukaryotes, including *S. cerevisiae*, *D. melanogaster*, and five vertebrates. These organisms were those containing the largest numbers of coding sequences known in their primary structures. Analyses were also done for the genomes of 87 other prokaryotes comprising smaller numbers of genes, also present in the same GenBank release.

Table I lists the prokaryotes from which the coding sequences analyzed were derived, the numbers of genes analyzed, their average GC levels, the average GC levels of first + second and third codon positions, as well as the ratios of GC levels of third over first + second codon positions.

The data of Table I show: (*i*) that the spread of mean GC values was greater for third codon positions (23–83%) than for first + second codon positions (37–51%), coding sequences showing intermediate values (33–64%); (*ii*) that the standard deviation of GC levels of third codon positions were systematically larger than those of the average GC levels of first + second codon positions (and of coding sequences); (*iii*) and that low GC genomes exhibited GC levels of third codon positions lower than those of first + second codon positions, whereas the opposite was true for high GC genomes; the transition between these two situations appeared to occur around 45% GC of coding sequences.

Figs. 1 and 2 show the compositional (GC) distributions of third codon positions for coding sequences of most bacterial genomes from Table I. These compositional distributions have been called compositional patterns and represent compositional genome phenotypes (Bernardi and Bernardi, 1986; other compositional phenotypes of the genome are represented by the distributions of large DNA fragments,

TABLE I

GC of codon positions from prokaryotes

Organism[a]	Number of genes examined	% GC[b]			
		I + II + III	I + II	III	III/I + II
1 *Staphylococcus aureus*	73	32.6 (3.6)	37.3 (4.4)	23.0 (5.1)	0.61
2 *Haemophilus influenzae*	34	37.3 (3.6)	42.2 (5.0)	27.4 (5.8)	0.65
3 *Lactococcus lactis*	41	37.5 (3.7)	40.3 (5.5)	31.4 (7.8)	0.77
4 *Chlamydia trachomatis*	45	41.6 (3.3)	45.7 (4.5)	33.2 (5.8)	0.72
5 *Bacillus subtilis*	305	42.8 (4.1)	43.6 (4.5)	41.1 (7.5)	0.94
6 *Escherichia coli*	1195	51.2 (5.5)	49.6 (4.8)	54.4 (9.9)	1.10
7 *Bacillus stearothermophilus*	51	49.7 (7.1)	46.6 (6.2)	56.1 (13.2)	1.17
8 *Salmonella typhimurium*	189	53.0 (5.0)	50.2 (3.9)	58.5 (9.3)	1.16
9 *Neisseria gonorrhoeae*	27	53.9 (3.5)	51.1 (6.1)	59.4 (8.2)	1.16
10 *Anacystis nidulans*	17	53.9 (5.0)	49.5 (3.4)	62.8 (11.8)	1.27
11 *Erwinia chrysanthemi*	28	54.7 (3.4)	49.3 (3.5)	65.5 (6.0)	1.32
12 *Klebsiella pneumoniae*	80	59.9 (5.8)	53.3 (4.6)	73.1 (10.6)	1.37
13 *Halobacterium halobium*	35	62.6 (4.9)	54.2 (4.3)	79.1 (12.8)	1.46
14 *Pseudomonas aeruginosa*	85	63.5 (6.0)	54.5 (5.0)	81.3 (13.5)	1.49
15 *Rhodobacter capsulatus*	49	64.5 (3.8)	55.0 (5.4)	83.1 (4.1)	1.52
16 *Azotobacter vinelandii*	46	62.8 (3.7)	51.8 (4.2)	84.5 (5.5)	1.63
	2300				

[a] Organisms are arranged in order of increasing GC of third codon positions.

[b] I, II and III indicate first, second, and third codon positions, respectively. Average values are given for I + II + III and I + II. Values in parentheses represent standard deviations.

like those shown in Fig. 1 of Bernardi et al., 1985; see also Bernardi, 1989).

The histograms of Fig. 1 concern genomes characterized by standard deviations of GC levels of third codon positions lower than 6% GC (with one exception, *B. subtilis*, which showed a standard deviation of 7.5% GC). In contrast, the histograms of Fig. 2 concern the genomes which exhibited a broad compositional distribution of their genes; these genomes were characterized by standard deviations of GC levels of third codon positions higher than 9% GC.

