Two Classes of Genes in Plants

Nicolas Carels*,† and Giorgio Bernardi*,†

**Laboratoire de Génétique Moléculaire, Institut Jacques Monod, F-75005 Paris, France and †Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, I-80121 Napoli, Italy*

Manuscript received August 30, 1999
Accepted for publication December 6, 1999

ABSTRACT

Two classes of genes were identified in three Gramineae (maize, rice, barley) and six dicots (Arabidopsis, soybean, pea, tobacco, tomato, potato). One class, the GC-rich class, contained genes with no, or few, short introns. In contrast, the GC-poor class contained genes with numerous, long introns. The similarity of the properties of each class, as present in the genomes of maize and Arabidopsis, is particularly remarkable in view of the fact that these plants exhibit large differences in genome size, average intron size, and DNA base composition. The functional relevance of the two classes of genes is stressed by (1) the conservation in homologous genes from maize and Arabidopsis not only of the number of introns and of their positions, but also of the relative size of concatenated introns; and (2) the existence of two similar classes of genes in vertebrates; interestingly, the differences in intron sizes and numbers in genes from the GC-poor and GC-rich classes are much more striking in plants than in vertebrates.

EUKARYOTIC genomes cover a large spectrum of haploid sizes [or C values; for angiosperms see Bennett and Smith (1991) and http://www.rbgkew.org.uk/cval/database1.html]. It has been realized for quite some time that, except for the case of polyploidy, the variation in genome size is essentially due to changes in the amount of noncoding (mainly intergenic) sequences (see Cavalier-Smith 1985, for a review).

In the case of vertebrates, this was demonstrated by showing that the very small genome (400 Mb) of *Arothron diadematus*, a fish belonging to the order *Tetraodontiformes*, is characterized by very small amounts of highly and moderately repetitive sequences (Pizon *et al.* 1984). Since *Tetraodontiformes* is a very recent fish order derived from *Perciformes*, which is characterized by genome sizes two to three times as large (Bernardi and Bernardi 1990), there can be little doubt that the small C value of *Tetraodontiformes* is the result of a genome contraction.

As far as introns are concerned, it was shown (Brenner *et al.* 1993; Mason *et al.* 1995) that genes from the small genome of *Fugu rubripes*, another fish belonging to the order *Tetraodontiformes*, are characterized by shorter introns compared to human genes; incidentally, this is also the case for avian genes compared to human genes, as expected from the smaller genome size of birds compared to mammals (Hughes and Hughes 1995). A comparison of homologous genes from Fugu and human showed that size differences are especially large for human GC-poor (GC is the molar ratio of guanine + cytosine in DNA) compared to human GC-rich genes (Bernardi 1995), a point recently confirmed on a larger set of genes (Villard *et al.* 1998). Since GC-poor genes from vertebrates are characterized by longer introns compared to GC-rich genes (Duret *et al.* 1995), these results indicate that GC-poor genes are subject to more extensive contraction/expansion phenomena compared to GC-rich genes.

In the case of plants, phenomena of genome contraction/expansion are also known to occur. To cite just one example, a wide range of genome sizes is known in Gramineae, C values ranging from 415 Mb in the case of *Oryza sativa* to 12,600 Mb in the case of *Avena sativa*. However, no investigation has been reported so far on intron sizes as related to GC richness.

In this work, we explored plant genomes from three Gramineae (maize, rice, barley) and six dicots (Arabidopsis, soybean, pea, tobacco, tomato, potato) to see whether intron size and GC richness are correlated with each other in these genomes, as is the case for the genomes of vertebrates. The plants studied were chosen so as to explore genomes characterized by two different compositional situations. Indeed, the genomes of Gramineae are GC rich and their coding sequences cover a broad compositional range, whereas the genomes of the dicots studied are GC poor and their coding sequences cover a narrow GC range (Salinas *et al.* 1988; Matassi *et al.* 1989; Carels *et al.* 1998).

