J Mol Evol (1991) 32:504–510

Journal of Molecular Evolution

88

Correlations between the Compositional Properties of Human Genes, Codon Usage, and Amino Acid Composition of Proteins

Giuseppe D'Onofrio,[1]* Dominique Mouchiroud,[2] Brahim Aïssani,[1] Christian Gautier,[2] and Giorgio Bernardi[1]

[1] Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France
[2] Laboratoire de Biométrie, Génétique et Biologie des Populations, U.R.A. 243, Université Claude Bernard, 69600 Villeurbanne, France

Summary. We have analyzed the correlation that exists between the GC levels of third and first or second codon position for about 1400 human coding sequences. The linear relationship that was found indicates that the large differences in GC level of third codon positions of human genes are paralleled by smaller differences in GC levels of first and second codon positions. Whereas third codon position differences correspond to very large differences in codon usage within the human genome, the first and second codon position differences correspond to smaller, yet very remarkable, differences in the amino acid composition of encoded proteins. Because GC levels of codon positions are linearly correlated with the GC levels of the isochores harboring the corresponding genes, both codon usage and amino acid composition are different for proteins encoded by genes located in isochores of different GC levels.

Furthermore, we have also shown that a linear relationship with a unity slope and a correlation coefficient of 0.77 exists between GC levels of introns and exons from the 238 human genes currently available for this analysis. Introns are, however, about 5% lower in GC, on average, than exons from the same genes.

Key words: Human genome — Amino acids — Isochores — Coding sequences — Introns — Codon positions

Introduction

Previous investigations (Bernardi and Bernardi 1985, 1986) showed that the GC levels of the three codon positions show strong linear correlations with those of the corresponding genomes. Such relationships were essentially identical for the genomes (or genome compartments, in the case of compositionally compartmentalized genomes) of all organisms explored, which comprised prokaryotes, uni- and multicellular eukaryotes, and viruses (small differences shown by the second codon positions of animal viruses, as well as those due to the intergenic sequences of vertebrates are neglected here; see Bernardi and Bernardi 1986). These results pointed to the existence of compositional correlations among different codon positions and to changes in codon usage and amino acid utilization associated with compositional changes in the corresponding coding sequences.

In the present work, we have studied in some detail the compositional correlations among codon positions of human genes in order to define them as precisely as possible, and we have investigated their consequences with regard to codon usage and to the amino acid composition of the encoded proteins.

The choice of human genes in this study was primarily dictated by the following reasons: (1) the availability (as of March 1990) of a large sample of about 1400 coding sequences with known primary structure, and of 238 genes with known exon and intron sequences; (2) the extremely widespread compositional pattern of human genes: the GC levels of human coding sequences cover a GC range of

* On leave from the Stazione Zoologica, Villa Comunale, 80121 Naples, Italy
Offprint requests to: G. Bernardi

33–77%, those of third codon positions a 27–97% range—these are extremely extended ranges, almost as wide as those exhibited by the genes from all bacterial species studied so far; (3) the similarity of the compositional pattern of the human genome with those of the genomes from mammals and warm-blooded vertebrates in general (see Bernardi 1989, for a review). Thus it is possible to extend the conclusions drawn here to these other genomes.

Another reason for this study was to compare the compositional correlation among codon positions as obtained for the 1400 coding sequences with those obtained for a small sample of genes used in defining the compositional correlation between human genes and the isochores in which the genes are located (Aïssani et al. 1991). This comparison was used to decide whether the sample of localized genes was representative of human genes in general. For the same reason, we investigated the correlation between GC levels of exons and introns from human genes.

Materials and Methods

Weight-average GC levels of all exons, different codon positions, and introns of any given gene were used. Exons (but not introns) preceding the initiation codon AUG were disregarded. The gene sequences used here were all those available from Release 63 (March 15, 1990) of GenBank. They were obtained using the ACNUC retrieval system (Gouy et al. 1985). Duplicate sequences were excluded.

Orthogonal regressions were used in plots of GC levels of third codon positions and introns against GC levels of first and/or second codon positions. This approach minimizes the sum of square distances between points and regression line. The use of orthogonal regression is appropriate in the cases under consideration. Indeed, large differences exist in the spread of points along the two cartesian axes; moreover, our aim here was to obtain a good representation of the scatter diagram, rather than to find (as with the linear regression) how one variable depended upon the other one.