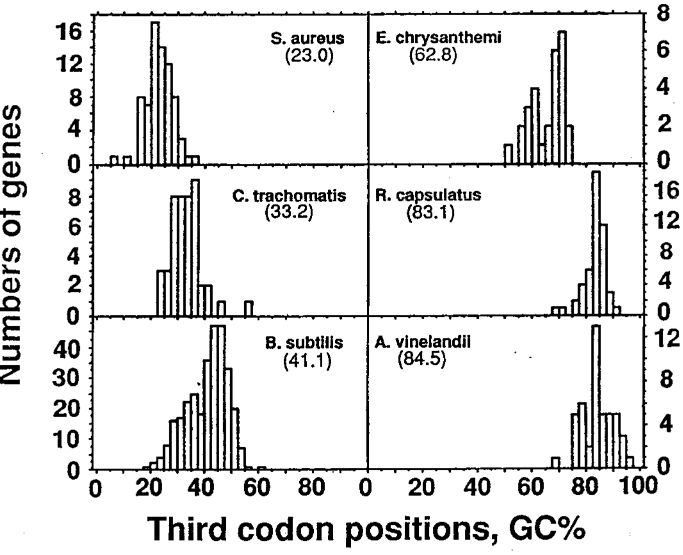

Fig. 1. Compositional patterns of 'homogeneous' prokaryotic genomes. Histograms show the distributions of GC levels of third codon positions of genes from the indicated bacterial species arranged in order of increasing GC levels (see Table I). Values in parentheses are the GC levels of third codon positions.

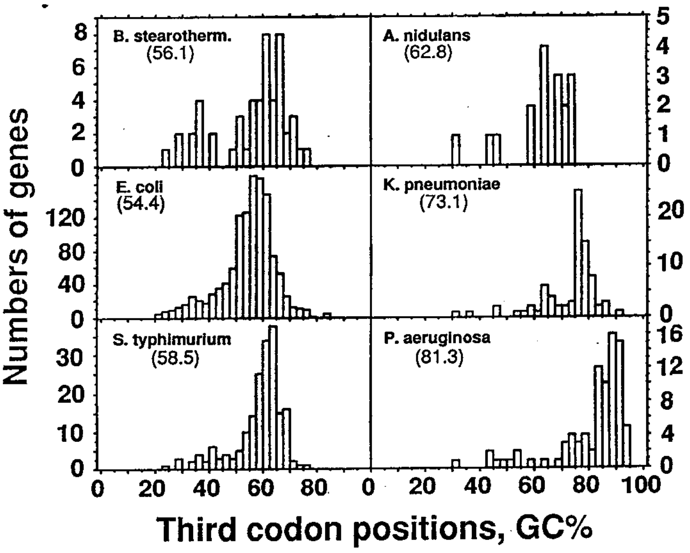

Fig. 2. Compositional patterns of 'heterogeneous' prokaryotic genomes. Histograms show the distributions of GC levels of third codon positions of genes from the indicated bacterial species arranged in order of increasing GC levels (see Table I). Values in parentheses are the GC levels of third codon positions.

The few intermediate cases of Table I are not represented in the figures, but they show the expected intermediate features.

A conclusion that can be drawn from the results of Figs. 1 and 2 is that, while a number of bacterial species are characterized by narrow compositional distribution of the corresponding genes, other species exist that exhibit a broad spectrum, trailing away from the peak towards high GC values, if the peak is low in GC, and vice versa; in one case at least, that of *B. stearothermophilus*, the compositional pattern is clearly bimodal. In the cases in which this point has been investigated, it has been observed that low and high GC coding sequences are clustered (Nomura et al., 1987; Sharp et al., 1989; Sharp, 1990). This is an indication that an isochore organization may exist even in prokaryotic genomes, as predicted (Bernardi et al., 1985).

The differential constraints on the three codon positions are indicated by the correlations between GC levels of the three codon positions. Fig. 3 shows that GC levels of third codon positions are higher (values above the diagonal line) or lower (values below the diagonal line) than those of first or second positions. On the other hand, for GC-rich and GC-poor genomes, respectively, the GC levels of first codon positions are always higher than those of second codon positions. For GC-rich genomes, first codon positions are intermediate in GC between third and second positions; for GC-poor genomes, this is not the case, the decreasing GC order being first, second and third codon position. This is due to the more common occurrence of aa whose codons have G or C in first codon positions (Leu, Val, His, Gln, Asp and Glu) and A or T in second codon positions compared to those in which the reverse is true (Ser, Thr, Cys, Trp, Arg doublet), as already noticed by Sueoka (1988). Fig. 4 shows, for example, the compositional patterns obtained for the three codon positions of *S. aureus* and *A. vinelandii*, two genomes characterized by extreme GC levels. It is clear that, while third codon positions do exhibit extreme GC levels in both cases, first codon positions are higher than second codon positions not only in the high GC genome, but also in the low GC genome. In other words, first codon positions occupy the expected intermediate level between third and second codon position in high GC genomes, but not in low GC genomes, for the reasons given above. Incidentally, the parallel behavior of GC levels in first and second codon positions (Fig. 3) justifies using the average GC levels of those positions throughout this work.