MATERIALS AND METHODS

Using the Infobiogen server (see http://www.infobiogen.fr), we extracted genomic DNA sequences encompassing complete genes from angiosperms (excluding seed-storage protein genes) from release 108 (August 1998) of GenBank with the

Corresponding author: Giorgio Bernardi, Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Napoli, Italy. E-mail: bernardi@alpha.szn.it

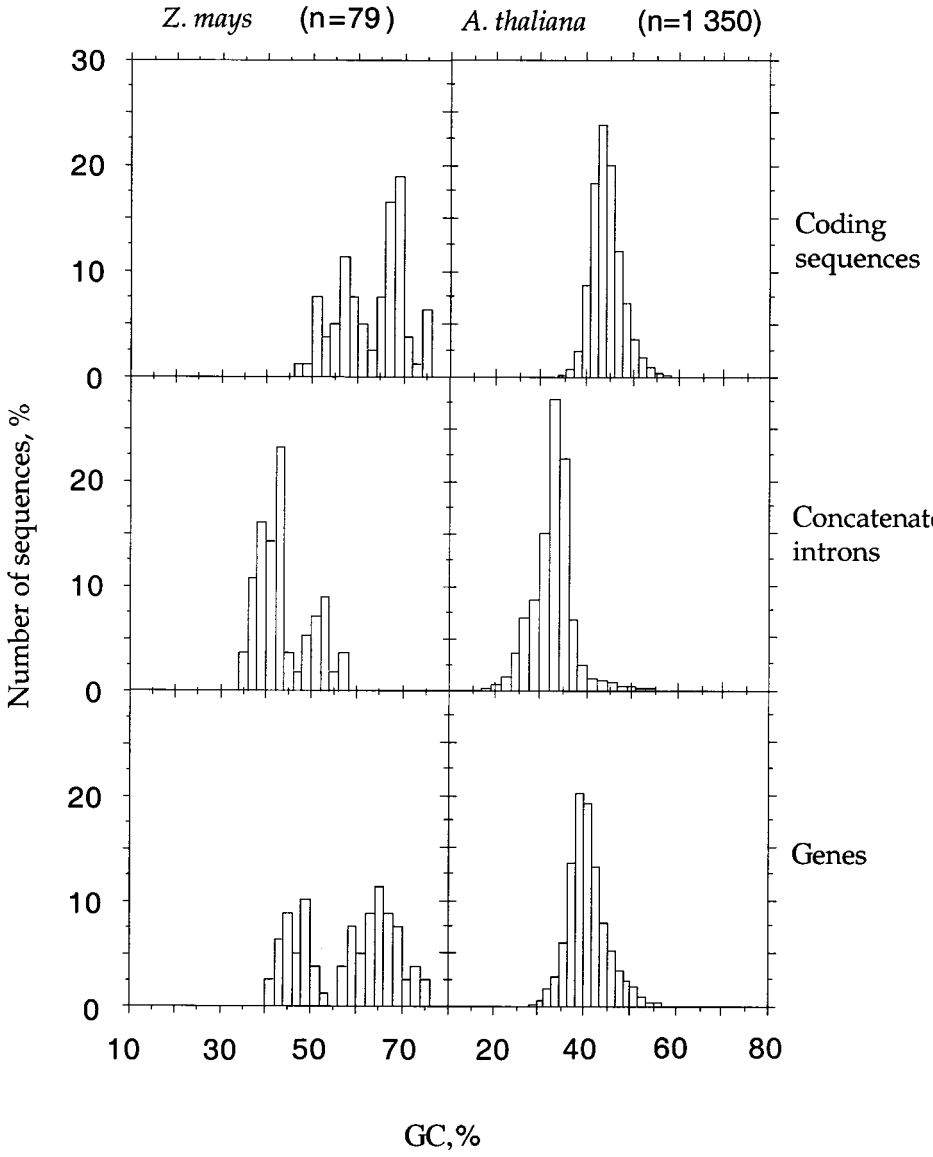

Figure 1.—Compositional distribution of coding sequences, concatenated introns, and genes in maize and Arabidopsis. *n* is the number of genes investigated.

