Volume 17 Number 13 1989 Nucleic Acids Research

The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants

Giorgio Matassi, Luis M.Montero[1], Julio Salinas[1] and Giorgio Bernardi*

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France and [1]Departamento de Proteccion Vegetal, Instituto Nacional de Investigaciones Agrarias, Carretera de la Coruna, Km. 7, 28040 Madrid, Spain

Received March 28, 1989; Revised and Accepted June 2, 1989

ABSTRACT

The isochore structure of the nuclear genome of angiosperms described by Salinas et al. (1) was confirmed by using a different experimental approach, namely by showing that the levels of coding sequences from both dicots and Gramineae are linearly correlated with GC levels of the corresponding flanking sequences. The compositional distribution of homologous coding sequences from several orders of dicots and from Gramineae were also studied and shown to mimick the compositional distributions previously seen (1)for coding sequences in general, most coding sequences from Gramineae being much higher than those of the dicots explored. These differences were even stronger for third codon positions and led to striking codon usages for many coding sequences especially in the case of Gramineae.

INTRODUCTION

Recent investigations on plant genomes (1) have led to two major conclusions : (i) that the nuclear genomes of angiosperms are mosaics of long (>100-200 Kb), compositionally homogeneous DNA segments, which were called isochores (2), and that belong to families characterized by different GC levels; and (ii) that the nuclear genomes of Gramineae exhibit strikingly different isochore patterns compared to those of the dicots investigated.

Differences in isochore patterns were seen at two different levels. (i) When DNA fragments (in the 50-100 Kb range) fromthree dicots (pea, sunflower and tobacco) were compared with those from three Gramineae (maize, rice and wheat), the compositional distribution of the fragments was centered around 41% GC for the former and around 45% GC for the latter. Moreover, in the case of maize and wheat, distributions trailed towards even higher GC values. (ii) Coding sequences from

several orders of dicots showed a narrow, symmetrical compositional distribution centered around 46% GC. In contrast, coding sequences from barley, maize and wheat showed a broad, asymmetrical distribution characterized by an upward trend towards high GC values, the majority of sequences being comprised between 60 and 70% GC. Similar, yet more striking, differences were found when GC levels of third codon positions were compared. Finally, introns exhibited compositional distributions that mimicked those of exons from the same genes, but were lower in GC levels, as were the intergenic non-coding sequences that form the vast majority of plant DNAs. The latter point is indicated by the lower GC levels of DNA fragments compared to coding sequences from the same genomes.

To sum up, the genome organization of angiosperms is very reminiscent of that previously described for vertebrates (2), with a striking and puzzling resemblance in compositional patterns between the dicots investigated and cold-blooded vertebrates, on the one hand, and of Gramineae and warm-blooded vertebrates, on the other. In the present work, we have investigated the compositional distributions of DNA fragments in the 50-100 Kb size range for two additional dicots and six additional monocots. We have then compared the GC levels of homologous coding sequences. (and of their different codon positions) from several dicots and from Gramineae, in order to better define the differences in compositional patterns between the genomes of these plants, and the consequences of such differences on codon usage. Finally, we have investigated the relationships of GC levels of coding sequences and introns with those of the corresponding flanking sequences.

MATERIALS AND METHODS

Preparation and fractionation of nuclear DNAs. Ethiolated seedlings from Hordeum vulgare(barley), Secale cereale(rye), and Asparagus officinalis, and mature leaves from Antirrhinum majus, Oenothera hookeri, Scindapsus aureus, Typha latifolia, and Allium cepa(onion), were used to prepare nuclear DNA as described elsewhere (1). Details on the characterization of the

DNAs and their fractions will be presented in a forthcoming paper.

GC levels of coding sequences (from the initial AUG to the terminal codon), of first, second and third codon positions, and of introns and flanking sequences were obtained from GenBank (Release 57, September 1988). Data concerning genes not yet available in GenBank were obtained from the literature. A reference list for the coding sequences from the literature and a list of the genes studied in both exons and introns will be provided upon request. The ACNUC retrieval system (3) was used.

Codon usage was investigated by considering the ratio of observed codon frequency over codon frequency expected for a statistical distribution of codons (this ratio has been called relative synonymous codon usage, or RSCU; see ref. 4).

The homologous genes investigated, the plants from which they derived, and the plant families were the following : **Actin** : Glycine max (Leguminosae, 2 genes); Zea mays(Gramineae).**Alcohol dehydrogenase(Adh)** : Arabidopsis thaliana (Cruciferae); Pisum sativum (Leguminosae). **Chlorophyll-a/b-binding protein(Cab)** : A. thaliana (3 genes); Cucurbita sp. (Cucurbitaceae; 2 genes); Lycopersicon esculentum (Solanaceae; 2 genes); Petunia hybrida (Solanaceae; 6 genes); P. sativum; Silene pratensis (Caryophillaceae); Lemna gibba (Lemnaceae); Triticum aestivum (Gramineae); Z. mays.**Chalcone synthase(Chs)** : Antirrhinum majus (Scrofulariaceae); Magnolia liliiflora (Magnoliaceae); Petroselium hortense (Umbelliferae); P. hybrida; Phaseolus vulgaris (Leguminosae); Ranunculus acer (Ranuncolaceae); Hordeum vulgare (Gramineae); Z. mays. **Glyceraldehyde-3-phosphate-dehydrogenase(GADPH)** : Sinapis alba (Cruciferae, 2 genes); Nicotiana tabacum (Solanaceae, 3 genes); Z. mays.**Histone H3(H3)** : A. thaliana; Oryza sativa (Gramineae); T. aestivum; Z. mays (2 genes). **Histone H4 (H4)** : A. thaliana; T. aestivum; Z. mays (2 genes).**Phytochrome(Phyt.)** : Cucurbita pepo; Avena sativa (Gramineae, 2 genes). **Ribulose-1,5-bisphosphate carboxylase, small subunit (Rubisco)** : Cucurbita sp.; G. max (2 genes); Helianthus annuus (Compositae); L. gibba; L. esculentum (3 genes); N. tabacum; N. sylvestris

(Solanaceae); O. sativa; P. hybrida (3 genes); P. sativum (4 genes); Raphanus sativus (Cruciferae); S. pratensis; T. aestivum; Z. mays.