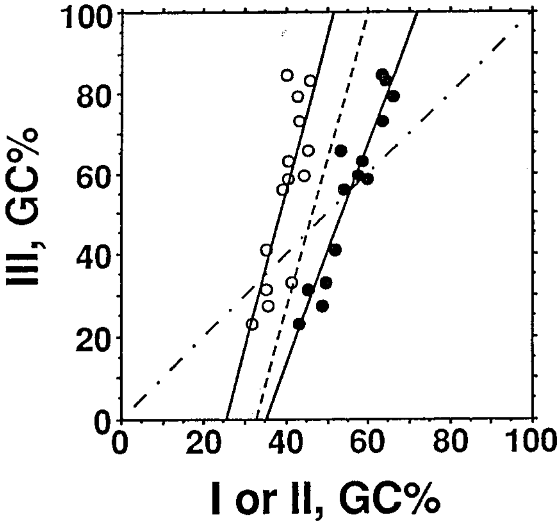

Fig. 3. Plots of GC levels of third codon positions from prokaryotic genes averaged per genome against the corresponding GC levels of first (blackened circles) and second (open circles) codon positions. The equations for the least-square lines and the squares of correlation coefficients are $y = 2.7x - 95.7$; $r^2 = 0.9$ for the first position plot, and $y = 3.8x - 97.0$; $r^2 = 0.6$ for the second position plot. The dashed line is the plot for first + second codon positions; in this case, $y = 3.6x - 119.6$; $r^2 = 0.89$. The diagonal (slope = 1) line is indicated by a point-dash line. I, II, III indicate first, second and third codon positions, respectively.

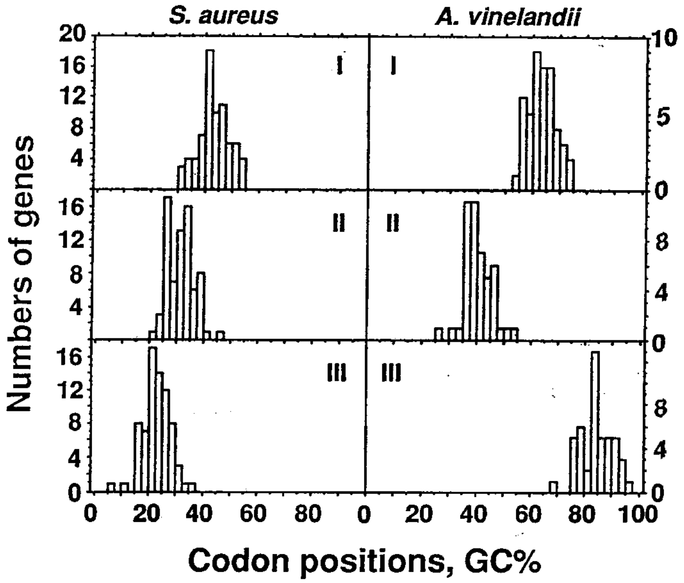

Fig. 4. Compositional patterns of the genomes of *S. aureus* and *A. vinelandii*. Histograms show the distributions of GC levels of the three codon positions (indicated as **I**, **II**, **III**, respectively).

(b) Compositional patterns of eukaryotic genomes

Table II presents the same sets of data as Table I for seven eukaryotic genomes. The most remarkable feature is that, with one exception (*S. cerevisiae*), the standard deviations of GC levels of third codon positions are very high, ranging from 10–17% GC.

The compositional patterns represented by the distribution of GC levels of third codon positions of the eukariotic genomes investigated are displayed in Figs. 5–7. Their main features are the following.