ACNUC/QUERY retrieval system (Gouy *et al.* 1985). The sequences were then analyzed for their exon and intron sizes, as well as for their base compositions, the latter being determined using ANALSEQ (Gautier and Jacobzone 1989). In *Arabidopsis thaliana*, complete sequences were also extracted from contig sequences in GenBank (release 111, March 1999). Sequences with self-contained CDS (coding sequence) fields (features) were retained if they had a total coding sequence length > 220 bp. ATG and stop codons implied by the feature specifications were checked, and the corresponding "complete" coding sequences were also checked to ensure that they contained an integral number of codons. Sequences without ATG or stop codons in the expected positions were not included in the analysis. Likewise, pseudogenes were excluded. To allow automation, only introns interrupting the coding sequences of genes were investigated. Obvious gene redundancies were eliminated on the basis of identity of sequence sizes and similarity of descriptions and base composition (within 2% GC) in the three codon positions.

Homologous gene pairs were obtained as previously described (Carels *et al.* 1998), except for maize and Arabidopsis in which the nonredundant sample of maize genes was used to recover the orthologous coding sequences from Arabidopsis using TFASTX (Pearson *et al.* 1997). The orthology of gene pairs was estimated as described in Carels *et al.* (1998).

Genes available for each organism were ordered according to GC levels and divided into two classes by taking as the cutting point the mode of their distribution (unless two classes were already obvious, as in the case of Gramineae).

All GC-poor and GC-rich genes were analyzed for size, number, and GC level of exons, introns, and coding sequences. The statistical significance of the differences was analyzed using the Student's test (Student 1908, 1925) at $\alpha = 0.05$.

RESULTS

The compositional distribution of genes from maize and other Gramineae: The compositional distribution of coding sequences of maize is very broad, 45–75% GC and at least bimodal (Figure 1; see also Salinas *et al.* 1988; Matassi *et al.* 1989; Carels *et al.* 1995, 1998). A clear bimodality was found in introns and in genes (exons + introns). In the case of introns, a major peak was

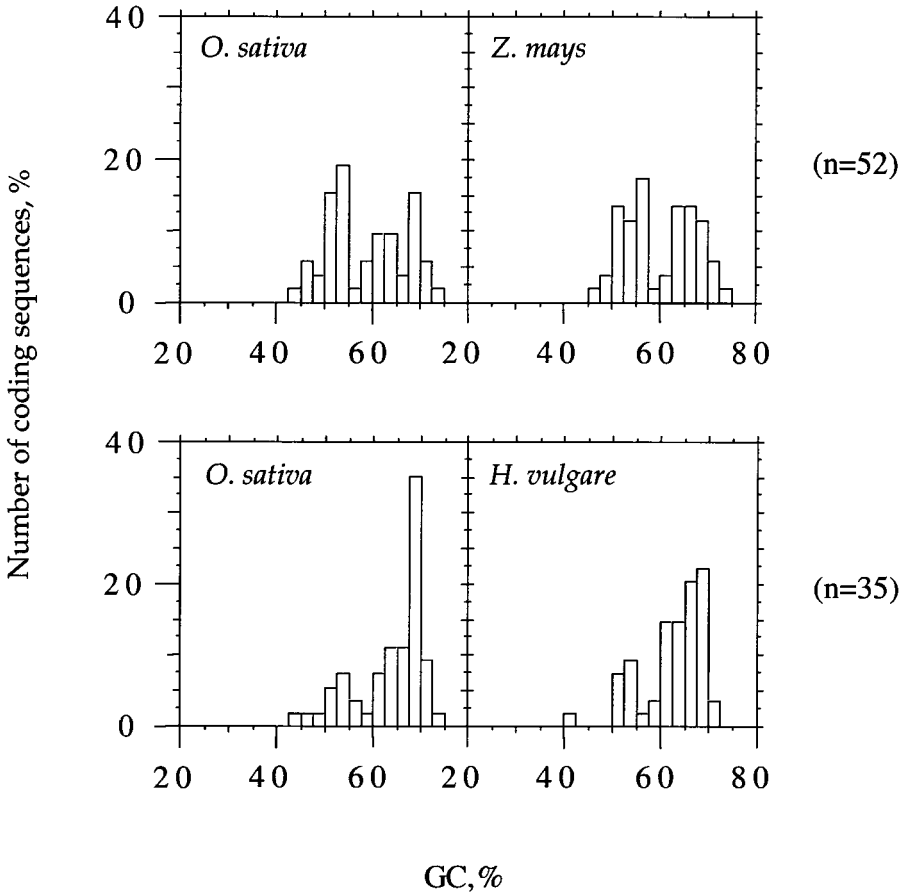

Figure 2.—Compositional distribution of coding sequences from rice and of their homologs from maize or barley. *n* is the number of genes investigated.