Results

When GC levels of third codon positions of all of the 1400 or so human coding sequences available in March 1990 were plotted against GC levels of first or second codon positions, positive relationships were found with slopes of 3.6 and 5.3, and with correlation coefficients of 0.50 and 0.29, respectively (Fig. 1). A linear relationship with a slope of 4.6 and a correlation coefficient of 0.52 was found when the GC level of third codon positions was plotted against that of first + second codon positions (Fig. 1). It should be noted that a small set of strongly deviating genes (comprising collagen, elastin, and proline-rich protein genes), which are represented in Fig. 1 as open circles (surrounded by a broken line in the plot of first + second codon positions), was not taken into account in calculating the correlation coefficients and the slopes.

The partial overlap of points in Fig. 1 obscures their actual distribution in the diagram. This distribution is shown in Fig. 2, which presents histograms of compositional classes corresponding to 2.5% GC intervals in first + second codon positions and 5% GC intervals in third codon positions. The data in Fig. 2 stress the fact that the positive correlation seen in Fig. 1 corresponds, as expected, to a line passing through the most abundant classes of genes. It should be noticed, however, that the abundant classes at 75–85% GC in third codon positions and 55–60% in first + second positions suggests that the linear relationship tends to bend to the right at such high values; and that only very few extreme genes lie outside lines that are parallel to the linear relationship, but are shifted by ±5% GC along the abscissa axis (note that the set of strongly deviating genes of Fig. 1, not represented in Fig. 2, would also belong to this class).

The scatter of points in the plots of Fig. 1 deserves to be analyzed further. Indeed, this scatter is obviously not due to any experimental error (because the points derive from primary structures), but to some intrinsic factor that needs to be understood. In fact, it can be shown easily that the scatter of points along the horizontal axis in Fig. 1 is due to particular frequencies of amino acids in the encoded proteins. Indeed, amino acids can be divided into three classes: (1) those that only contain G and/or C in the first and second positions of their codons; (2) those that only contain A and/or T; and (3) those that contain G and/or C as well as A and/or T. The GC class comprises four amino acids, alanine, arginine (quartet codons), glycine, and proline; the AT class comprises seven amino acids, asparagine, isoleucine, leucine (duet codons), lysine, methionine, phenylalanine, and tyrosine; the intermediate class contains all other amino acids. Remarkably, although the GC class corresponds only to quartet codons, the AT class corresponds only to duet codons, except for isoleucine and methionine codons (a triplet codon and a singlet codon, respectively); in contrast, the intermediate class corresponds to both quartet and duet codons. The GC and AT classes comprise 11 amino acids (two of them, arginine and leucine, only in some of their codons) and 30 codons (see Fig. 3).

Different GC levels of first + second codon positions correspond to variations in the molar ratio of the GC class over the AT class, namely to changes concerning half of the codons. Indeed, a plot of the GC level of first + second codon position against the logarithm of the GC class/AT class codon ratio yields, as expected, a straight line with a correlation coefficient of 1 (not shown).