(*i*) The coding sequences from *S. cerevisiae* are charac-

TABLE II

GC of codon positions from eukaryotes

Organism[a]	Number of genes examined	% GC[b]			
		I + II + III	I + II	III	III/I + II
1 *Saccharomyces cerevisiae*	833	41.2 (3.8)	41.8 (4.0)	40.0 (6.9)	0.96
2 *Drosophila melanogaster*	407	55.4 (5.5)	49.0 (4.8)	68.1 (12.0)	1.39
3 *Xenopus laevis*	213	49.5 (5.3)	48.2 (5.0)	52.0 (12.7)	1.08
4 *Gallus gallus*	288	51.1 (7.6)	48.1 (4.9)	66.2 (16.9)	1.38
5 *Mus musculus*	944	52.7 (6.1)	48.2 (5.3)	61.5 (11.7)	1.28
6 *Rattus norvegicus*	835	52.2 (5.2)	47.5 (4.5)	61.4 (10.2)	1.30
7 *Homo sapiens*	2019	53.9 (7.4)	49.1 (5.8)	63.1 (15.0)	1.28

[a] Organisms are arranged in a taxonomical order.

[b] See footnote b of Table I.

terized by a low GC level and a low standard deviation (41.2 ± 3.8% GC). Third codon positions (Fig. 5) have low GC levels and a low standard deviation (40.0 ± 6.9% GC); they trail towards high GC levels, and have a slightly lower average GC level and a slightly higher standard deviation than first + second codon positions (41.8 ± 4.0% GC). Overall, this eukaryotic genome is very similar in its compositional pattern to a 'homogeneous' bacterial genome of similar GC level (Fig. 1).

(*ii*) The coding sequences from *D. melanogaster* (55.4% GC; 5.5% GC) are characterized by a very wide spread of compositional distributions (Fig. 5) of third codon positions (68.1 ± 12.0% GC), which are, on the average, much higher in GC than first + second codon positions (49.0 ± 4.8% GC), and trail towards the low GC side of the histogram.

(*iii*) The coding sequences from *X. laevis* (49.5 ± 5.3% GC) present a wide spread of compositional distribution

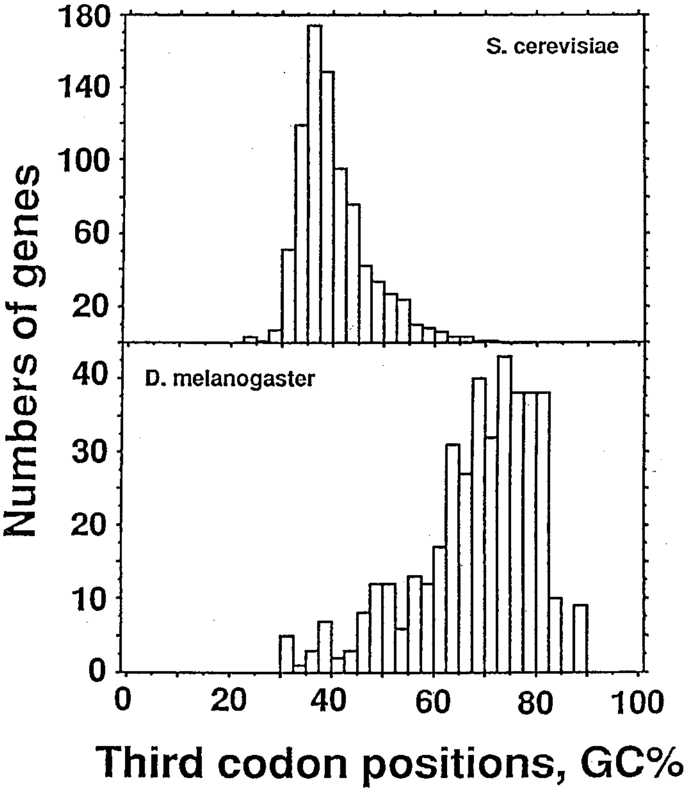

Fig. 5. Compositional patterns of eukaryotic genomes. Histograms show the distributions of GC levels of third codon positions of genes from the eukaryotic species indicated (see Tables II and III).