RESULTS

Compositional distribution of DNA fragments

Fig. 1shows histograms providing the buoyant densities in CsCl and the relative amounts of compositional fractions as obtained from Cs_2SO_4/BAMD fractionation of nuclear DNAs from five dicots and nine monocots (BAMD is 3,6 bis (acetatomercurimethyl) dioxane).

As far as dicots are concerned, two additional DNAs (from Antirrhinum majus and Oenothera hookeri, respectively) were

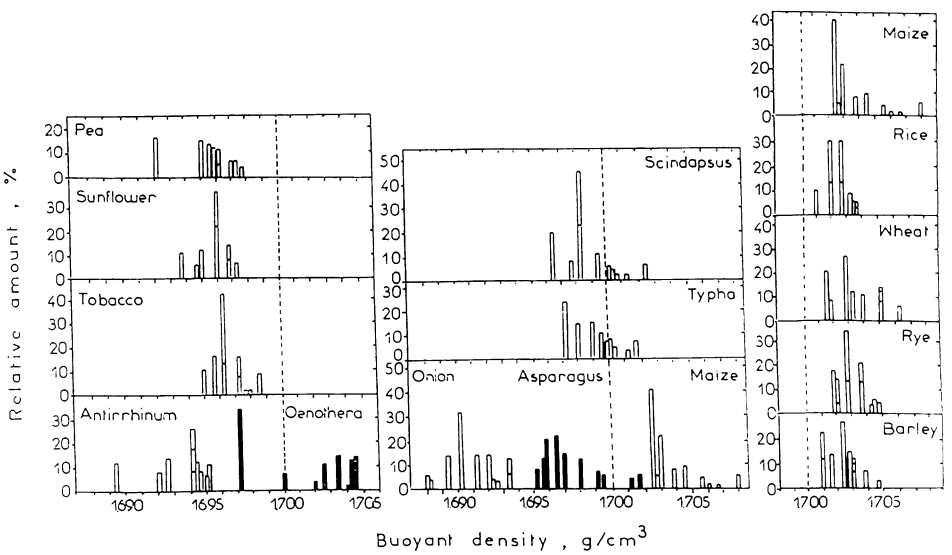

Fig. 1. Histograms showing the relative amounts and buoyant densities in CsCl of DNA fractions obtained by preparative Cs SO /BAMD density gradient centrifugation from dicots (left panel) and monocots (middle and right panels). Data for pea, sunflower, tobacco, maize, rice and wheat are from ref.1 (where additional details can be found). The vertical broken line at 1.700 g/cm3 is shown to provide a reference. Black bars are used whenever required for discriminating fractions from different DNAs. Horizontal lines on some bars separate DNA fractions showing the same buoyant densities.

studied compared to our previous investigations (1). These DNAs were chosen because literature data (5,6) indicated that they corresponded to the lowest and highest buoyant densities of all dicot DNAs investigated so far.

The new monocot DNAs studied were from two additional Gramineae, rye and barley, and from four other monocots, Scindapsus aureus, Typha latifolia, Allium cepa (onion) and Asparagus officinalis. Rye and barley were chosen because a number of their coding sequences are known in primary structure. The choice of the other four monocots was made in order to cover a larger spectrum of families (Scindapsus belongs to Araceae,

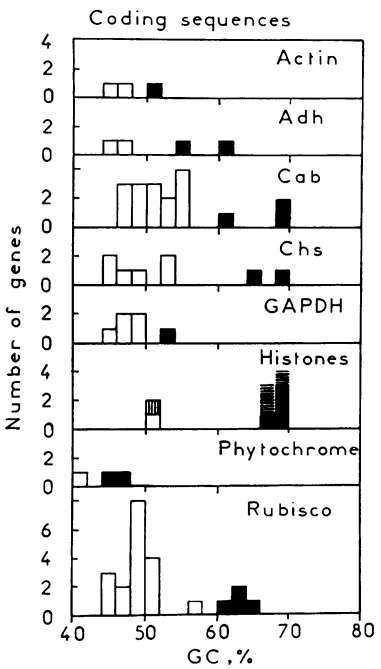

Fig. 2. The numbers of homologous genes are plotted against the GC levels of the corresponding coding sequences; a 2% GC window was used. Values for dicots and monocots (Gramineae, except for two genes, cab and rubisco, from L. gibba) are represented by the open and black bar histograms, respectively. Genes for histones H3 and H4 are presented in the same histogram. H3 genes from Gramineae are indicated by horizontally hatched bars, their homologs from dicots (Arabidopsis) by vertically hatched bars. See Materials and Methods.