centered at ~40% GC, a minor peak at 53% GC. In the case of genes, GC-poor and GC-rich genes covered the 40–55% and the 55–75% GC range, respectively. Similar features were found in the other two Gramineae that could be tested, rice and barley, by analyzing homologous genes (Figure 2). Not surprisingly, slight differences were found in the GC level of the boundary between the two classes of maize genes, as derived from the different (and rather small) gene sample investigated. For instance, the boundary of coding sequences from the gene (exons + introns) sample of Figure 1 is at 62–65% GC, that from the genes homologous to rice genes (Figure 2) is at 57–60% GC.

The two classes of maize genes are characterized by distinct features, which are presented in Figure 3 and Table 1 and can be summarized as follows. GC-rich genes contain exons that are short relative to exons of vertebrates, no introns in 43% of the cases, and very few, short introns in those genes that contain them. In contrast, GC-poor genes contain even shorter exons and more numerous, longer introns. Interestingly, the difference in average exon size between the two classes of genes is not accompanied by a significant difference in average size of whole coding sequences (see below). Needless to say, the features just described for GC-rich and GC-poor genes account for the bimodality of the compositional distribution of maize genes. Moreover, in GC-rich genes, GC levels of exons and introns are 67 and 48%, respectively, whereas in GC-poor genes, exons and introns exhibit GC levels of 56 and 40%, respectively. The GC levels of GC-rich genes were also found to be significantly higher than those of GC-poor genes by as much as 5, 9, and 20% for first, second, and third codon positions, respectively (see Table 2). These large compositional differences of introns and exons are typical of plants, the corresponding differences in vertebrate genes being much smaller (Carels *et al.* 1998).

The compositional distribution of genes from Arabidopsis and other dicots: In contrast with maize, coding sequences, introns, and genes from Arabidopsis are characterized (Figure 1) by unimodal distributions and by smaller compositional differences of exons (45 and 49%) and introns (31 and 33%) in GC-poor and GC-rich genes, respectively (Table 1; see also Salinas *et al.* 1988; Matassi *et al.* 1989; Barakat *et al.* 1998; Carels *et al.* 1998). Higher GC levels in the three codon positions were, however, also observed for GC-rich genes relative to GC-poor genes of Arabidopsis, although to a lesser degree compared to maize genes (Table 2). Indeed, the differences were found to be of only 2, 3, and 4%, on the average, in first, second, and third codon positions, respectively.

The other distinguishing features between the two classes are similar to those described above for maize