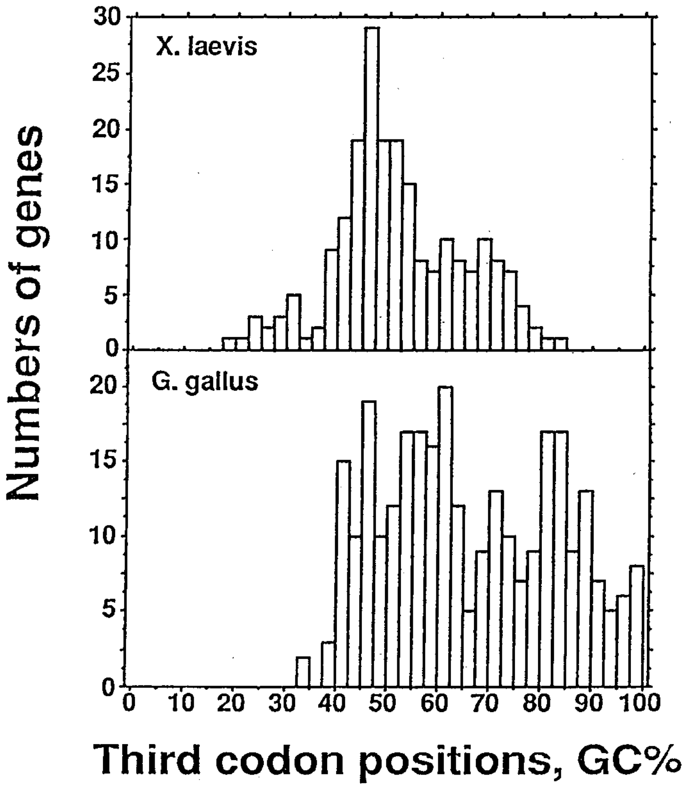

Fig. 6. Compositional patterns of eukaryotic genomes. Histograms show the distributions of GC levels of third codon positions of genes from the eukaryotic species indicated (see Tables II and III).

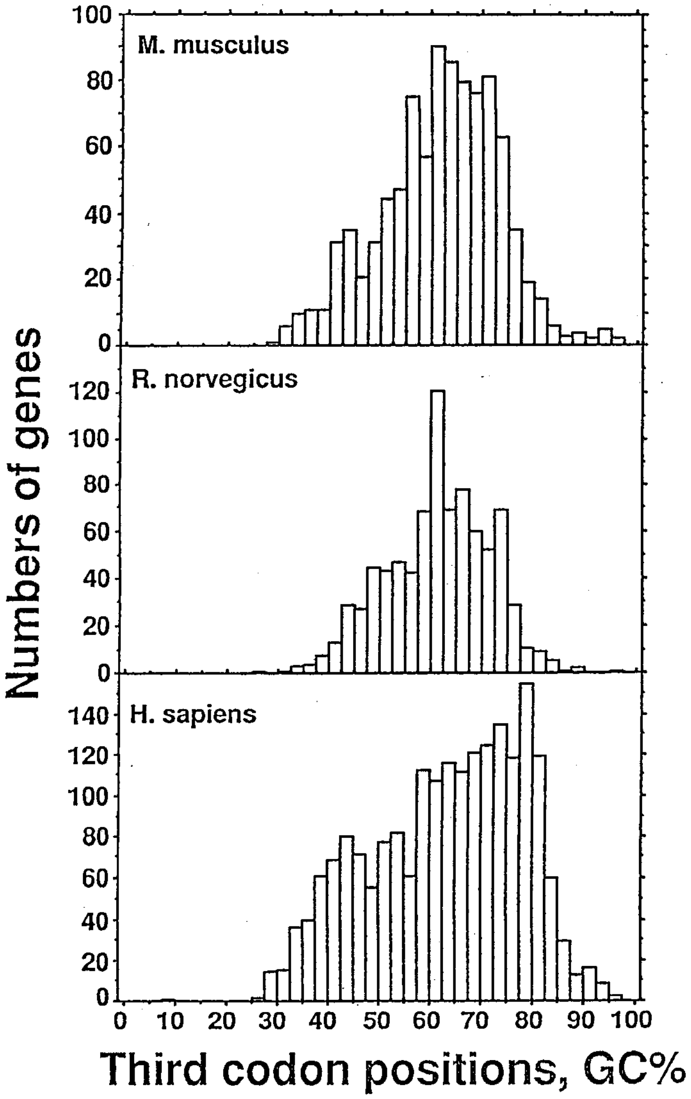

Fig. 7. Compositional patterns of eukaryotic genomes. Histograms show the distribution of GC levels of third codon positions of genes from the eukaryotic species indicated (see Tables II and III).

(Fig. 6) of third codon positions (52.0 ± 12.7% GC), but their average value is higher than that of first + second codon positions (48.2 ± 5.0% GC), and the skewness is mainly on the high GC side; a limited skewness on the low GC side is due to genes encoding small peptides.

