Gene distribution and isochore organization in the nuclear genome of plants

Luis M. Montero, Julio Salinas, Giorgio Matassi[1] and Giorgio Bernardi[1]
Departamento de Proteccion Vegetal, Instituto Nacional de Investigaciones Agrarias, Carretera de La Coruna, Km 7, 28040 Madrid, Spain and [1]Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

Received September 29, 1989; Revised and Accepted January 3, 1990

ABSTRACT

The genomic distribution of 23 nuclear genes from three dicotyledons (pea, sunflower, tobacco) and five monocotyledons of the *Gramineae* family (barley, maize, rice, oat, wheat) was studied by localizing these genes in DNA fractions obtained by preparative centrifugation in Cs_2SO_4/BAMD density gradients. Each one of these genes (and of many other related genes and pseudogenes) was found to be located in DNA fragments (50 – 100 Kb in size) that were less than 1 – 2% GC apart from each other. This definitively demonstrates the existence of isochores in plant genomes, namely of compositionally homogeneous DNA regions at least 100 – 200 Kb in size. Moreover, the GC levels of the 23 coding sequences studied, of their first, second and third codon positions, and of the corresponding introns were found to be linearly correlated with the GC levels of the isochores harboring those genes. Compositional correlations displayed increasing slopes when going from second to first to third codon position with obvious effects on codon usage. Coding sequences for seed storage proteins and phytochrome of *Gramineae* deviate from the compositional correlations just described. Finally, CpG doublets of coding sequences were characterized by a shortage that decreased and vanished with increasing GC levels of the sequences. A number of these findings bear a striking similarity with results previously obtained for vertebrate genes (1, 2).

INTRODUCTION

Previous investigations (1, 2) demonstrated that the nuclear genomes of vertebrates consist of long (>300 Kilobase, Kb), compositionally homogeneous DNA segments, the isochores, that belong to a small number of families characterized by different GC levels. An important question raised by these findings, as well as by the evolutionary significance of isochores (3), concerns the spread of an isochore organization in the nuclear genomes of eukaryotes. The nuclear genomes of plants were studied to provide information on this point. These investigations (4, 5) indicated (i) that the nuclear genomes of Angiosperms are characterized by an isochore organization, and by a compositional compartmentalization; and (ii) that the nuclear genomes of *Gramineae* exhibit strikingly different compositional patterns compared with those of other families of monocots and of the dicots explored.

As far as the first point is concerned, exons and introns were found to be compositionally correlated with their flanking intergenic sequences (5); moreover, a remarkable compositional homogeneity was found in vast (>100–200 Kb) genome regions around the two genes investigated (4).

Concerning the second point, we found that the compositional distributions of nuclear DNA fragments (in the 50–100 Kb size range) from five *Gramineae* (maize, rice wheat, rye and barley) are centered about higher GC levels compared with four other monocots (scindapsus, typha, onion and asparagus) and four dicots (pea, sunflower, tobacco and antirrhinum). In addition, the compositional distribution of coding sequences from several orders of dicots is narrow, symmetrical and centered around 46% GC, whereas that from *Gramineae* (essentially barley, maize and wheat) is broad, asymmetrical and characterized by an upward trend towards high GC values, with the majority of sequences between 60 and 70% GC (4). Sets of homologous coding sequences revealed the same features (5), ruling out the possibility that the differences found were due to differences in the gene samples investigated. Introns exhibited similar compositional distributions, compared to the exons from the same genes, but were systematically lower in their GC levels (4).

In the present work, we have very considerably extended the number of genes (from 2 to 23 genes) localized in density gradient fractions and we have investigated the correlations between the GC levels of coding sequences (and of their different codon positions), and the GC levels of the DNA fractions in which the genes were localized. We have also studied these correlations for all available introns of the genes investigated, as well as the correlations between the levels of the CpG doublet and the GC levels of coding sequences.

MATERIALS AND METHODS

DNAs from ethiolated seedlings of *Hordeum vulgare* (barley) and *Avena sativa* (oat) were obtained as previously described (4, 5). These DNA preparations were in the 50–100 Kb size range, as judged by gel electrophoresis with appropriate size markers.

DNA preparations and fractionations from other plants were described elsewhere (4, 5). In the case of maize, the fractions

described in the legend of Fig. 1 of ref. 4 were used. In this work, DNA fractionation was carried out by preparative density gradient centrifugation in Cs_2SO_4/BAMD, as already described (4); BAMD is 3,6 bis (acetato-mercuri-methyl) dioxane.

EcoRI restriction endonuclease digestion of total DNAs and their fractions, and hybridization experiments were carried out as already described (4). Hybridization results were used to identify coding sequences known in their primary structure from previous work of other laboratories (see refs. in Table 1). The probes used and their sources are given in Table 1.

GC levels of coding sequences (from the initial AUG to the termination codon), of first, second and third codon positions, and of introns were obtained from GenBank, Release 59 (March 1989) and from the literature. The ACNUC retrieval system (6) was used.

RESULTS

Fractionation of DNAs

Fractionations by preparative density gradient centrifugation in Cs_2SO_4/BAMD of the DNAs from *Pisum sativum* (pea), *Helianthus annuus* (sunflower), *Nicotiana tabacum* (tobacco), *Zea mays* (maize), *Oriza sativa* (rice) and *Triticum aestivum* (wheat) were already described (4). Results for *Hordeum vulgare* (barley) and *Avena sativa* (oat) are shown in Fig. 1. Compared to wheat DNA, barley DNA is slightly shifted towards lower buoyant density values in the case of high GC fractions, whereas the buoyant densities of oat DNA begins and ends at significantly lower values.

Localization of genes in fractionated DNAs

Table 1 summarizes the hybridization results as obtained on the 23 genes tested. Hybridization patterns can be classified in two groups according to whether hybridization bands show one (group *a*) or more (group *b*) peaks of intensity across the density gradient. By definition, a single-copy gene can only give a hybridization pattern belonging to the first group, whereas multiple-copy genes may belong to either group. Multiple-copy genes of the first group usually are physically clustered (although the possibility exists of scattered genes hybridizing to the same isochore family). Multiple-copy genes of the second group are physically scattered (the only exceptions would correspond to clustered genes located on contiguous isochores having different GC levels). Table 1 indicates to which group the genes tested belong.