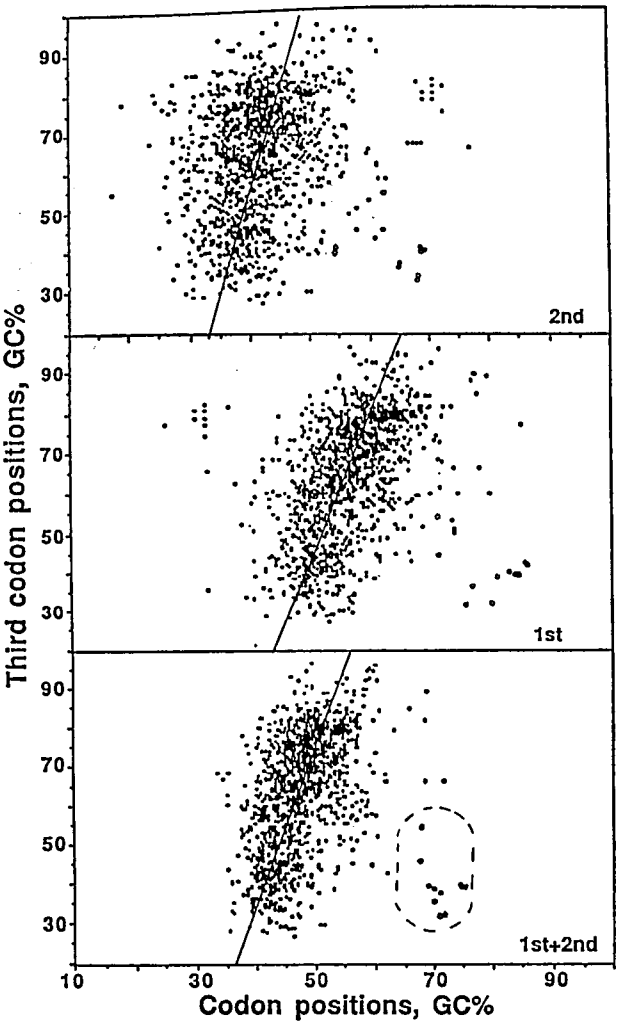

Fig. 1. Plot of GC levels of third codon positions against GC levels of first (**a**), second (**b**), and first + second codon positions (**c**). A set of 1380 human coding sequences was used. The solid lines are the orthogonal regression lines (see Materials and Methods) through the points (neglecting the points represented by open circles surrounded by a broken line in the plot of first + second codon positions; see text).

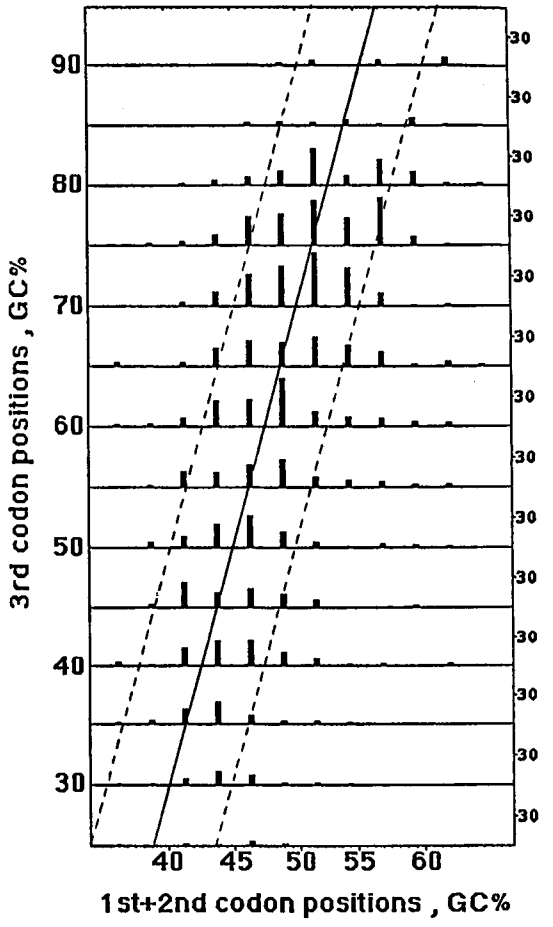

Fig. 2. Number of coding sequences belonging to 5% GC intervals in third codon positions and 2.5% GC intervals in first + second codon positions. The solid line is the orthogonal regression line; the broken lines correspond to a shift of the regression line by ±5% GC along the abscissa. Points corresponding to open circles in Fig. 1 were not represented, because they would fall outside the frame of Fig. 2.