(*iv*) The coding sequences of chicken (51.1 ± 7.6% GC) exhibit a very wide spread of GC levels of third codon positions (Fig. 6), which attain 100% GC, and are, on the average, much higher than those of first + second codon positions (66.2 ± 16.9 vs. 48.1 ± 4.9% GC).

(*v*) The coding sequences from mouse and rat (52.7 ± 6.1 and 52.2 ± 5.2% GC, respectively) are characterized by broad compositional distributions (Fig. 7) of third codon positions (61.5 ± 11.7% GC; 61.4 ± 10.2% GC) and higher GC values compared to those of first + second codon positions (48.2 ± 5.3% GC; 47.5 ± 4.5% GC).

(*vi*) The coding sequences from the human genome (53.9 ± 7.4% GC) show the same features as the other two mammalian genomes in the compositional distributions (Fig. 7) of third codon positions (63.1 ± 15.0% GC), which reach higher GC values than first + second codon positions (49.1 ± 5.8% GC).

The very broad compositional distributions of GC levels of third codon positions exhibited by the genomes of all eukaryotes tested, except *S. cerevisiae*, prompted an analysis of their compositional compartments. If these genomes were divided in three compartments (except *D. melanogaster* which was divided in two compartments; Table III), the standard deviations of GC levels found for each compartment not only were smaller than for total genomes, as expected, but reached values comparable to those exhibited by 'homogeneous' prokaryotic genomes. This justifies the division of such genomes into compartments for the analyses.

(c) The compositional correlation among codon positions from different genomes

When the GC levels (averaged per genome) of third codon positions from the prokaryotic genomes of Table I were plotted against those of first + second codon positions (Fig. 8A), the least-square line through the points exhibited a slope of 3.6 and an excellent correlation coefficient ($r^2 = 0.88$). A practically identical relationship (Fig. 8B) was obtained by plotting data from 87 other prokaryotic genomes, each one of which comprises a smaller number of genes than the genomes of Table I.

Only a slightly higher slope was obtained (Fig. 8C) if values from eukaryotic genomes were plotted using the data for compartments (see Table III). Under these circumstances, it is not surprising that, when plotted together, all points from eukaryotes and prokaryotes were found to follow the same relationship (Fig. 8D).

The same conclusion can be drawn by plotting the ratios of GC levels of third over first + second codon positions against GC levels of coding sequences (as given in Tables I–III), in that these plots were identical for prokaryotes and eukaryotes (not shown).

(d) The compositional correlation among codon positions from compartments of the same genomes

As far as the existence of a compositional correlation among codon positions from coding sequences of the same genome, and its identity with the inter-genomic correlation discussed above are concerned, expectedly one does not find such correlation within compositionally 'homogeneous' genomes or genome compartments, because the points in such cases are too clustered to lead to significant correlations.

The situation is, however, different for heterogeneous

TABLE III

GC levels of coding sequence from compositional compartments of eukaryotic genomes

Organism[a]	Number of genes examined	% GC[b]				
		Range	I + II + III	I + II	III	III/I + II
1 *Saccharomyces cerevisiae*	833	22.5–72.5	41.2 (3.8)	41.8 (4.0)	40.0 (6.9)	0.96
2 *Drosophila melanogaster*	102	30.0–63.0	49.2 (6.4)	48.1 (6.6)	51.1 (8.8)	1.0
3 *Drosophila melanogaster*	305	63.0–90.0	57.5 (3.2)	49.3 (4.2)	73.7 (6.2)	1.5
4 *Xenopus laevis*	16	17.5–35.0	44.6 (5.0)	53.0 (9.0)	27.5 (4.6)	0.5
5 *Xenopus laevis*	139	35.0–60.0	47.5 (3.4)	47.2 (4.3)	47.9 (5.4)	1.0
6 *Xenopus laevis*	58	60.0–85.0	55.6 (3.8)	49.1 (4.0)	68.5 (5.6)	1.4
7 *Gallus gallus*	88	32.5–55.0	46.2 (2.8)	45.7 (3.5)	46.9 (5.0)	1.1
8 *Gallus gallus*	103	55.0–75.5	53.8 (3.7)	48.8 (4.3)	63.6 (5.9)	1.3
9 *Gallus gallus*	97	75.5–100	62.6 (4.3)	50.6 (5.3)	86.3 (6.4)	1.7
10 *Mus musculus*	248	27.5–55.0	46.0 (4.8)	45.7 (6.0)	46.2 (6.3)	1.0
11 *Mus musculus*	614	55.0–75.5	54.4 (3.9)	48.8 (4.6)	65.0 (5.6)	1.3
12 *Mus musculus*	82	75.5–97.5	61.2 (4.4)	51.1 (5.5)	81.1 (5.5)	1.6
13 *Rattus norvegicus*	219	25.0–55.0	46.6 (3.8)	45.8 (5.2)	47.9 (5.2)	1.0
14 *Rattus norvegicus*	565	55.0–75.5	53.6 (3.4)	47.8 (4.0)	64.9 (5.5)	1.4
15 *Rattus norvegicus*	51	75.5–97.5	60.4 (5.0)	50.7 (6.4)	79.5 (3.9)	1.6
16 *Homo sapiens*	670	25.0–57.5	46.3 (4.9)	46.6 (6.1)	45.3 (7.5)	1.0
17 *Homo sapiens*	690	57.5–72.5	54.6 (4.0)	49.2 (5.2)	65.2 (4.4)	1.3
18 *Homo sapiens*	659	72.5–97.5	61.0 (4.0)	51.7 (5.0)	79.1 (4.8)	1.5