Figures 2–4 show examples of gene localization in DNA fractions. The hybridization of the oat phytochrome gene probe on oat DNA showed two main bands in several fractions (Fig. 2A). The 5.2 Kb band corresponds to the sequence for which the primary structure is available. This band was mainly localized in fraction 5 (ϱ = 1.7013 g/cm^3) and in fraction 6 which had an almost identical density (ϱ = 1.7016 g/cm^3). Interestingly, the other band was mainly localized in fractions 6 (ϱ = 1.7016 g/cm^3) and 7 (ϱ = 1.7024 g/cm^3). The two hybridization bands show, therefore, different distributions in density fractions. This result indicates that the phytochrome multigene family belongs to group *b* (see above). Fig. 2B presents the results obtained by hybridizing a maize alcohol dehydrogenase (ADH) gene probe on maize DNA. Two bands were localized in contiguous fractions 3 (ϱ = 1.7018 g/cm^3) and 4 (ϱ = 1.7032 g/cm^3), and were assigned a 1.7022 g/cm^3 density since the lighter fraction showed more intense hybridization signals. The *ADH* hybridization pattern belongs to group *a* (see above).

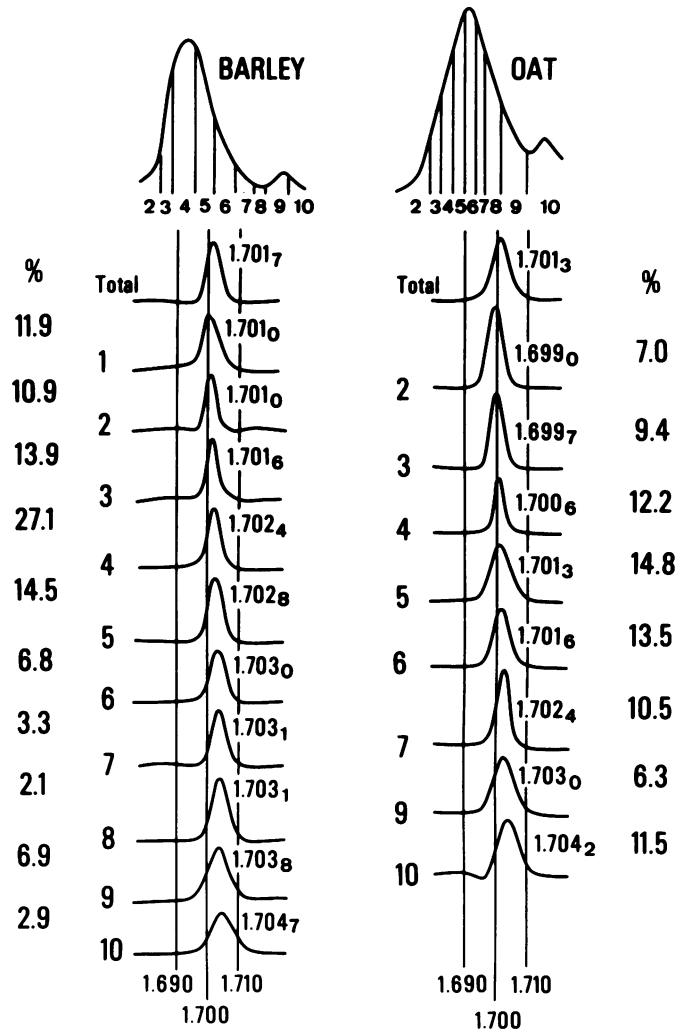

Figure 1. Fractionation by preparative Cs_2SO_4/BAMD density gradient centrifugation of DNAs from two *Gramineae*, barley and oat. Experimental conditions were identical to those already described (4). The top curves display the transmission profiles at 254.7 nm as recorded by an LKB ultraviolet recorder. The numbers indicate the fractions; fraction 1 corresponds to the pelleted DNA. The analytical CsCl profiles of unfractionated (total) DNA and of fractions pooled from pairs of centrifugation tubes derived from a single run are also shown. Modal buoyant densities (g/cm^3) and relative amounts of DNAs are indicated. In the oat DNA fractionation, the profiles of the pellet and of fraction 8 are not shown because these fractions were lost accidentally.

The hybridization of a wheat chlorophyll-a/b-binding protein (*Cab*) gene probe on digests of DNA fractions from pea (Fig. 3A) indicates that the gene hybridizing the probe is localized in fraction 9 (ϱ = 1.6977 g/cm^3), weaker signals being shown by the neighboring fractions 8 and 10, which are characterized by extremely close modal buoyant densities, 1.6970 and 1.6973 g/cm^3. In contrast, the same *Cab* probe revealed (Fig. 3B) a very large number of bands in wheat, all of them essentially localized in a single fraction, 4 (ϱ = 1.7029 g/cm^3). The gene corresponding to the sequence of interest was that present on a 1.6 Kb fragment. The two *Cab* hybridization results just described obviously belong to group *a*; in the latter case, in all likelihood, this is due to gene clustering.

The hybridization results obtained with a pea ribulose-1, 5-biophosphate carboxylase, small sub-unit (*Rubisco*) gene probe on tobacco DNA fractions (Fig. 4A) showed a large number of

Table I. Genes localized in plant DNA fractions

N°	GENE	PLANT	PROBE name	HYBRIDIZATION ref.	pattern (g/cm³)	fraction	ref.
1	GS	Pea	pR1	47	a	1.6955	13
2	rbcS	"	pSS15	48	b	1.6951	14,15
3	ADH-1	"	pH2.3	49	a	1.6951	16
4	Leg A	"	pDUB6	50	a	1.6956	17–19
5	Cab	"	CabIIA	45	a	1.6977	20
6	rbcS	Sunflower	pSS15	48	a	1.6959	21
7	rbcS	Tobacco	pSS15	48	a	1.6961	22
8	Zein	Maize	pML1	51	b	1.7013	23, 24
9	ADH-1	"	pH2.3	49	a	1.7018	25
10	Ch.s.	"	pLC30	26	a	1.7032	26, 27
11	H4	"	H4C14	28	b	1.7032	28, 29
12	Wx	"	pWXC17	30	a	1.7032	30, 31
13	rbcS	"	SS7	32	b	1.7038	32, 33
14	PEPCase	"	pH1	34	a	1.7032	34, 35
15	B1-Hordein	Barley	pBHR184	36	b	1.7016	36
16	α-Amylase	"	2128	46	a	1.7024	37
17	Phytochrome	Oat	pAP3	52	b	1.7013	38, 39
18	Glutelin	Rice	pREE61	40	b	1.7020	40, 41
19	α/β-gliadins	Wheat	WJZ35A1	pers.com.	b	1.7016	42
20	H3	"	H3C4	29	a	1.7029	43
21	rbcS	"	SS7	32	a	1.7029	44
22	Cab	"	CabIIA	45	a	1.7029	45
23	α-Amylase	"	2128	46	a	1.7039	46