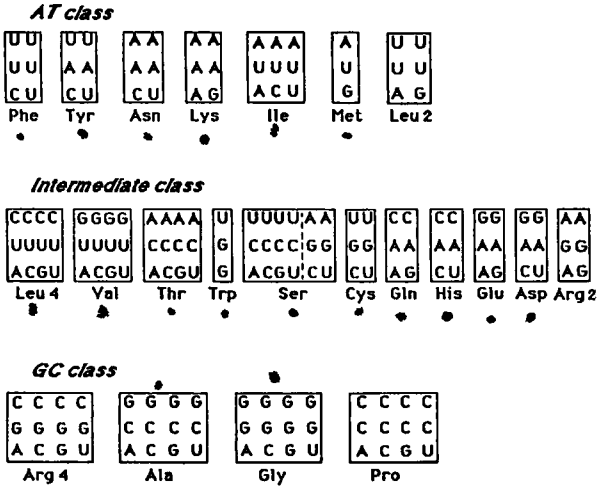

Fig. 3. Table of codons modified from Grantham (1980) to show the AT, GC, and intermediate classes (these classes concern first and second codon position, see text)

When molar percentages for individual amino acids were plotted against the GC levels of first + second codon positions (see Fig. 4) absolute correlation coefficients were found (Table 1) to be in the 0.25–0.6 range for amino acids of the GC class or the AT class (with, however, a lower value for lysine, 0.14). In the first case slopes were positive; in the second they were negative. In contrast, the intermediate class showed very poor absolute correlation coefficients, below 0.25. As a consequence, the corresponding slopes (in italics in Table 1) had weak or no significance.

The genes used in Fig. 4 and in Table 1 comprise the two sets studied in the preceding paper (see Table 1 of Aïssani et al. 1991), as well as a set of extreme genes (see Table 2 for a list), namely the genes that most deviate from the linear correlation of Fig. 2 along the abscissa. As it appears from Fig. 4, there

Table 1. Slopes and correlation coefficients from Fig. 4

Amino acid	Slope	*R*
Phe	−1.36	0.53
Ile	−1.52	0.47
Met	−0.57	0.30
Tyr	−0.63	0.26
Asn	−1.02	0.44
Lys	−0.71	0.14
Leu 2	−0.89	0.39
Leu 2 + 4	−2.01	0.38
Leu 4	*−1.19*	0.23
Val	*−0.29*	0.09
Trp	*−0.30*	0.19
Thr	*−0.47*	0.17
Ser	*−0.56*	0.14
Cys	*0.03*	0.01
Gln	*0.44*	0.13
His	*−0.43*	0.20
Asp	*−0.85*	0.22
Glu	*−0.90*	0.18
Arg 2	*0.55*	0.21
Arg 2 + 4	1.49	0.36
Arg 4	0.94	0.26
Ala	1.51	0.30
Gly	4.21	0.57
Pro	4.27	0.63

Absolute values are given for correlation coefficients. Slopes in italics are not statistically significant

is a considerable scatter of points when individual amino acids are considered, instead of whole classes. Some exceptions do exist, however, the most remarkable one being that of proline. In the case of leucine and arginine, relationships are shown for both classes of codons, encoding each one of these amino acids, as well as for the sums of the two.

Finally, diagrams of GC levels of introns plotted against GC levels of the corresponding exons or third codon positions are shown in Fig. 5. In the first case, a slope of 1.02 and a correlation coefficient of 0.77 were found; the least-square line is slightly lower, by about 5% GC, than the unity slope line through the origin. In the second case, a slope of 0.56 and a correlation coefficient of 0.81 were found. A plot of intron GC level against the GC level of first + second codon positions exhibited a correlation coefficient of 0.53 and a slope of 2.75 (Fig. 6).

Discussion

The present work dealt mainly with the problem of the compositional correlations among codon positions using a large sample of about 1400 coding sequences from a single genome, the human genome. It is clear that GC levels of third codon positions are positively correlated with those of first and/or second positions, as expected from previous work (Bernardi and Bernardi 1985, 1986). The correlations of Figs. 1 and 2 indicate that, on the average, large GC increases in third codon positions are accompanied by small GC increases in first and second positions. In other words, large codon usage changes are accompanied by smaller, yet very remarkable, amino acid changes in the proteins encoded by the corresponding genes. It should be noted that the relationship of Figs. 1 and 2 is a universal one, as indicated by early results (Bernardi and Bernardi 1985, 1986) and by recent work (Bernardi and Bernardi 1991; and unpublished).