[a] See footnote [a] of Table II.

[b] See footnote [b] of Table I.

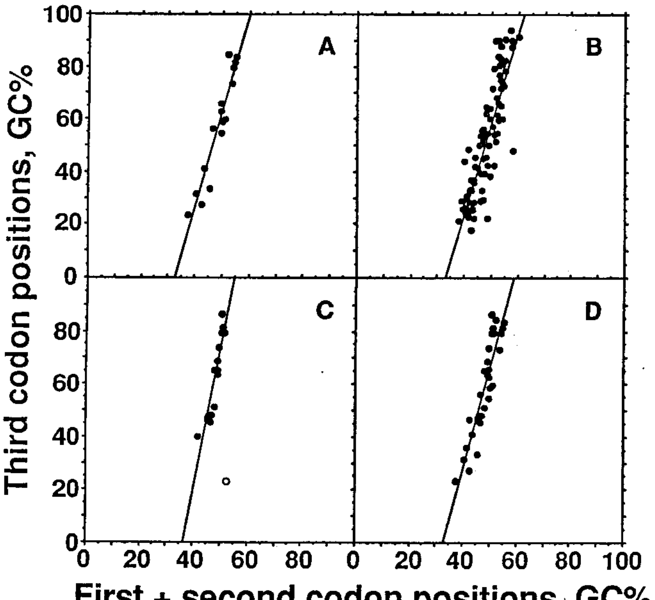

Fig. 8. Plots of GC levels of third codon positions against GC levels of first + second positions (all values are averaged per genome or genome compartment) of prokaryotic and eukaryotic genomes. The equations of least-square lines and the squares of the correlation coefficients are given in parentheses. Plots concern: **A**, the prokaryotic genomes of Table I ($y = 3.6\,x - 119$; $r^2 = 0.88$); **B**, 87 additional prokaryotic genomes ($y = 3.4\,x - 112$; $r^2 = 0.73$); **C**, the compositional compartments of eukaryotic genomes of Table III ($y = 5.4\,x - 197$; $r^2 = 0.82$); the deviating point (open circle) was not taken into account in calculating the least-square line and the correlation coefficient, this point corresponds to the low GC compartment of *X. laevis*; as indicated in the text, this compartment comprises genes for small peptides; **D**, the prokaryotic genomes of Table I and the compositional compartments of eukaryotic genomes of Table III ($y = 3.9x - 127$; $r^2 = 0.81$); the deviating point of panel **C** is not shown.

genomes. For instance, in the case of human coding sequences, a good correlation very close to that of Fig. 8D was found (D'Onofrio et al., 1991), and in the case of *E. coli* the least-square-line equation was $y = 3.5\,x - 116$, and the correlation coefficient was 0.81. A similar correlation had been found for coding sequences from *E. coli* bacteriophages (Wada and Suyama, 1985).

(e) Conclusions

Here, we have investigated the compositional patterns exhibited by a number of prokaryotic and eukaryotic genomes, as well as the compositional correlations holding between third and first + second codon positions, both among different genomes and within given genomes (those characterized by large compositional heterogeneities).

The compositional patterns, as investigated through the distributions of GC levels of third codon positions, have revealed a number of novel features. (*i*) As far as prokaryotic genomes are concerned, a number of them are characterized by narrow compositional patterns, but others are not. The compositional homogeneity of the former set fits with the traditional picture derived from the CsCl profiles of prokaryotic genomes (Rolfe and Meselson, 1959; Sueoka, 1959), but clearly prokaryotic genomes exist which are remarkably heterogeneous. (*ii*) In the case of eukaryotic genomes, the characteristic features are not only their very large compositional heterogeneities (with, however, the exception of *S. cerevisiae*), but also the varieties of compositional patterns (distribution profiles).