This Table provides the numbering of the genes tested (as used in Figs. 5–8), the plant sources, the probes used and the corresponding literature references,the hybridization patterns (see below), the modal buoyant densities of the fractions in which the genes, whose primary structure was available, were localized, and the literature references concerning the size of hybridizing EcoRI fragments that correspond to the sequence considered. Concerning hybridization patterns, groups a and b correspond to bands showing respectively one or more than one peak of intensity across the density gradient; the first group practically corresponds to either single-copy or multiple-copy clustered genes, the second one to multiple-copy scattered genes (see also text). Gene abbreviations not already given in the text are: GS, Glutamine synthetase; Leg. A, Legumin A; Ch.s. Chalcone synthase; H3 and H4, Histones H3 and H4; Wx, waxy, UDP-glucose starch glycosyl transferase; PEPCase, Phosphoenolpyruvate carboxylase.

bands in the unfractionated DNA and in fraction 4 (ϱ = 1.6961 g/cm³), with some weaker bands also appearing in fraction 5, which had an essentially identical buoyant density (ϱ = 1.6963 g/cm³). In this case, the sequence under consideration corresponds to a fragment characterized by a molecular size of 4.4 Kb. Fig. 4B shows the more complex patterns that were obtained when the same *Rubisco* probe was hybridized on EcoRI digests of pea DNA fractions. In this case too, the hybridization pattern comprised many bands on unfractionated DNA, but different hybridization bands appeared to have a different distribution in density fractions. Some bands showed their highest intensity in fraction 1 (ϱ = 1.6922 g/cm³), other ones in fractions 3–5 (that range in modal buoyant density from 1.6951 to 1.6959 g/cm³), and a faint band appears in fraction 8 (ϱ = 1.6970 g/cm³). The pea *Rubisco* sequence of interest corresponds to a fragment of about 5 Kb, that is present in fractions 2–4. In conclusion, the hybridization patterns of Figs. 4A and 4B belong to groups *a* and *b*, respectively.

It should be stressed that a total of about 230 hybridization bands were detected on the filters used to localize the 23 genes under investigation. All these bands showed narrow distributions over the DNA fractions covering, on the average, a range of 1.4% GC.

Compositional correlations between nuclear genes and isochores

Figure 5 presents a plot of GC levels of coding sequences against the buoyant densities of the fractions in which the genes were localized. If one neglects for the time being a group of five genes from *Gramineae* (that are mentioned below), a straight line, (with a slope s = 2.3 and a correlation coefficient R = 0.93), can be drawn by the least square method through the points from both dicots and *Gramineae*. It should be stressed that such slope was obtained by using GC values of DNA fractions derived from modal buoyant densities under the assumption (known to be wrong) of no DNA methylation. This point will be further commented upon in the Discussion.

The group of genes from *Gramineae* which remarkably deviate from the other ones comprise the phytochrome gene from oat and four genes for seed storage proteins, zein from maize, B1-hordein from barley, glutelin from rice and α/β-gliadins from wheat. The behavior of these genes will be further commented upon in the Discussion.

The analysis of the correlation between GC levels of third codon position and buoyant density (Fig. 6) reveals features that are largely similar to those just described for coding sequences, (under the same no-methylation assumption as above, and again neglecting the group of five genes mentioned in the previous paragraph). In this case, the correlation coefficient is as high as in the case of Fig. 5 (R= 0.92), but the slope is much higher (s = 5.2). While the third codon positions of dicot genes were, on the average, about 10% higher in GC than the line of slope 1, those of genes from *Gramineae* were almost 50% higher in GC. In contrast, third codon position of storage protein genes and phytochrome practically fell on the unity slope line passing through the origin, with two points, those for the genes of B1-hordein and α/β-gliadins, exhibiting even lower GC values. The plot for first codon positions (Fig. 6) showed a lower slope (s = 0.98; R = 0.66) than those of coding sequences, with two

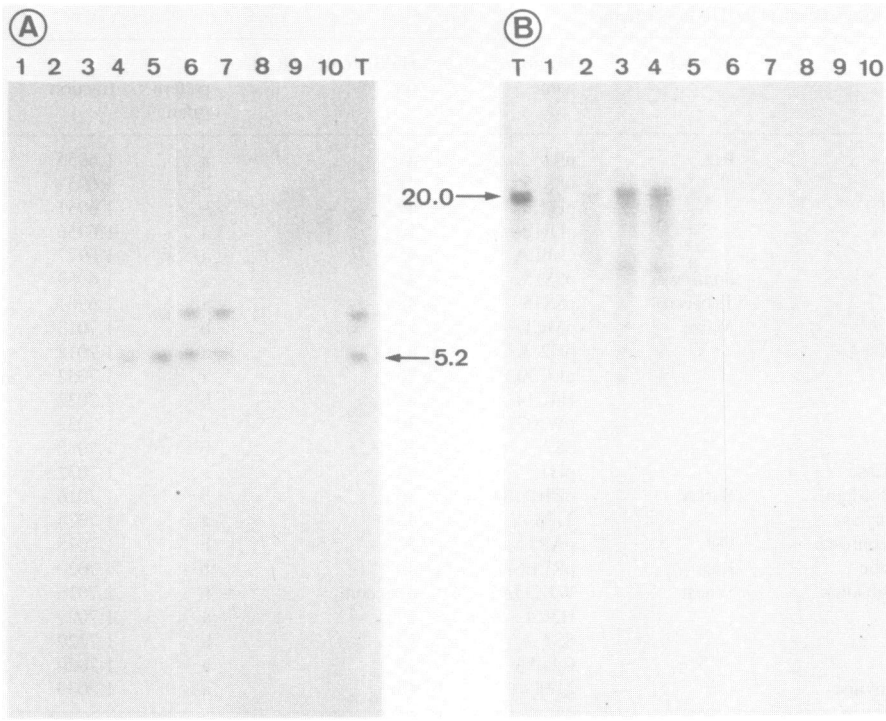

Figure 2. Localization of the phytochrome gene from oat (a) and of the ADH gene from maize (b) on total, unfractionated DNAs (T) and on Cs_2SO_4/BAMD DNA fractions. 10 μg of total DNA and DNA fractions in amounts corresponding to 10 μg of total DNA were processed as indicated in Materials and Methods. Arrows indicate the hybridization band(s) identifying the gene of interest.