Table 2. Compositional properties of extreme genes

Genes (mnemonics)	GC%				
	Exons	I	II	III	I + II
CFANT	46.32	42.11	27.37	68.42	34.74
MRP8	46.45	47.87	23.40	67.02	35.64
FABPLA	47.14	43.75	28.91	68.75	36.33
IFNB1F	44.68	43.62	29.26	60.64	36.44
FP	44.20	60.27	17.41	54.46	38.84
CAATP1	51.85	53.33	24.44	77.78	38.89
MAS	48.16	42.02	36.81	65.64	39.42
CYCAA	44.34	44.34	34.91	53.77	39.62
ARM	43.73	47.14	30.95	52.86	39.85
CRBP1	53.68	55.15	27.94	77.21	41.54
TNCS	54.53	59.26	24.07	79.63	41.67
XEH	54.75	51.32	36.62	76.10	43.97
TC2	57.76	59.01	29.81	84.47	44.41
OPS	57.78	49.86	39.26	84.24	44.56
GNAZ	58.52	52.81	37.36	85.11	45.08
MGPH	58.53	58.33	35.19	81.94	46.76
IDB	62.56	59.59	34.93	93.15	47.26
DLDH	41.37	53.33	41.37	29.22	47.35
LPDH	41.44	53.53	41.57	29.02	47.55
ACADAM	40.48	52.84	42.42	27.25	47.63
SAACY	61.64	54.76	40.74	89.15	47.75
ETFA	45.21	57.19	43.71	34.43	50.45
CKB	64.31	64.14	37.70	90.84	50.92
CREB	47.15	58.23	45.43	37.80	51.83
HMG14A	50.83	56.44	47.52	48.51	51.94
RNPB1A	44.92	54.52	49.72	30.23	52.12
RNPC2A	45.32	55.26	50.29	30.12	52.78
HBA4	64.80	63.64	41.96	88.81	52.80
SNEXIN	49.46	59.10	49.89	39.19	54.50
H33G2	51.46	61.03	49.26	44.12	55.15
GAP43A	53.97	66.53	49.37	45.61	57.95
YB1A	56.18	64.47	52.52	51.26	58.49
HMG17G	51.28	61.54	48.35	42.86	59.45
MEA	57.71	73.66	45.70	53.23	59.68
GFB	61.45	63.99	56.40	63.51	60.19
IGHBP1	61.11	66.67	55.56	58.33	61.11
A1ATR3	56.22	64.18	58.21	44.78	61.19
SPR2A	60.27	60.27	63.01	57.53	61.64
SNRAA	56.98	66.80	61.00	42.74	63.90
C4A2	63.11	72.45	62.11	54.70	67.28

Extreme genes are defined here as the genes located outside the broken line of Fig. 2. Genes are listed in order of increasing GC levels in first + second codon positions.

Codon Usage

Several explanations were provided for given codons being preferred to other synonymous codons.