If compositional correlations between third and first + second codon positions are considered for different genomes (genome compartments being considered as genomes in this respect), all points fall on the same line for both prokaryotes and eukaryotes. This also occurs for points belonging to different compartments within the same compartmentalized genome.

The general conclusion of the present work is that there is, indeed, a universal compositional correlation among the three codon positions. This correlation indicates not only that compositional constraints exist on genomes or genome compartments, but also that their relative effects on the different codon positions are the same, on the average, whatever the genome under consideration. One should expect deviations from the correlations due to special constraints associated with the coding of certain aa or with particular codon usages associated with gene expressivity but these effects are practically cancelled by the fact that average values of codon positions for homogeneous genomes or genome compartments are used. It has been pointed out elsewhere that this universal correlation amounts to a genomic code (Bernardi, 1990), which also encompasses the compositional correlations between coding sequences (and their codon positions) and isochores of eukaryotes, namely between coding and non-coding sequences (Bernardi et al., 1985; Aïssani et al., 1991).

The origin of the compositional constraints (also called GC/AT pressures) is still a matter of debate. One extreme view, first proposed 30 years ago, is that compositional constraints (at least as seen in the homogeneous bacterial genomes studied at that time) are the result of mutational biases (Freese, 1962; Sueoka, 1962). This view requires additional ad hoc hypotheses (Sueoka, 1988; Wolfe et al., 1989), in the case of compartmentalized genomes. The opposite view is that natural selection plays the main role leading to preferential fixation of directional mutations (Bernardi and Bernardi, 1986; Bernardi et al., 1988). A discussion on this point will be presented elsewhere.

ACKNOWLEDGEMENTS

We thank Dominique Mouchiroud and Christian Gautier for having kindly provided the data files extracted from GenBank and for their interest, Giacomo Bernardi and Giorgio Matassi for useful discussions, the Association Française contre les Myopathies (AFM), the Association pour la Recherche contre le Cancer (ARC) and the Ministry for Research and Technology (MRT) for financial support.

REFERENCES

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

Bernardi, G.: The isochore organization of the human genome. Annu. Rev. Genet. 23 (1989) 637–661.

Bernardi, G.: Le genome des vertébrés: organisation, fonction, évolution. Biofutur 94 (1990) 43–46.

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

Bernardi, G. and Bernardi, G.: Compositional constraints and genome evolution. J. Mol. Evol. 24 (1986) 1–11.

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

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

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

Freese, E.: On the evolution of the base composition of DNA. J. Theor. Biol. 3 (1962) 82–101.

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

Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C. and Bernardi, G.: The distribution of genes in the human genome. Gene 100 (1991) 181–187.

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

Nomura, M., Sor, F., Yamagishi, M. and Lawson, M.: Heterogeneity of GC content within a single bacterial genome and its implication for evolution. Cold Spring Harbor Symp. Quant. Biol. 52 (1987) 658–663.

Rolfe, R. and Meselson, M.: The relative homogeneity of microbial DNA. Proc. Natl. Acad. Sci. USA 45 (1959) 1039–1043.

Sharp, P.M.: Processes of genome evolution reflected by base frequency differences among *Serratia marcescens* genes. Mol. Microbiol. 4 (1990) 119–122.

Sharp, P.M., Shields, D.C., Wolfe, K.H. and Li, W.-H.: Chromosomal location and evolutionary rate variation in enterobacterial genes. Science 246 (1989) 808–810.

Sueoka, N.: Statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. Proc. Natl. Acad. Sci. USA 45 (1959) 1480–1490.

Sueoka, N.: On the genetic bases of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA 48 (1962) 582–592.

Sueoka, N.: Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85 (1988) (2653–2655.

Wada, A. and Suyama, A.: Third letters in codons counterbalance the (G + C) content of their first and second letters. FEBS Lett. 188 (1985) 291–294.

Wada, A., Suyama, A. and Hanori, R.: Phenomenological theory of GC/AT pressure on DNA base composition. J. Mol. Evol. 32 (1991) 374–378.

Wolfe, K.H., Sharp, P.M. and Li, W.-H.: Mutation rates differ among regions of the mammalian genome. Nature 337 (1989) 283–285.