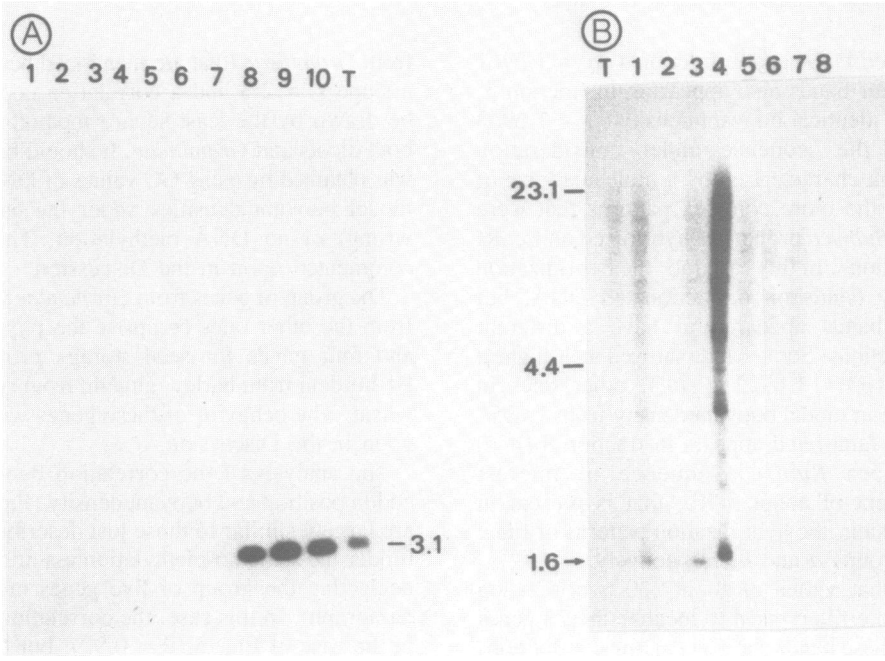

Figure 3. Localization of *Cab* gene from pea (a) and wheat (b) DNA fractions. For other indications, see legend of Fig. 2.

genes, those for B1-hordein and α/β-gliadins, standing out, again on the low GC side. The plot for second position GC (Fig. 6) showed an extremely low slope (s = 0.16; R = 0.28), with two points, those for B1-hordein and α/β-gliadins, standing out, but on the high side relative to the linear relationship exhibited by all other genes.

Finally, the intron line (Fig. 5) is practically parallel to the exon line (s = 2.1; R = 0.92), but lower by 20% GC. In fact,

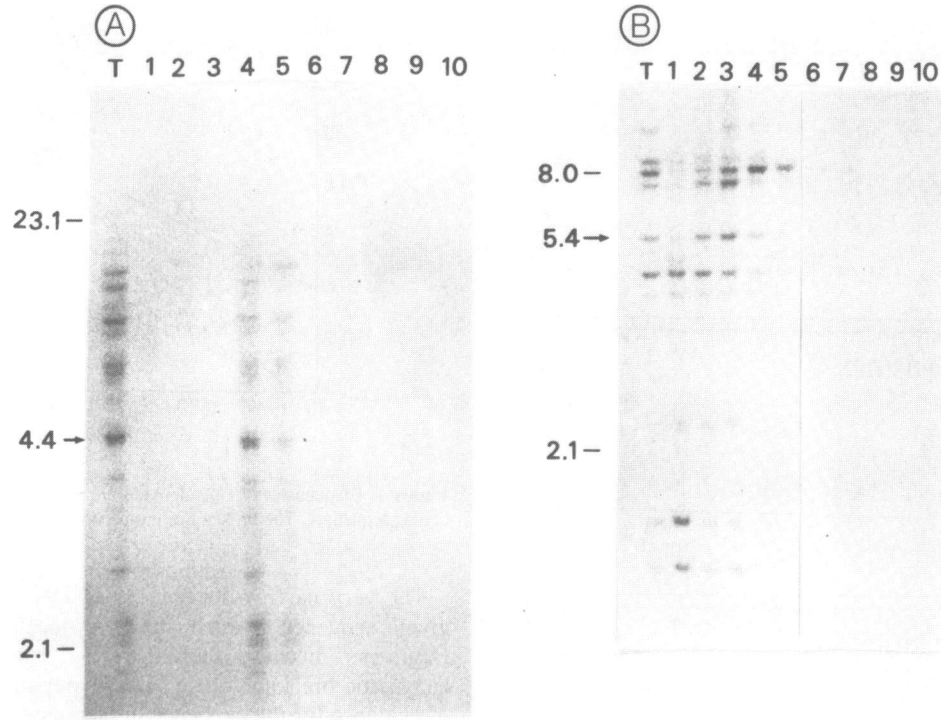

Figure 4. Localization of Rubisco genes from tobacco (a) and from pea (b) DNA fractions. For other indications see legend of Fig. 2.

intron GC levels are even lower than those of the sequences flanking the genes (these correspond to the unity slope; see Discussion). In the case of the only storage protein gene for which intron data are available, that of glutelin from rice, the point stands out on the low GC side; this point was obviously neglected in calculating the intron line.

CpG levels in coding sequences

The diagram of Figure 7 shows that when the CpG levels of all coding sequences examined are plotted against the GC levels of the sequences, points fall on a line (R = 0.98) that is characterized by a shortage of the doublet relative to statistical expectations. This shortage, however, decreases and vanishes with increasing GC of coding sequences. Expectedly, if the CpG level of plant coding sequences are plotted against the GC levels of the DNA fragments in which the sequences were located, a positive correlation was found (not shown) and the points were essentially distributed as in the exon plot (Fig. 5).

DISCUSSION

The fractionation and hybridization patterns

The fractionation results of barley DNA are rather similar to those previously obtained for wheat DNA (4), with modal buoyant densities of the high density fractions only slightly shifted towards lower values. In contrast, the compositional pattern of oat DNA is characterized by a rather large shift towards lower values, particularly on the GC-poor side of the distribution, compared to wheat DNA. In connection with these results, it should be recalled that while barley and wheat belong to the same sub-family *Triticeae* of *Gramineae*, oat belongs to a different sub-family, *Aveneae*. It should be stressed that the CsCl profiles of DNA