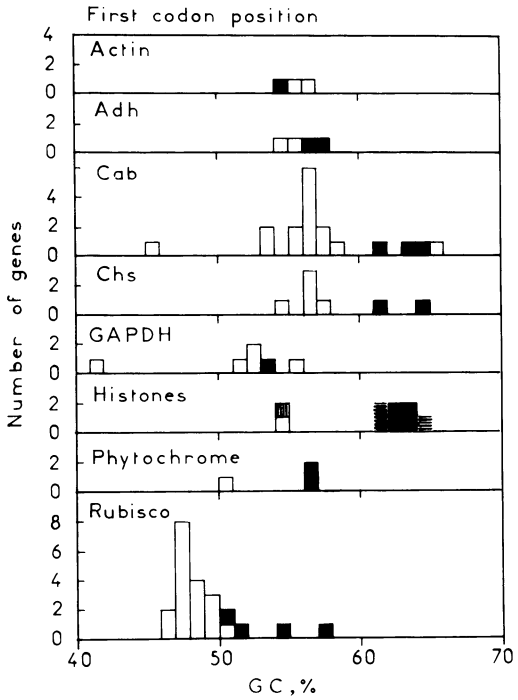

Fig. 3.The numbers of homologous genes of Fig. 2 are plotted against the GC levels of the first codon position of the corresponding coding sequences. The highest value among cab genes belongs to Silene pratensis. Other indications as in Fig. 2.

Typha to Typhaceae, Asparagus to Asparagaceae and Allium to Alliaceae) compared to those previously explored, and to analyze a DNA (that of Allium) characterized by one of the lowest modal buoyant densities among plant DNAs (5).

The data of Fig. 1 stress the fact that DNAs from both dicots and monocots cover a wide range of modal buoyant densities (namely, of the buoyant densities at the peak of analytical CsCl profiles). In the case of the dicots studied here, the limits of the range are 1.6942 (Antirrhinum majus) and 1.7035 g/cm3 (Oenothera hookeri). In the case of monocots, the range of modal buoyant densities is wider, from 1.691 g/cm3 for Allium cepa to 1.7026 g/cm3 for T. aestivum; all values for Gramineae are in a higher narrow range, 1.7107 to 1.7026 g/cm3.

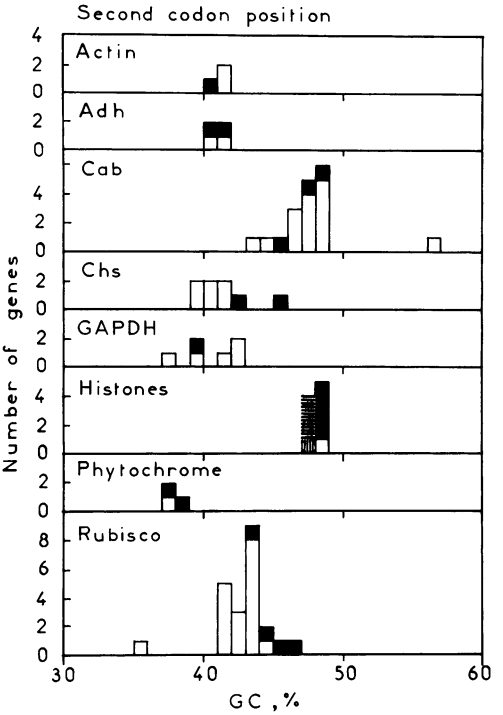

Fig. 4. The numbers of homologous genes of Fig. 2 are plotted against the GC levels of the second codon position of the corresponding coding sequences. The highest value among cab genes belongs to Cucumis sativus. Other indications as in Fig. 2.

Compositional distribution of homologous coding sequences from dicots and Gramineae

Fig. 2 displays the compositional distribution for all available homologous coding sequences from dicots and monocots. As indicated in the Materials and Methods section, the nine sets of homologous genes investigated were from several dicots, belonging to ten different families, but from only five monocots, four of which belonged to a single family , Gramineae (only two genes, cab and rubisco, were from another monocot, Lemna gibba). GC levels of coding sequences were higher in monocots than in dicots for all nine genes tested. When comparisons were made for first, second and third codon positions (Figs. 3-5), the differences found in coding sequences

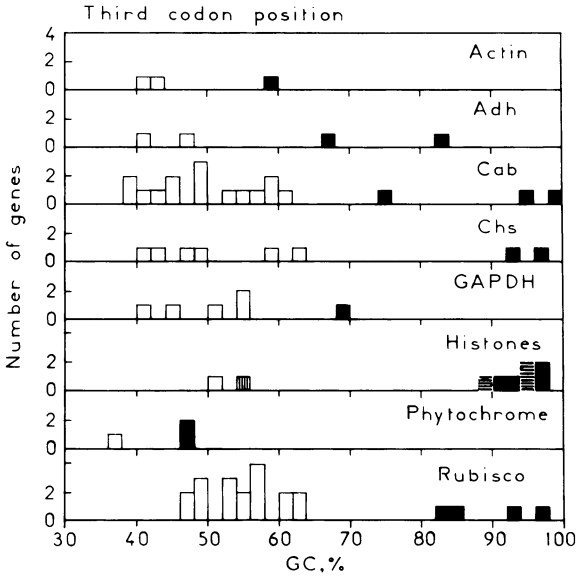

Fig. 5. The numbers of homologous genes of Fig. 2 are plotted against the GC levels of the third codon position of the corresponding coding sequences. Other indications as in Fig. 2.

became much larger when third codon positions were compared, whereas they became smaller for first codon positions, and almost disappeared for second codon positions.

Codon usage

The codon usage of all genes shown in Fig. 2 was analysed. Table 1 displays an example of this analysis for nine pairs of homologous genes from dicots and monocots, respectively. In the case of monocots, Zea mays was presented systematically (except for the phytochrome gene which is not available), in order to provide information on codon usage patterns within a single species. These results will be commented upon in the Discussion.