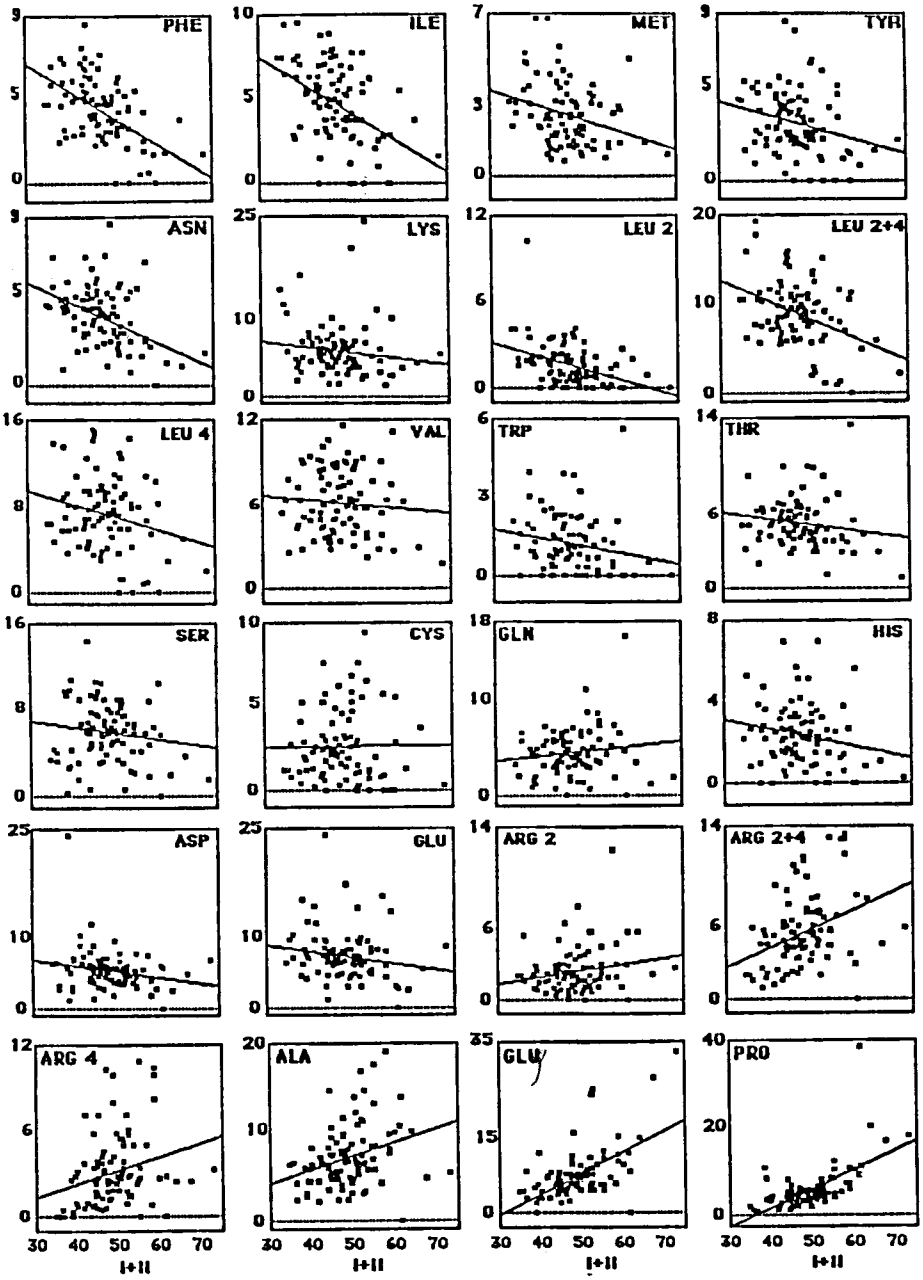

Fig. 4. Plot of relative amounts (moles percent) of individual amino acids of proteins against GC levels in first + second codon positions of the corresponding genes

These explanations (see Grantham 1980; Bernardi and Bernardi 1986) include (1) the optimization of codon–anticodon interaction energy and the consequent optimization of translation efficiency in highly expressed genes; (2) the fulfillment of requirements for mRNA secondary structure and stability; (3) the adaptation of codons to the actual populations of isoaccepting tRNAs; (4) the regulation of replication or transcription by RNA secondary structure; and (5) a metabolic discrimination between nucleotides at the time of DNA replication. Although some of these explanations may be correct, they are clearly incomplete, as they do not account, for instance, for the different codon usages found in different species of living organisms, nor for the different codon usages within a single organism (see below).

An important step toward a better understanding of codon usage was made when it was realized that codon usage is a strategy associated with a given genome or genome type (Grantham 1980; Grantham et al. 1980). This genome hypothesis subsequently underwent two major conceptual changes as a consequence of later investigations (Bernardi and Bernardi 1985, 1986; Bernardi et al. 1985): (1) Codon usage, as studied through the GC levels of third codon positions, is not simply related to the genome of a species, but to the base composition of the genome. (2) Multiple codon usages are found in compositionally compartmentalized genomes, like those of most if not all eukaryotes, including warm-blooded vertebrates. Obviously, these findings further stress the correlation between codon usage and