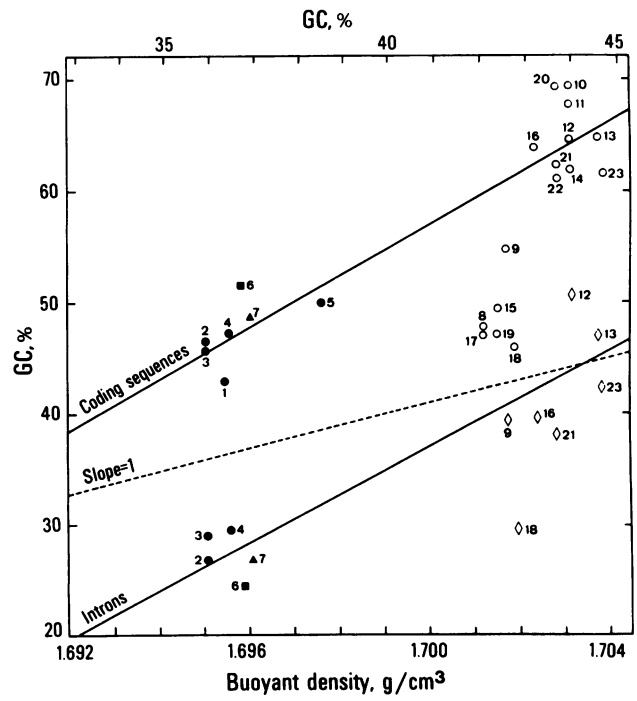

Figure 5. GC levels of coding sequences and introns of plant genes are plotted against the modal buoyant density of the Cs_2SO_4/BAMD fractions in which the coding sequences were localized. The upper scale provides GC levels of the fractions, as obtained from buoyant densities, under the assumption of no DNA methylation. For the numbering of genes, see Table 1. Filled symbols refer to genes from dicots (circles, pea; squares, sunflower; triangles, tobacco), empty symbols to genes from monocots belonging to the *Gramineae* family (circles, coding sequences; lozenges, introns).

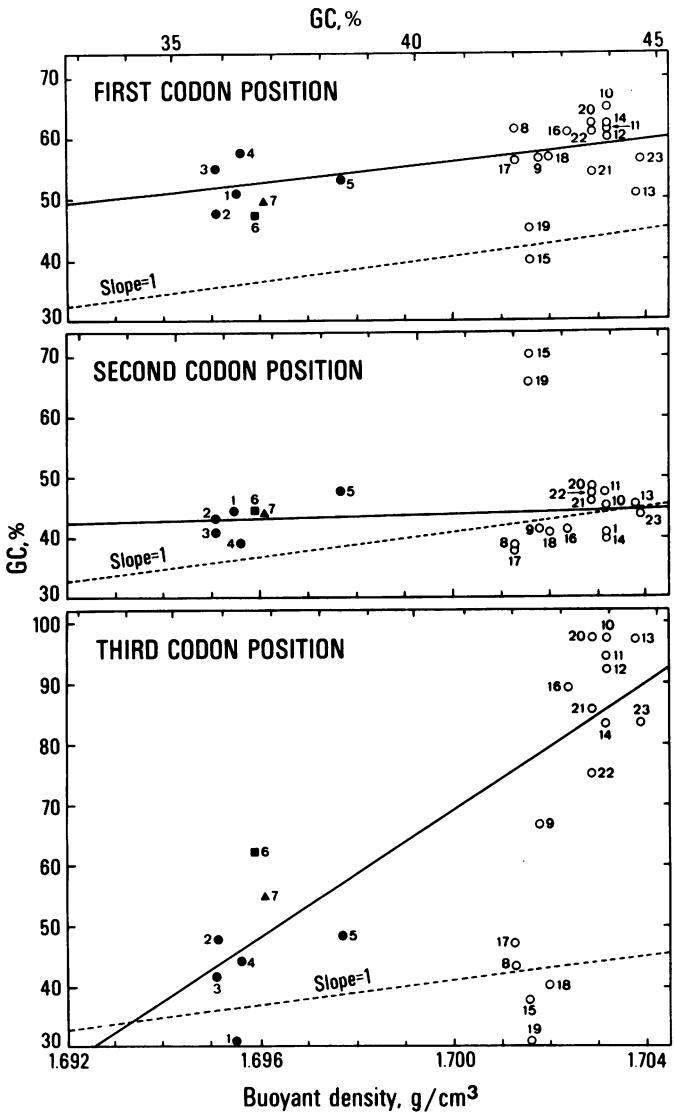

Figure 6. GC levels of first, second and third codon positions of plant genes are plotted as indicated in the legend of Fig. 5, using the same numbering and symbols.

fractions (like those of Fig. 1) are of crucial importance as far as the interpretation of hybridization results is concerned. Indeed, they provide the modal buoyant density as well as the distribution of buoyant densities of the DNA fragments present in each fraction from the preparative ultracentrifugation that was used in hybridization experiments.

The hybridization results confirm and considerably extend our initial observation (4) that a given gene is located on DNA fragments that are extremely close in GC levels. In all cases studied, the GC range of the fractions carrying the gene tested was less than 1–2%. In fact, the conclusion about regional compositional homogeneity does not only derive from the results concerning the genes studied in more detail. The results obtained on all genes and pseudogenes from the multigene families investigated (corresponding to 230 hybridization bands) lead to the same conclusion, since the average spread of any hybridization band over neighboring fractions is only 1.4% GC. Obviously, a rather large number of both clustered and scattered coding sequences and pseudogenes were tested in this way.

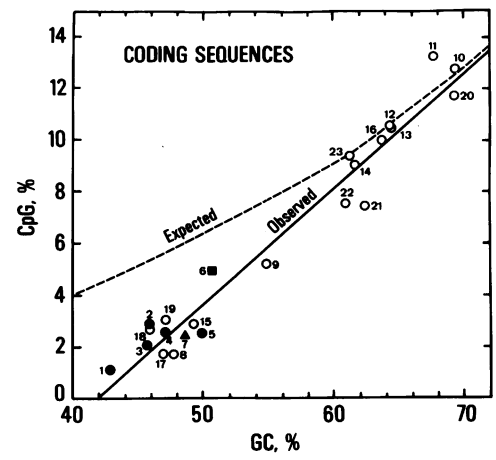

Figure 7. Frequencies of CpG doublets are plotted against the GC levels of plant coding sequences. The broken line corresponds to the statistically expected values.

The very narrow distribution of DNA fragments carrying a given sequence (which may be located anywhere on the fragments, because they arise by random mechanical and enzymatic breakage during DNA preparation) is the main line of evidence for the existence in plant genomes of DNA regions, at least 100–200 Kb in size, that are very homogeneous in composition, the isochores.