Comparison of GC levels of exons and introns with GC levels of flanking sequences

The GC levels of coding sequences were compared with those of the corresponding flanking sequences. In this case, the gene sample used was largely different from that of homologous genes and comprised all coding sequences (from the GenBank and the literature) for which pooled 5'and 3' flanking sequences were longer than 1 Kb. A straight-line relationship with a slope of

		As Phyt.	Cp Phyt.	Zm Act1	Gm Act1	Zm Adh1	Ps Adh1	Zm GAPDH	Sa GAPDH	Zm H3	At H3	Zm H4	At H4	Zm Cab	Ph Cab	Zm Rub.	Gm Rub.	Zm Chs	Ml Chs
Phe	TTT	1.14	1.24	0.73	0.91	0.50	0.72	0.43	0.57		0.40	1.00	1.00		1.06		0.45		0.77
	TTC	0.86	0.76	1.27	1.09	1.50	1.28	1.57	1.43	2.00	1.60	1.00	1.00	2.00	0.94	2.00	1.55	2.00	1.23
Leu	TTA	0.34	1.38		0.20	0.30	0.50								0.57				0.49
	TTG	1.11	2.05	1.15	1.24	0.30	1.00	0.30	1.43		1.50		1.50	0.23	1.99		1.72		0.49
	CTT	1.60	1.14	1.39	1.86	1.80	2.75	1.50	1.72		1.99		0.75	0.46	1.99		2.14		1.30
	CTC	0.82	0.29	1.85	0.83	0.90	1.00	3.00	2.87	4.00	1.00	4.50	3.00	3.23	1.14	1.85	0.86	3.08	1.78
	CTA	0.92	0.62	0.46	0.20				0.29		0.50				0.29		0.43	0.16	0.65
	CTG	1.21	0.52	1.15	1.66	2.70	0.75	1.20	0.29	2.00	1.00	1.50	0.75	2.08		4.15	0.86	2.75	1.30
Ile	ATT	1.07	1.48	1.50	2.04	0.73	0.99	0.82	0.68	0.43	1.29		0.86	0.38	1.33		1.00		0.30
	ATC	1.34	0.87	1.37	0.72	2.20	1.50	2.18	2.32	2.57	1.71	3.00	2.14	2.62	1.33	2.50	2.00	3.00	2.40
	ATA	0.59	0.65	0.13	0.24	0.13	0.50								0.33	0.50			0.30
Val	GTT	1.40	1.43	1.42	2.06	0.92	1.94	1.67	2.16		2.00		1.00	0.24	1.80		1.46		0.93
	GTC	0.77	0.64	1.55	0.52	0.92	0.11	1.89	1.89	2.67	1.33	2.00	2.00	2.44	1.00	2.18	0.36	1.89	1.20
	GTA	0.72	0.84	0.39	0.39	0.31	1.89								0.40		0.36		0.40
	GTG	1.11	1.09	0.64	1.03	1.85	0.76	0.44	0.65	1.33	0.67	2.00	1.00	1.56	0.80	1.82	1.82	2.81	1.47
Ser	TCT	1.28	1.38	0.64	1.28	0.35	1.14	0.50	1.36			3.00			2.21		1.00		1.11
	TCC	1.21	0.48	1.28	0.86	0.35	1.43	1.75	1.09	2.40	1.20			2.33	0.95	2.25	2.50	3.14	1.11
	TCA	0.94	1.58	0.43	1.50	1.06	1.14	0.25	0.82		1.20				0.63		1.00		0.67
	TCG	0.40	0.41	0.64		1.76	0.29	1.00	0.55	2.40		3.00	3.00	1.00	0.32	1.50		1.72	1.11
Pro	CCT	1.56	2.24	1.90	1.90	1.20	1.20	1.33	1.09	0.67	1.33		4.00		1.33		0.73	0.21	1.26
	CCC	0.52	0.49	0.57	0.21	0.60	0.40	1.33	1.09	2.67				1.05	0.44	1.67	1.09	1.68	0.63
	CCA	1.65	0.98	1.14	1.47	0.80	1.60	0.67	1.82		1.33			0.42	2.22		2.18	0.21	1.47
	CCG	0.26	0.29	0.38	0.42	1.40	0.80	0.67		0.67	1.33	4.00		2.53		2.33		1.90	0.63
Thr	ACT	1.30	1.72	1.09	0.95	1.80	1.40	1.09	1.64		1.20		2.28		2.40		2.00		0.92
	ACC	1.02	0.48	1.27	1.33	1.80	1.00	2.00	1.46	2.40	2.00	4.00	1.71	1.78	0.80	2.67	2.00	2.77	1.23
	ACA	1.40	1.52	0.73	1.33	0.20	0.80	0.73	0.54		0.80				0.80				0.77
	ACG	0.28	0.28	0.91	0.38	0.20	0.80	0.18	0.36	1.60				2.22		1.33		1.23	1.08