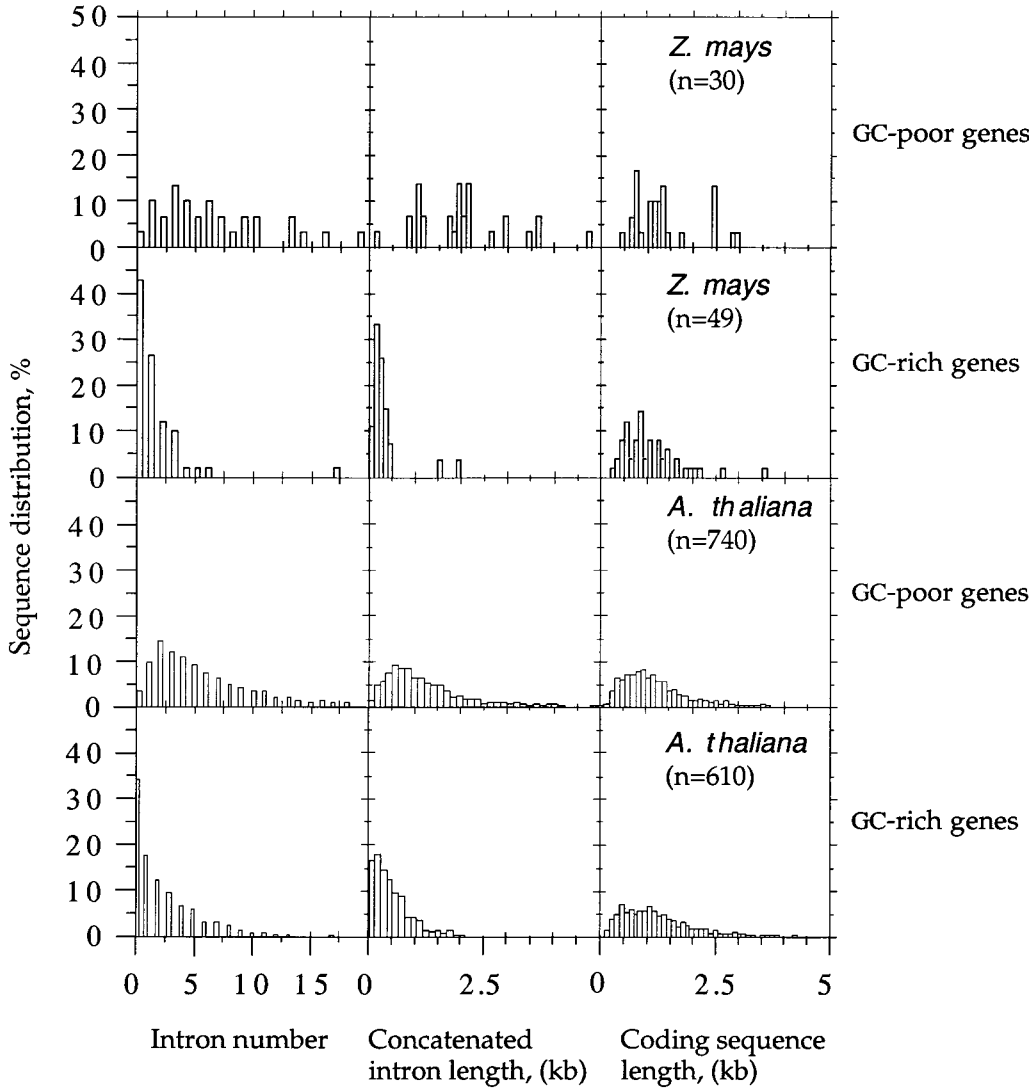

Figure 3.—Distributions of intron number, length of concatenated introns, and coding sequence length in GC-poor and GC-rich genes of maize and Arabidopsis. *n* is the number of genes investigated. An extremely small number of very long concatenated introns in the GC-poor genes from Arabidopsis are not represented.

(see Table 1 and Figure 3). Indeed, the proportion of intron-less genes in GC-rich genes of Arabidopsis (38%) is close to that of maize (43%), and intron numbers are also similar in both classes of genes in the two genomes. The smaller intron sizes in GC-poor genes (but not in GC-rich genes) were to be expected because of the much smaller genome size of Arabidopsis (~120 Mb) compared to maize (~2500 Mb).

The features described for Arabidopsis genes were also found in the genes of other dicots, soybean, tobacco, tomato, and potato (see Table 1), with minor differences, probably due to differences in the gene samples.

As far as homologous genes from Arabidopsis and maize are concerned, intron number and the size of concatenated introns are correlated with coefficients of 0.96 and 0.68, respectively (Figure 4). In contrast, the GC levels of concatenated introns of these homologous genes are not correlated (data not shown). This is not surprising in view of the very narrow compositional distribution of introns in Arabidopsis genes (Carels *et al.* 1998). Finally, as already mentioned, the lower GC level of introns compared to exons appears to be a general feature of plant genes, the difference between exons and introns being, however, larger for GC-rich genes, especially in Gramineae.