The hybridization results lead to yet another conclusion. Except for the storage protein and the phytochrome genes, the genes from *Gramineae* that were tested are clustered between the buoyant densities of 1.702 and 1.704 g/cm^3. Yet, the corresponding coding sequences and third codon positions range from 60 to 70% GC, and from 70 to 95% GC, respectively. In both cases, the higher values (70% and 95%, respectively) are among the highest ones found in genes from *Gramineae* (4) and yet they correspond to genes located in DNA fragments not higher than 1.704 g/cm^3. Now, GC levels of coding sequences and codon third positions are correlated with the GC levels of the isochores in which the corresponding genes are contained (see following section). These data strongly suggest, therefore, that DNA higher than 1.704 g/cm^3 in buoyant density does not contain coding sequences, but is rather made up of repetitive DNA and ribosomal DNA. Another finding in favor of this interpretation is the variability from species to species in the family of *Gramineae* of DNA fractions having buoyant densities higher than 1.704 g/cm^3 (see Fig. 1, and also Fig. 2 from ref. 4).

The points just made are of interest in another respect as well. In the case of warm-blooded vertebrates, the genes that are highest in GC in their coding sequences (up to 70%) and third codon positions (up to 90%) are located in the DNA fractions that are highest in GC (ribosomal DNA being, however, still higher) and that are in the 1.708–1.712 g/cm^3 buoyant density range, which is much higher than that found in the case of *Gramineae*. In the case of *Gramineae*, two possibilities should, therefore, be considered, namely either (i) that the difference in GC levels between coding sequences (and third codon positions) and the DNA fragments containing them is much smaller than in vertebrates; or (ii) that this smaller difference is only apparent and due to the very high methylation of the DNA of *Gramineae*. Arguments will be provided below in favor of this second explanation.

The compositional correlations between genes and isochores

The compositional correlations between coding sequences, or introns, and the DNA fractions carrying them, deserve several comments.

(i) *The meaning of compositional correlations.* A general comment is that the relationships of Figs. 5 and 6 concern, in fact, on the one hand, coding sequences and coding positions or introns and, on the other hand, intergenic non-coding sequences, since the latter represent the vast majority of plant genomes. For this reason, the unity slope line passing through the origin corresponds to the identity of GC levels of coding sequences, codon positions and introns with GC levels of intergenic non-coding sequences. As such, this unity slope line represents a very useful reference, since points above and below it correspond to sequences that have higher and lower GC levels, respectively, than intergenic, non-coding sequences.

(ii) *Existence of a single compositional correlation for monocot and dicot genes.* It may be useful to explain why a line was drawn in Fig. 5 through the points of coding sequences from both dicots and *Gramineae*, leaving as a separate cluster the phytochrome and seed storage protein genes, instead of either drawing two separate lines through the dicot points and the *Gramineae* points, respectively, or a line through all the points. (a) First of all, the gap between the dicot and the *Gramineae* points is only due to the DNAs investigated. Our previous results (5) show the existence of monocot DNAs that are poorer in GC than the dicot DNAs investigated, of dicot DNAs that are richer in GC than some DNAs from monocots, as well as of DNAs from monocots and dicots that cover the intermediate GC range. There is, therefore, no compositional gap between DNAs from dicots and monocots. If such is the case, there must be a continuum of GC levels of coding sequences, provided that appropriate plant species are investigated. Such GC levels would fill the gap present in Fig. 5. (b) The lines obtained for first and second codon positions are non-ambiguous (the deviating points being explained under v, below) and they exhibit the expected lower slopes relative to the third codon position line (7). Drawing two lines, for dicots and *Gramineae*, through third codon position points, would not make sense in view of the compositional correlations between different codon positions as found in all other genomes (7). (c) Drawing a line using all the points would lower its slope and its correlation coefficient. Ignoring the clustering of a group of genes would not make sense, however, in view of other considerations (see below).

(iii) *The effect of DNA methylation on the compositional correlations.* It was already mentioned that the slopes of Figs. 5 and 6 were calculated from the buoyant densities of the fractions, under the assumption that intergenic, non-coding DNA is non-methylated . It is well known, however, that plant DNAs have relatively high levels of methylation, and that methylation causes a decrease of buoyant density of about 0.7 mg/cm^3 per 1% of 5-methylcytosine (8–10). For the plant DNAs under consideration, methylation levels are in the 5–7% range of 5-methylcytosine, with rice DNA exhibiting a lower level, 3%. In other words, the GC levels on the abscissa axis might be shifted to the right by 3.5–5% GC. The difference in methylation of the DNAs used are, however, less than 1–2%, with a larger value for rice. No serious effects of methylation can therefore be expected on the slopes of Figs. 5 and 6, except for a slight lowering due to the generally higher methylation levels of *Gramineae* compared to the dicots under consideration. On the other hand, the fact that methylation may not be uniform through

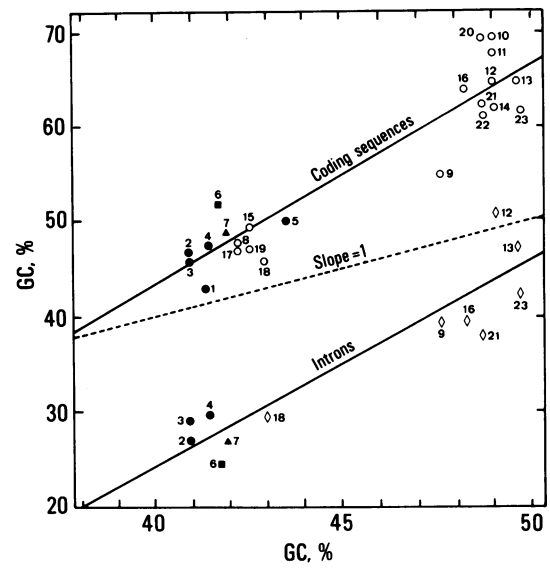

Figure 8. GC levels of coding sequences and introns of plant genes are plotted against the estimated GC levels of the Cs_2SO_4/BAMD fractions in which the coding sequences were localized. Estimation of the latter GC levels from the modal buoyant densities in CsCl of the fractions was made assuming a 5% level of 5-methylcytosine for the fractions containing all genes except for those containing the phytochrome and storage protein genes (these were assumed to be methylation-free). See also the Discussion. Numbering and Symbols as in Fig. 5.

the genome may probably account for the relatively larger scatter of points, particularly in the case of the genes from *Gramineae*.