Ala	GCT	1.75	1.84	1.67	1.52	1.41	2.34	1.93	2.52	0.40	1.20	0.57	2.28	0.24	1.75	0.33	2.15		0.57
	GCC	0.75	0.58	1.67	0.96	0.94	0.28	1.79	1.04	1.60	0.60	1.14	1.14	2.70	1.62	3.33	0.92	2.04	1.60
	GCA	1.21	1.16	0.17	1.52	0.82	1.10	0.28	0.44	0.20	1.00				0.50		0.92		0.92
	GCG	0.29	0.42	0.50		0.82	0.28			1.80	1.20	2.28	0.57	1.18	0.12	0.33		1.96	0.92
Tyr	TAT	1.24	1.52	0.67	1.29	0.40	0.33	0.67	0.50		1.00				0.57				0.67
	TAC	0.76	0.48	1.33	0.71	1.60	1.67	1.33	1.50	2.00	1.00	2.00	2.00	2.00	1.43	2.00	2.00	2.00	1.33
His	CAT	1.29	1.52	0.80	1.56	0.91	1.33	0.75	1.00		1.00				0.67				0.40
	CAC	0.71	0.48	1.20	0.44	1.09	0.67	1.25	1.00	2.00	1.00	2.00	2.00	2.00	1.33	2.00	2.00	2.00	1.60
Gln	CAA	0.73	1.29	0.44	1.09	0.40	0.57				0.75		2.00		2.00		1.14		0.91
	CAG	1.27	0.71	1.56	0.91	1.60	1.43	2.00	2.00	2.00	1.25	2.00		2.00		2.00	0.86	2.00	1.09
Asn	AAT	0.93	1.50	0.22	0.57	0.44	1.07	0.46	0.40		2.00				0.91		0.29		1.09
	AAC	1.07	0.50	1.78	1.43	1.56	0.93	1.54	1.60	2.00		2.00	2.00	2.00	1.09	2.00	1.71	2.00	0.91
Lys	AAA	0.50	0.86	0.21	0.42	0.38	0.92	0.29	0.32		0.71		0.20		0.29	0.20		0.10	0.38
	AAG	1.50	1.14	1.79	1.58	1.62	1.08	1.71	1.68	2.00	1.29	2.00	1.80	2.00	1.71	1.80	2.00	1.90	1.62
Asp	GAT	1.29	1.37	1.39	1.36	0.77	1.60	0.54	0.88	0.50	1.50		1.33		0.91		1.14		0.83
	GAC	0.71	0.63	0.61	0.64	1.23	0.40	1.46	1.12	1.50	0.50	2.00	0.67	2.00	1.09	2.00	0.86	2.00	1.17
Glu	GAA	0.99	1.14	0.57	0.64	0.47	0.93	0.27	0.24		1.14			0.13	0.57		0.67		0.45
	GAG	1.01	0.86	1.43	1.36	1.53	1.07	1.73	1.76	2.00	0.86	2.00	2.00	1.87	1.43	2.00	1.33	2.00	1.55
Cys	TGT	1.13	1.40	0.80	1.33	0.62	1.64												1.00
	TGC	0.87	0.60	1.20	0.67	1.38	0.46	2.00	2.00	2.00	2.00			2.00	2.00	2.00	2.00	2.00	1.00
Arg	CGT	0.68	1.11	1.33	4.24		0.92	0.55	1.50	0.35	1.76		2.40		2.57		1.99	0.26	2.63
	CGC	0.11	0.22	0.34		2.57	0.46	1.09		4.59	0.35	4.00		5.25		4.00	1.33	3.65	
	CGA	0.56	0.89	0.34			0.92						0.40						1.13
	CGG	0.91	0.34	0.34	0.35	0.86				0.35		0.80		0.75	0.86	0.67		1.30	0.75
Ser	AGT	1.10	1.38	0.64	0.43	0.35	0.29		1.09				3.00		0.63				0.44
	AGC	1.10	0.76	2.36	1.93	2.12	1.72	2.50	1.09	1.20	3.60			2.66	1.27	2.25	1.50	1.14	1.55
Arg	AGA	1.58	1.99	1.99	0.71	0.43	2.77	1.09	2.50		1.76		1.20		1.72	0.67	0.67		0.75
	AGG	2.15	1.45	1.67	0.71	2.14	0.92	3.27	1.99	0.71	2.12	1.20	1.99		0.86	0.67	1.99	0.78	0.75
Gly	GGT	1.39	1.01	1.60	1.62	1.62	1.84	1.87	1.86	0.57		0.47	1.65	0.12	1.88		1.14		1.29
	GGC	0.64	0.64	1.07	0.50	0.97	0.32	1.47	0.28	2.86	0.57	2.59	0.24	3.12	0.38	3.27	1.43	2.86	1.16
	GGA	1.22	1.39	0.53	1.50	0.54	0.86	0.40	0.26		3.43	0.24	2.12	0.12	1.50	0.36	1.43	0.34	0.78
	GGG	0.75	0.96	0.80	0.38	0.86	0.97	0.27		0.57		0.70		0.62	0.25	0.36		0.80	0.78
G+C% c.s.		47.1	41.9	51.4	47.0	54.9	45.9	54.0	48.6	68.6	51.1	61.5	51.9	69.4	50.1	64.6	50.8	69.3	53.7
G+C% 3rd		46.8	37.7	59.6	42.9	66.6	41.7	68.7	56.0	93.4	50.4	94.2	53.9	95.6	45.7	97.0	61.4	97.7	62.1