Statistical analysis: All data of Table 1 were found to be significantly different between the two classes of genes, except for coding sequence size and GC level of concatenated introns. In the case of coding sequence size, the Student's test indicated no significant difference whatever the species under consideration. Moreover, exons were significantly shorter, on the average, in GC-poor genes compared to GC-rich genes in all angiosperms tested (Table 1). Differences in GC levels between GC-poor and GC-rich genes were also found to be statistically significant for maize as well as Arabidopsis when individual codon positions were compared (see Table 2). The GC levels of concatenated introns were found to be different between the two classes of genes from Gramineae, but the situation was less clear in dicots. In tobacco and tomato, GC levels of concatenated introns were found to be different on the average, but this was not the case in the other dicots. This situation

TABLE 1

Some features of genes in angiosperms

	Gramineae			Brassicaceae:	Fabaceae		Solanaceae		
	Maize	Rice	Barley	Arabidopsis	Soybean	Pea	Tobacco	Tomato	Potato
GC-poor genes									
Coding sequences									
Size, kb	1.3	1.2	1.0	1.4	1.2	1.1	1.0	1.2	0.9
GC, %	55.8	54.2	55.2	43.4	45.8	41.9	41.8	41.7	39.9
Exons									
No. per gene	7.5	5.3	4.3	7.6	4.5	5.1	3.3	4.5	4.0
Size, kb	0.3	0.3	0.3	0.2	0.3	0.3	0.3	0.3	0.3
Introns									
No. per gene	6.5	4.3	3.3	6.6	3.5	4.1	2.3	3.5	3.0
No. per gene with introns	6.7	4.8	3.7	6.7	3.7	4.5	2.5	3.7	3.8
Size, kb	0.5	0.4	0.6	0.4	0.4	0.5	0.6	0.5	0.3
Size per gene, kb	2.0	1.4	1.4	1.8	1.1	1.3	1.2	1.5	1.2
GC, %	39.8	35.6	36.9	31.4	27.1	26.1	30.1	28.0	27.1
Sample size	30	54	28	740	37	24	32	42	39
GC-rich genes									
Coding sequences									
Size, kb	1.0	1.0	1.0	1.4	0.9	1.0	1.1	1.2	1.1
GC, %	67.4	67	66	46.6	49.1	44.9	47.4	47.2	46.4
Exons									
No. per gene	2.5	2.0	2.2	3.2	1.4	2.6	1.6	2.3	3.2
Size, kb	0.6	0.6	0.5	0.7	0.7	0.5	1.0	0.8	0.7
Introns									
No. per gene	1.5	1.0	1.2	2.2	0.4	1.6	0.6	1.3	2.2
No. per gene with introns	2.6	1.9	1.9	3.3	1.6	2.7	2.0	2.8	4.6
Size, kb	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.2
Size per gene, kb	0.3	0.3	0.3	0.3	0.3	0.4	0.5	0.5	0.6
GC, %	48.0	40.5	44.7	33.5	30.0	29.3	32.7	30.8	32.5
Sample size	49	36	45	610	48	22	44	31	33

is probably due to several factors, such as the low levels of genome heterogeneity and rather small sample sizes.

DISCUSSION

The main finding of this work is the identification in plants of two classes of genes, which were first observed in maize (Carels *et al.* 1998). In the case of maize, the compositional distribution of genes is strikingly bimodal, and the two classes of genes, GC-rich and GC-poor, are very distinct, in that the former exhibit either no introns or very few, short introns, whereas the latter are characterized by many long introns. In maize, the two classes differ by 12% GC in their coding sequences and by a 10-fold factor in intron size. Similar distinctive features were also found not only in the other two Gram-

TABLE 2

Average GC composition in the three codon positions of genes from maize and Arabidopsis

Species	*n*	$\overline{GC_1}$	$\overline{GC_2}$	$\overline{GC_3}$
Maize				
GC poor	49	57.7 (2.85)	41.0 (3.90)	68.6 (9.15)
GC rich	30	63.1 (8.22)	50.1 (10.74)	89.0 (10.58)
Difference		5.4	9.1	20.4
Arabidopsis				
GC poor	740	49.8 (4.35)	39.0 (4.55)	41.2 (5.42)
GC rich	610	52.0 (4.70)	42.1 (6.16)	45.6 (6.81)
Difference		2.2	3.1	4.4

n is the number of sequences analyzed in each compositional class studied; $\overline{GC_1}$, $\overline{GC_2}$, and $\overline{GC_3}$ are the average GC levels; the standard deviations in the three codon positions are given in parentheses.