(iv) *Explanations for the deviation of some monocot genes from the compositional correlations.* (a) Non-uniformity of DNA methylation may in fact account for the remarkable deviation of the genes for oat phytochrome and storage proteins genes of *Gramineae* from the relationships exhibited by coding sequences, introns and third codon positions. In fact, if these genes were located in methylation-free or methylation-poor regions of the genome, the shift to the right of the plots (to take into account the effect of methylation on buoyant density; see Fig. 8) would put the five genes together with the genes of the dicots investigated, with which they share both GC levels and CpG levels (see Fig. 7). (b) An alternative explanation might be that the 'deviating' genes have in fact been translated to a different genomic environment, that is much closer in GC. This explanation, which cannot be ruled out for the isolated case of oat phytochrome, is unlikely to be correct for the seed storage protein genes. Indeed, these genes comprise multiple copies that are scattered over different isochores and in the genomes of four monocots (maize, barley, wheat, and rice) that belong to three different sub-families of *Gramineae*. Under these circumstances, a common pattern in the genome environment of these genes would be a most surprising coincidence. A remarkable difference between the two classes of genes of *Gramineae* just described is that the GC-poor class (like the genes of the dicots under consideration) is very close in GC levels to the intergenic non-coding sequences, whereas the GC-rich class is characterized by much higher GC levels (about 20%). Likewise, third codon position of GC-poor genes would be very close in GC levels to the intergenic non-coding sequences, whereas those of GC-rich genes would be almost 50% higher in GC. The much lower CpG level of the 'deviating' coding sequences relative to the other genes from *Gramineae* (Fig. 7) is also in agreement with the

explanation provided under (a) above. Since CpG is one of the main sites of methylation in plant genomes, the potential methylation level of these gene is lower compared to the other genes.

(v) *Some other features of the seed storage protein genes.* Two of these genes, those for B1-hordein and for α/β-gliadins, also exhibit a very low GC level in first codon position and an extremely high GC level in second codon position. The latter phenomena are certainly due to the very high levels of alanine and, more so, of threonine in the corresponding proteins. Indeed, alanine and threonine codons have only C in their second positions, and threonine codons have only A in the first codon positions.

(vi) *The compositional correlation of introns.* Introns are characterized by very low GC values and by a slope close to that of coding sequences. The difference between GC levels of introns and exons is very large, about 20% GC, and puts intron GC levels below the GC level of intergenic non-coding sequences. Interestingly, the very low GC value of the introns of the only storage protein gene available, that of rice glutelin, is more than 10% lower than the corresponding intergenic, non-coding sequences. It should be noted that these local differences are not relevant as far as the long range homogeneity of isochores is concerned.

Comparison of the genomes of angiosperms and vertebrates

The homogeneous composition found around 23 genes from angiosperms (and around a much larger number of other genes and pseudogenes present in the multigene families tested) provides a definitive demonstration for the existence of isochores in plant genomes. In this respect, the organization of the plant genomes resembles very closely that of vertebrate genomes. More specifically, the organization of the genomes of the dicots studied resembles that exhibited by cold-blooded vertebrates, whereas that of *Gramineae* is more similar to that of warm-blooded vertebrates (4). It has been suggested (3) that the very high GC levels of coding sequences and third codon positions of most genes from warm-blooded vertebrates and *Gramineae* are due to directional changes towards high GC under the selective pressure of higher environmental temperature. Such directional changes lead to a situation in which homologous genes from the dicots investigated and from *Gramineae* remarkably differ in GC contents in third codon position and also, to a lesser extent, in first and second position. In any case, both genomic systems are characterized by a strikingly non-uniform distribution of genes, and by the effects of compositional compartmentalization on codon usage and even on amino acid composition of encoded proteins.

Another point of similarity in the organization of the genomes of plants and vertebrates is the existence of relationships between the GC levels of coding sequences and introns and the GC levels of the DNA segments harboring them. The two systems are not identical, however, the main differences being the higher slope of the exon relationship and the systematically lower GC levels of introns compared to the corresponding exons. In vertebrates, the exon slope is close to one and the difference in GC levels between exons and non-coding intergenic sequences is constant and equal to about 10% GC. In contrast, in plants the difference ranges from 10% for low GC genes to over 20% GC for GC-rich genes. On the other hand, introns of vertebrates are lower in GC than exons for low GC exons, but reach higher GC values than exons for GC rich exons, whereas in plants the difference is constant and very large, about 20% GC.

Finally, the correlations found between the levels of CpG doublets and the GC levels of coding sequences (and, therefore, of the corresponding isochores) of plants are practically identical with those previously reported for vertebrate genes (1). Obviously, the difference in this case concerns the methylation level which is so much lower in vertebrate DNAs (2–8% of total cytosines) than in plant DNAs (up to 30% of all cytosines), where it concerns not only the doublet CpG, but also the doublets CpA, CpT and the triplets CNG (11). In the case of vertebrates, the decreasing discrimination against CpG is accompanied by an increase in CpG islands (2). It is possible that the same situation exist in the genomes of plants, where CpG islands have been recently found (12).

ACKNOWLEGEMENTS

We thank the Institute Nacional de Investigationes Agrarias of Spain for a fellowship to L.L.M., EMBO for a fellowship to J.S. and the French Ministry of Foreign Affairs for a fellowship to G.M. We also thank Drs. B.G. Forde, A.R. Cashmore, M. Freeling, R.R.D. Croy, N.-H. Chua, J.A. Pintor-Toro, U. Wienand, G. Gigot, R.B. Klösgen, J.Y. Sheen, J.W. Grula, A.A. Gatenby, P.H. Quail, F. Takaiwa and J.V. Torres for the gift of DNA probes. We also thank Dr. W.F. Martin for critical comments.