Table 1. Codon usage exhibited by nine pairs of homologous genes from monocots and dicots, respectively. Values are ratios of observed codon frequency over codon frequency expected for a statistical distribution of codons (RSCU; ref. 4). Met, Trp and termination codons were not taken into account. Dashes stand for absent codons. Gene pairs were ordered from left to right according to increasing GC content in third codon position of genes from monocots (A. sativa and Z. mays). The two bottom lines present GC levels of coding sequences and third codon positions.

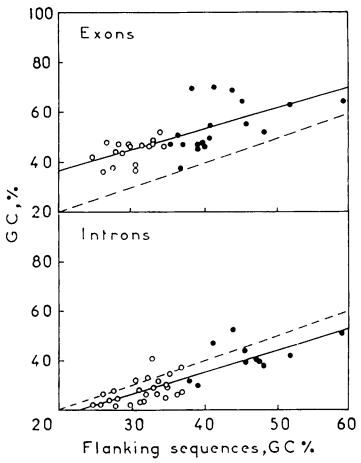

Fig. 6. GC levels of exons (upper frame) or introns (lower frame) are plotted against the GC levels of the corresponding flanking sequences. Open and closed points correspond to dicots and monocots (Gramineae except for one point from L. gibba, respectively. The sequence list will be provided upon request.

0.80 and a correlation coefficient of 0.72 was obtained by using the least square procedure (Fig. 6). Points for genes from dicots corresponded to lower GC values and were less scattered compared to those of Gramineae, (except for one point corresponding to Lemna). When all available 5' and 3' flanking sequences (regardless of size) were plotted separately, results were similar. Slopes were 0.77 (5') and 0.73 (3') with correlation coefficients of 0.73 and 0.60, respectively (not shown). Fig.6 also indicates that the flanking sequences of Gramineae are characterized by higher GC levels than those of dicots.

When GC levels of introns from the same genes were plotted against GC levels of flanking sequences, points were less scattered and fell on a straight line with a slope of 0.86 and a correlation coefficient of 0.82. The straight line of introns was displaced towards the bottom of the diagram compared to that of coding sequences (Fig. 6).

If GC levels of individual codon positions of the same genes were plotted against GC levels of corresponding flanking sequences, points showed a larger scatter compared to the plot

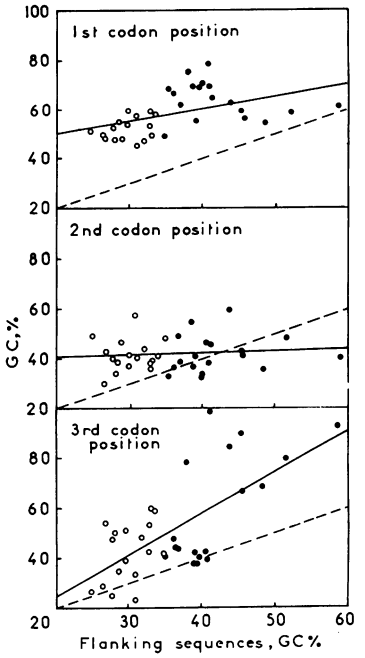

Fig. 7. GC levels of codon positions from the genes of Fig. 6 are plotted against the GC levels of the corresponding flanking sequences. The highest GC value in third codon position is the cab gene from L. gibba. See legend of Fig. 6 for other indications.

of coding sequences (Fig. 7). Slopes ranged from 0.1 for the second, to 0.52 for the first and to 1.78 for the third codon position respectively (Fig. 7) and correlation coefficients were 0.11, 0.47 and 0.70, respectively.

DISCUSSION

Compositional distributions of DNA fragments from dicots and monocots

The range of modal buoyant densities of DNAs from dicots deserves several comments. (i) The range of all dicots investigated so far is narrower, 1.694-1.697 g/cm3, than that of monocots, with the only exception of Oenothera. Needless to say, exploration of DNAs from other dicots may reveal that the range is broader than that seen at present. (ii) The lowest value of dicot DNAs found here, that of Antirrhinum, is remarkably higher

than previously reported, 1.691 g/cm3 (5). Antirrhinum DNA exhibits, however, a remarkable heterogeneity with a very GC-poor component (possibly a satellite) as low as 1.6895 g/cm3 in buoyant density. Unfortunately, no data are available on the methylation of this DNA, a fact preventing any precise assessment of its GC content. Since methylation leads to a decrease in buoyant density of about 0.7 mg/cm3 per 1% 5-methyl cytosine (1), the GC content of Antirrhinum DNA (which would be about 34% in the absence of methylation) can only be underestimated if calculated from its buoyant density. (iii) The upper limit of modal buoyant densities of dicot DNAs is that of Oenothera. Our value for modal buoyant density, 1.7035 g/cm3, is in good agreement with previous ones for DNAs from several Oenothera species, 1.703 g/cm3 (5). Our Cs_2SO_4 analysis (Fig. 1) has shown, however, the existence of a remarkably lighter peak, 1.6972 g/cm3, representing as much as 30% of total DNA. While the latter is likely to correspond to a satellite, this needs to be established. Again, since no information is available on the methylation of Oenothera DNA, no estimate of its GC content can be given. In conclusion, the GC levels of the DNAs from dicots investigated so far range from a minimal value of 34% to a value of 43% (if the 1.703 g/cm3 peak is indeed the main band) or higher if Oenothera DNA is methylated.

In the case of monocot DNAs, two points should be made. (i) Since methylation data are available for both onion and wheat DNA (6), the GC range of the DNAs studied can be estimated as comprised between 38 and 48% (1). (ii) An interesting observation concerns the difference in the compositional patterns exhibited by Allium cepa and Asparagus officinalis, two species belonging to the same order, Liliales. The corresponding DNAs differ by as much as 5.3 mg/cm3 in modal buoyant density, and buoyant density distributions show no overlap. This situation stresses the possible lack of correlation between taxonomy and genome compositional patterns and is similar to that found for a number of fish orders and families (G.Bernardi and G.Bernardi, paper in preparation; see also ref. 8).