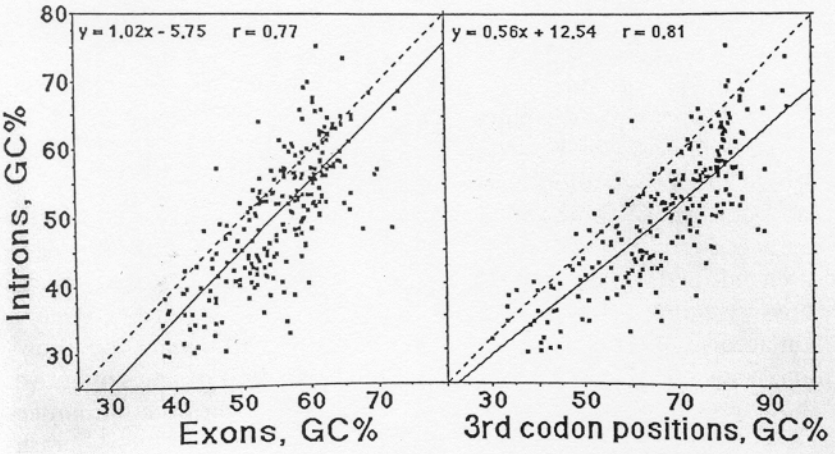

Fig. 5. Plot of GC levels of introns against GC levels (**A**) of exons and (**B**) of third codon positions from the same genes; 238 genes available for this purpose were used

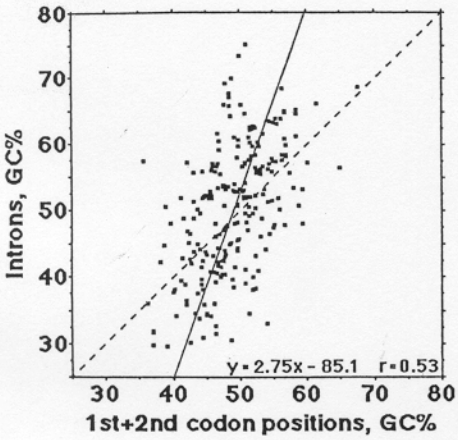

Fig. 6. Plot of GC levels of introns against GC levels of first + second codon positions from the same genes. The orthogonal regression line is shown. Data concern the gene sample of Fig. 5.

DNA composition of the genome, or genome compartments.

Figures 1 and 2 confirm that codon usage can vary greatly within a single genome, when this genome is compositionally compartmentalized. This conclusion, already drawn from previous investigations (Grantham 1980; Bernardi and Bernardi 1985, 1986; Bernardi et al. 1985; Bernardi 1989), is reinforced by the analysis of the much larger gene sample used in the present work. That differences in codon usage may be extremely striking within a single genome is stressed by the fact that, at 100% GC in third codon position, a value that is closely approached by a number of genes, 50% of the codons simply cannot be used because they are absent.

Amino Acid Composition of Proteins

The existence of a correlation between third and first + second codon positions (Figs. 1 and 2) leads to the following considerations. First and second codon positions are associated with specific amino acids. Therefore, as third codon positions increase or decrease in GC levels, amino acid compositions tend to change toward amino acids corresponding to codons characterized by higher or lower GC levels in second and first codon positions. Such differences concern the relative abundance of alanine, arginine, glycine, and proline, on the one hand, and of asparagine, isoleucine, leucine, lysine, methionine, phenylalanine, and tyrosine, on the other hand, in the proteins encoded by genes that are GC-poor or GC-rich in their first + second codon positions. The analysis shown in Fig. 4 was extended to all genes available and found in the extended sample as well. There is, therefore, no doubt about the existence of particular trends in amino acid compositions of proteins encoded by genes characterized by different GC contents.

This conclusion has an important implication. Because GC levels of coding sequences and of their different codon positions are correlated with those of the isochores in which they are located (Bernardi et al. 1985; Bernardi 1989; Aïssani et al. 1991), the amino acid composition of proteins is also correlated with genome (isochore) localization (and even the localization in Giemsa-positive, Giemsa-negative, or telomeric chromosomal bands) of the corresponding genes. In turn, this implies variations of external (lysine and arginine), internal (phenylalanine, leucine, isoleucine, methionine), and ambivalent amino acids (proline, alanine, glycine), which are relevant factors in determining protein structure. This raises an interesting point—that of the existence of some extent of correlation between functional properties (in the broad sense) of proteins and the isochore localization of the corresponding genes.