REFERENCES

1. Bernardi,G., Olofsson,B., Filipski,J., Zerial,M., Salinas, J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) *Science* **228**, 953–958.
2. Bernardi,G. (1989) *Ann. Rev. Genet.* **23**, 637–661.
3. Bernardi,G., Mouchiroud,D., Gautier,C. and Bernardi,G. (1988) *J. Mol. Evol.* **28**, 7–18.
4. Salinas,J., Matassi,G., Montero,L.M. and Bernardi,G. (1988) *Nucl. Acids Res.* **16**, 4269–4285.
5. Matassi,G., Montero,L.M., Salinas,J. and Bernardi,G. (1989) *Nucl. Acids Res.* **17**, 5273–5290.
6. Gouy,M., Milleret,F., Mugnier,C., Jacobzone,M. and Gautier,C. (1984) *Nucl. Acids Res.* **12**, 121–127.
7. Bernardi,G. and Bernardi,G. (1986) *J. Mol. Evol.* **24**, 1–11.
8. 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.
9. Kemp,J.D. and Sutton,D.W. (1976) *Biochim. Biophys. Acta* **425**, 148–156.
10. Wagner,I. and Capesius,I. (1980) *Biochim. Biophys. Acta* **654**, 52–56.
11. Gruenbaum,Y., Navey-Many,T., Cedar,H. and Razin,A. (1981) *Nature* **329**, 860–862.
12. Antequera,F. and Bird,A. (1988) *EMBO J.* **8**, 2295–2299.
13. Tingey,S.V., Walker,E.L. and Coruzzi,G.M. (1987) *EMBO J.* **6**, 1–9.
14. Cashmore,A.R. (1983) In: 'Genetic engineering of plants. An agricultural perspective' (Kosuge,T., Meredith,C.P. and Hollander, A., Eds.), pp. 29–38, Plenum, New York.
15. Coruzzi,G., Broglie,R., Edwards,E. and Chua,N.-H. (1984) *EMBO J.* **3**, 1671–1679.
16. Llewellyn,D.J., Finnegan,E.J., Ellis,J.G., Dennis,E.S. and Peacock,W.J. (1987) *J. Mol. Biol.* **195**, 115–123.
17. Lycett G.W., Croy,R.R.D., Shirsat,A.M. and Boulter,D. (1984) *Nucl. Acids Res.* **12**, 4493–4506.
18. Domoney,C., Barker,D. and Casey,R. (1986) *Plant Mol. Biol.* **7**, 467–474.
19. Domoney,C., Ellis,T.H.N. and Davis,D.R. (1986) *Mol. Gen. Genet.* **202**, 280–285.
20. Cashmore,A.R. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 2960–2964.
21. Waksman,G. and Freyssinet,G. (1987) *Nucl. Acids Res.* **15**, 1328.
22. Mazur,B.J. and Chui,C.F. (1985) *Nucl. Acids Res.* **13**, 2373–2386.
23. Kridl,J.C., Vieira,J., Rubenstein,I. and Messing,J. (1984) *Gene* **28**, 113–118.
24. Wilson D.R. and Larkins,B.A. (1984) *J. Mol. Evol.* **20**, 330–340.
25. Dennis,E.S., Gerlach,W.L., Pryor,A.J., Bennetzen,J.L., Inglis,A.,

Llewellyn,D., Sachs,M.M., Ferl,R.J. and Peacock,W.J. (1984) *Nucl. Acids Res.* **12**, 3983–4000.
26. Wienand,U., Weydemann,U., Niesbach-Klösgen,U., Peterson,P.A. and Saedler,H. (1986) *Mol. Gen. Genet.* **203**, 202–207.
27. 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.
28. Philipps,G., Chaubet,N., Chaboute,M.-E., Ehling,M. and Gigot,C. (1986) *Gene* **42**, 225–229.
29. Chaubet,N., Philipps,G., Chaboute,M.-E., Ehling,M. and Gigot,C. (1986) *Plant Mol. Biol.* **6**, 253–263.
30. Schwarz-Sommer,Zs., Gierl,A., Klösgen,R.B., Wienand,U., Peterson,P.A. and Saedler,H. (1984) *EMBO J.* **3**, 1021–1028.
31. Klösgen,R.B., Gierl,A., Schwarz-Sommer,Zs. and Saedler,H. (1986) *Mol. Gen. Genet.* **203**, 237–244.
32. Sheen,J.-Y. and Bogorad,L. (1986) *EMBO J.* **5**, 3417–3422.
33. Lebrun,M., Waksman,G. and Freyssinet,G. (1987) *Nucl. Acids Res.* **15**, 4360.
34. Hudspeth,R.L., Glackin,C.A., Bonner,J. and Grula,J.W. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2884–2888.
35. Izui,K., Ishijima,S., Yamaguchi,Y., Katagiri,F., Murata,T., Shigesada,K., Sugiyama,T. and Katsuki,H. (1986) *Nucl. Acids Res.* **14**, 1615–1628.
36. Forde,B.G., Heyworth,A., Pywell,J. and Kreis,M. (1985) *Nucl. Acids Res.* **13**, 7327–7339.
37. Knox,C.A.P., Sonthayanon,B., Chandra,G.R. and Muthukrishnan, S. (1987) *Plant Mol. Biol.* **9**, 3–17.
38. Hershey,H.P., Barker,R.F., Idler,K.B., Lissemore,J.L. and Quail,P.H. (1985) *Nucl. Acids Res.* **13**, 8543–8559.
39. Hershey,H.P., Barker,R.F., Colbert,J.T., Lissemore,J.L. and Quail,P.H. (1985) In: 'Molecular form and function of the plant genome' (van Vloten-Doting L., Groot G.S.P. and Hall,T.C. Eds.), pp. 101–111. Plenum, New York.
40. Takaiwa,F., Kikuchi,S. and Oono,K. (1986) *FEBS Lett.* **206**, 33–35.
41. Takaiwa,F., Kikuchi,S. and Oono,K. (1987) *Mol. Gen Genet.* **208**, 15–22.
42. Okita,T.W., Cheesbrough,V. and Reeves,C.D. (1985) *J. Biol. Chem.* **260**, 8203–8213.
43. Tabata,T., Fukasawa,M. and Iwabuchi,M. (1984) *Mol. Gen. Genet.* **196**, 397–400.
44. Broglie,R., Coruzzi,G., Lamppa,G., Keith,B. and Chua,N.-H. (1983) *Bio-Technology* **1**, 55–6l.
45. Lamppa,G.K., Morelli,G. and Chua,N.-H. (1985) *Mol. Cell Biol.* **5**, 1370–1378.
46. Baulcombe,D.C., Huttly,A.K., Martienssen,R.A., Barker,R.F. and Jarvis,M.G. (1987) *Mol. Gen. Genet.* **209**, 33–40.
47. Gebhardt,C., Oliver,J.E., Forde,B.G., Saarelainen,R. and Miflin,B.J. (l986) *EMBO J.* **5**, 1429–1435.
48. Broglie,R., Bellemare,G., Bartlett,S.G., Chua,N.-H. and Cashmore,A.R. (1981) *Proc. Natl. Acad. Sci. USA.* **78**, 73041–7308.
49. Bennetzen,J.L., Swanson,J., Taylor,W.C. and Freeling,M. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 4125–4128.
50. Lycett,G.W., Delauney,A.J., Zhao,W., Gatehouse,J.A., Croy, R.R.D. and Boulter,D. (1984) *Plant Mol. Biol.* **3**, 91–96.
51. Pintor-Toro,J.A., Langridge,P. and,Felix,G. (1982) *Nucl. Acids Res.* **10**, 3845–3860.
52. Hershey,H.P., Colbert,J.T., Lissemore,J.L., Barker,R.F. and Quail,P.H. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 2332–2336.