The expanded set of plant genomes analyzed indicates, in agreement with previous literature data (5,6), that both dicots

and monocots cover a wide range of GC contents, which does not seem to differ very much in its extreme values.

Comparisons of homologous coding sequences from dicots and Gramineae

Fig. 2 shows that the GC levels of the coding sequences from the dicots analyzed range from 40 to 56% GC, with most values comprised between 44 and 54%. In other words, they correspond to the center of the compositional distribution previously seen for coding sequences from dicots (1). The coding sequences from Gramineae cover a much broader range, from 44 to 70% GC, most values being, however, in the 60-70% GC range. This distribution also largely reflects the general distribution of coding sequences from Gramineae, which comprises both a low GC and a high GC range, with most coding sequences in the latter (1).

When third codon positions of homologous genes are examined, the GC range covered in the case of dicots was 36-64%, which again, expectedly, corresponds to the center of distribution of third codon positions of all available sequences for dicots (1). In the case of Gramineae, the range was 46-100%; this covered both the low and the high GC range of all third codon positions. Expectedly again, the lowest and the highest GC values corresponded to the coding sequences which were lowest and highest in GC, respectively.

Second codon positions showed GC levels which were very close in dicots and Gramineae. The latter exhibited, however, slightly higher values in two cases and a slightly lower value in one case.

Finally, the case of first codon positions is similar to that of third codon positions, but values from Gramineae were not as much higher than dicot values; in one case, that of the actin gene, the value was even slightly lower.

A conclusion to be drawn from the data of Figs. 2-5 is that the differences in GC levels previously found in coding sequences from dicots and Gramineae (1) are also found in the set of homologous coding sequences under consideration (and in their codon positions). In other words, the differences were not due to the gene samples investigated, but were real differences,

implying compositional divergences from ancestral genes, caused by a number of directional mutations. A similar situation was already described for homologous genes from warm-blooded and cold-blooded vertebrates(7,8). These results confirm and extend those previously described by Niesbach-Klösgen et al.(9).

Since the dicots studied in their coding sequences are characterized by DNAs having relatively low GC levels, whereas the monocots most investigated belong to some sub-families of Gramineae that are characterized by high GC values, comparisons of genes from dicots and monocots, as currently possible, are strongly biased. If comparisons were done among DNAs (and coding sequences) from the monocot family Alliaceae and the same dicots, one would find compositional differences in the opposite direction.

Codon usage

A detailed statistical analysis of codon usage patterns in plant genes is being carried out and will be presented elsewhere in due time. Some conclusions can, however, be already drawn on the basis of the results obtained in this work.

First of all, as expected, all coding sequences exhibiting very high GC levels in codon third positions are characterized by the absence of a very large number (23 to 31) of codons and by very low levels (RSCU < 0.30) of some additional codons. This applies to all coding sequences higher than 90% GC in third codon position. The effect is, however, very evident also in the 80-90% GC range and perhaps already above 70% GC (this cannot be decided, however, since only one value in the 70-80% GC range is available). One should expect a similar phenomenon of codon absence also for very low GC values in third positions. Indeed, the only GC value in third codon position lower than 30% (18.4%, in the lignin-forming peroxidase gene from tobacco) shows an important level of absent or strongly underrepresented codons.

Coding sequences lower than 70% GC in third codon positions showed a different extent of codon avoidance which was, however, markedly different for different genes. Indeed, as shown by the data of Table 1, some genes with 50-60% GC in third codon position may show a strong codon avoidance (this is the case of the genes for histone H3, histone H4 and rubisco from dicots),

whereas other genes in the same GC range may show no, or almost no, avoidance of codons (this is the case of chalcone synthase from dicots and of actin and alcohol dehydrogenase from Z. mays).

Second, two opposite situations can be found for homologous genes exhibiting large differences in GC levels of third codon positions. In the first case, exemplified by the chalcone synthase genes, the gene exhibiting a very high GC level in third codon position shows the expected codon avoidance, whereas its homologous with a low GC level does not. In the second case, exemplified by rubisco and histones H3 and H4, the genes with low GC levels exhibit almost the same codon avoidance as their counterparts with high GC levels. This phenomenon has been seen for homologous genes from as many as 14 species for the rubisco gene, suggesting the possible existence of gene-specific features in codon avoidance. Needless to say, a detailed statistical analysis is required to allow a more precise assessment of this situation.

Third, codon avoidance can cover an extremely wide range, 0-52 %, within a single genome, in the case of plants (like Zea mays and, in all likelihood, of other Gramineae) that have genes characterized by an extremely broad GC range in third codon positions. In these genomes an extreme codon avoidance may concern a very large number of genes. Expectedly codon usage can also exhibit very large variations within the same genome (Table 1). This appears to be the case not only for Zea mays, but generally for plant genomes (as observed for tobacco and soybean; data not shown). This situation had already been found for vertebrate genomes (2,10).

Finally, the present results provide additional data against some misconceptions. Indeed, it is wrong to think that in compartmentalized genome, like those of plants, "in general, genes within a taxonomic group exhibit similarities in codon choice," (11). This situation, first described by Grantham et al. (12-14), only applies to compositionally homogeneous genomes, like those of most bacteria, but not to compartmentalized genomes (15). For these reasons, pooling codon usages for dicots and, even more so, for monocots (11) does not

make sense. Likewise, concluding that "the relative use of synonymous codons differs between the monocots and the dicots" (10) ignores the fact that differences are not due to differences between codon choices of monocots and dicots, but mainly to differences between GC-rich genes from Gramineae and GC-poor genes from the dicots investigated.