Interestingly, the correlation between amino acid composition and GC levels of genomes was first found by Sueoka (1961) in his pioneering study on these properties in 11 bacterial species. This work, performed before the genetic code was elucidated, showed that, out of 18 amino acids tested, alanine, arginine, glycine, and proline are positively correlated with increasing GC content; isoleucine, lysine,

aspartic acid plus asparagine, glutamic acid plus glutamine, tyrosine, and phenylalanine are negatively correlated; and histidine, valine, leucine, threonine, serine, and possibly methionine show no detectable evidence of correlation. Although this is in overall agreement with the present findings, some differences also exist: in fact, both glutamic acid and glutamine are not correlated, asparagine is negatively correlated, but aspartic acid is not correlated, and methionine is negatively correlated. Moreover, both arginine and leucine are correlated or not according to the codons used; finally, the two amino acids not analyzed, tryptophan and cysteine, are both not correlated.

Compositional Correlation between Introns and Exons

This relationship, first detected when studying a small number of genes from different vertebrates (Bernardi et al. 1985), was then shown to exhibit a high correlation coefficient (0.76) using a set of 56 human genes (Bulmer 1987). The present results on 238 genes show (Fig. 5) that the correlation is not only characterized by a high correlation coefficient ($R = 0.77$), in full agreement with the previous report, but also by a unity slope ($s = 1.02$), and by the fact that, on the average, intron points lie on a line about 5% lower than the unity slope line through the origin; this indicates a slightly lower GC level for introns relative to exons from the same genes. A plot of intron GC levels against third codon position GC levels exhibits a linear correlation with a slope of 0.56 and a correlation coefficient of 0.81. More interestingly, a linear correlation with a coefficient of 0.53 and a slope of 2.71 (as judged by the orthogonal regression approach) is exhibited when intron GC levels are plotted against GC levels of first + second codon positions (Fig. 6). Likewise, linear relationships between GC levels of intergenic sequences [which basically correspond to the compositional fractions investigated in Aïssani et al. (1991)] or 5′ flanking sequences (Aïssani et al. 1991) and GC levels of first + second codon positions were also found, with correlation coefficients of 0.61 and 0.54, respectively.

Comparison of Data from the Present Work with Data for Localized Genes

Two points raised by the present results will be discussed elsewhere. The first point is that a comparison of the data obtained on the human genes investigated in the preceding paper (Aïssani et al. 1991) and in the present one indicate that the genes studied previously are representative of human genes in general. Indeed, the compositional correlations between third and first + second codon positions, as well as that between introns and exons are essentially the same for the set of human genes studied in the preceding paper (Aïssani et al. 1991) and for the 1400 or so genes investigated here. (Needless to say, in such comparisons, the same approach was used in calculating the slopes of the third versus first + second codon positions.) This conclusion allows reconstruction of the distribution of genes of known coding sequence in the isochores of the human genome and even in the Giemsa-positive, Giemsa-negative, or telomeric chromosomal bands (unpublished).

Acknowledgments. We thank the Association Française contre les Myopathies (AFM) for a Fellowship to B.A., FEBS and the Di Capua Foundation, Naples, Italy, for Fellowships to G.D., and the Association pour la Recherche contre le Cancer (ARC) and the Ministry for Research and Technology (MRT) and AFM for financial support.

References

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

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

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

Bernardi G, Bernardi G (1986) Compositional 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 (in press)

Bernardi G, Olofsson 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

Bulmer M (1987) A statistical analysis of nucleotide sequences of introns and exons in human genes. Mol Biol Evol 4:395–405

Gouy M, Gautier C, Attimonelli N, Lanave C, Di Paola G (1985) ACNUC—portable retrieval system for nucleic acid sequence database: logical and physical design and usage. Cabios 1: 167–172

Grantham R (1980) Workings of the genetic code. Trends Biochem Sci 5:327–333

Grantham R, Gautier C, Gouy M, Mercier R, Paré A (1980) Codon catalogue usage and the genome hypothesis. Nucleic Acids Res 8:r49–r62

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

Received December 12, 1990/Revised February 2, 1991