Comparisons of GC content of exons and introns with GC content of flanking sequences

Previous investigations (1) had indicated an isochore structure in the nuclear genome of angiosperms. The evidence rested on the following findings. (i) Extended regions (> 100-200 Kb) around the few genes which were localized on large DNA fragments fractionated by preparative Cs_2SO_4/BAMD centrifugation were shown to be compositionally homogeneous. Indeed, such fragments were the result of random breakage of DNA during its preparation, and the gene probed could therefore be present at any location on the fragments, yet the compositional distribution of the fragments carrying the gene was within 1% GC. (ii) The nuclear DNAs of the dicots and of the Gramineae investigated showed very different compositional patterns. A major difference was the presence in Gramineae of DNA fragments that showed a high GC level not represented in the dicot genomes analyzed. The GC distribution of coding sequences in Gramineae was similarly highly biased towards high GC values and covered a GC range that was not overlapped by dicot coding sequences, (a similar difference was found in introns). Taken together, these results indicated that GC-rich genes of Gramineae were present in the GC-rich regions of those genomes, whereas GC-poor genes corresponded to GC-poor regions.

In the present work, the isochore structure of the nuclear genome of plants was probed in a different way, namely by comparing the GC levels of coding sequences and introns of given genes with the corresponding flanking sequences. A certain level of variability in the results should be expected for two reasons: the small size of flanking regions, and the fact that different relative amounts of 5' and 3' flanking sequences (which may comprise different amounts of CpG islands; see

ref.16) were pooled together. The first problem could be alleviated, but not eliminated, by using only sequences longer than 1 Kb. As far as the second problem is concerned, separate plots of 5' and 3' flanking sequences revealed no significant deviation from the relationships with pooled flanking sequences.

The correlations found between GC levels of coding sequences and introns, and GC levels of flanking sequences, provide yet another evidence for an isochore structure of the nuclear genomes of plants. In apparent contrast, Brinkmann et al. (17) claimed that there was "little if any correlation" between GC levels of introns and flanking sequences, and GC levels of third codon position. Their data show, however, a slope of 0.30 and a correlation coefficient of 0.79, which is in disagreement with their conclusions, but in agreement with ours.

The data on the compositional relationships among GC levels of exons, introns and flanking sequences lead to some other conclusions. (i) The results of Fig. 6 suggest that a single correlation exists between plant genes and flanking sequences independently of whether dicots or Gramineae are considered. (ii) The correlations are similar to those found for vertebrate genomes. The possibility definitely exists, therefore, that a single general correlation holds in both cases. (iii) As in the case of vertebrates (2), coding sequences although correlated with flanking sequences, are systematically higher in GC, whereas introns are much closer to flanking sequences. (iv) The different relationships found for first, second and third codon positions in plants match similar findings apparently valid for all genomes (15). (v) Flanking sequences of Gramineae exhibit a higher GC level than those of dicots, in agreement with the distribution of DNA fragments shown in Fig.1.

ACKNOWLEDGEMENTS

We thank the French Ministry of Foreign Affairs for a fellowship to G.M., EMBO, Heidelberg (FRG), for a short-term fellowship to J.S. and INIA for financial support to L.M.M.

*To whom correspondence should be addressed

REFERENCES

1. Salinas J., Matassi G., Montero L.M. and Bernardi G. (1988) Nucl. Acids Res. 16:4269-42852.
2. Bernardi G, Olofsson B., Filipski J., Zerial M., Salinas J., Cuny G, Meunier-Rotival M. and Rodier F. (1985) Science 228:953-958.
3. Gouy M., Milleret F., Mugnier C., Jacobzone M. and Gautier C. (1984) Nucl. Acids Res. 12:121-127.
4. Sharp P.M., Tuohy T.M.F. and Mosurcki K.R. (1986) Nucl. Acids Res. 14:5125-5143.
5. Ingle J., Pearson G.G. and Sinclair J. (1973) Nature 242:193-197.
6. Shapiro, H.S. (1976). In: Fasman, G.D. (ed.) Handbook of Biochemistry and Molecular Biology, 3rd edn, vol. II, pp. 241-281, CRC Press Inc.
7. Perrin P. and Bernardi G. (1987) J. Mol. Evol. 26:301-310.
8. Bernardi G., Mouchiroud D., Gautier C. and Bernardi G. (1988) J. Mol. Evol. 28: 7-18.
9. Niesbach-Klösgen, U., Barzen, E., Bernhardt, J., Rohde, W., Schwarz-Sommer, Zs., Reif, H.J., Wienand, U. and Saedler, H. (1987) J. Mol. Evol. 26, 213-225.
10. Ikemura, T. (1985) Mol. Biol. Evol. 2:13-34.
11. Murray E.E., Lotzer J. and Eberle M. (1989) Nucl. Acids Res. 17:477-498.
12. Grantham R., Gautier C. and Gouy M. (1980) Nucl. Acids Res. 8:1893-1912.
13. Grantham R., Gautier C., Gouy M, Mercier R. and Pave A. (1980) Nucl. Acids Res. 8:r49-r62.
14. Grantham R., Gautier C., Gouy M, Jacobzone M. and Mercier R. (1981) Nucl. Acids Res. 9:r43-r74.
15. Bernardi, G. and Bernardi, G. (1986) J. Mol. Evol. 24:1-11.
16. Antequera F. and Bird A. (1988) EMBO J. 8:2295-2299.
17. Brinkmann H., Martinez P., Quigley F., Martin W. and Cerff R. (1987) J. Mol. Evol. 26:320-328.