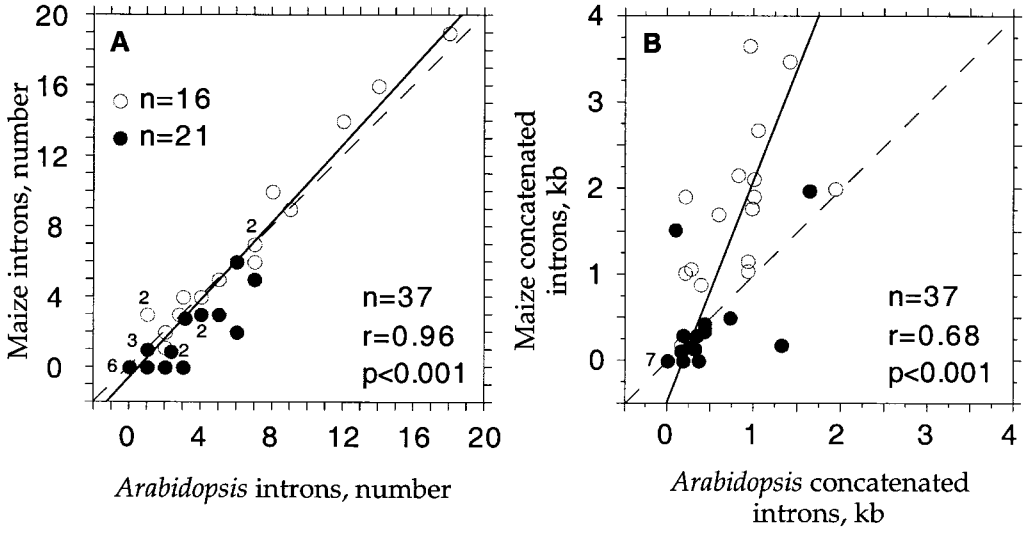

Figure 4.—Number of introns and length of concatenated introns in homologous genes of maize and Arabidopsis. GC-rich genes are represented with solid circles and GC-poor genes with open circles. Numbers close to circles are the number of genes represented. The solid line is the orthogonal regression line through the points, and the broken line is the diagonal.

ineae, rice and barley, for which reasonably large sequence databases were available, but also in dicots, which are characterized, in contrast to Gramineae, by a very narrow compositional distribution of genes. The most striking case is that of Arabidopsis, which is characterized by a very compact genome. Even in this case, in which the two classes differ by only 3% GC in their coding sequences, intron sizes show a 4-fold difference.

The fact that genes are, on the average, interrupted by either a large number of long introns or a small number of short introns and that GC levels are different in the two classes of genes is far from trivial for two reasons. First, these properties are found not only in angiosperms, but also in vertebrates (Duret *et al.* 1995) and might, therefore, be quite general properties of the genomes of multicellular eukaryotes. Second, total intron sizes in homologous genes from maize and Arabidopsis are correlated, as shown in Figure 4B. In other words, the Arabidopsis genes that are homologous to the genes having large total intron size in maize also are endowed with large total intron sizes. Therefore, in homologous genes, not only the number of introns (Figure 4A) is conserved, as expected, but also the total size of introns.

Moreover, GC-rich genes were observed to be, on the average, richer in GC in all codon positions in all species tested, but especially in Gramineae, compared to GC-poor genes (Table 2). GC differences in second codon positions obviously have repercussions on the amino acid composition of the encoded proteins, a subject currently under investigation.

As far as the functional meaning of the two classes of genes is concerned, we would like to speculate that since housekeeping genes were found to be associated with GC-rich genes not only in Arabidopsis and maize (Chiapello *et al.* 1998), but also in vertebrates (see Bernardi 1995, for a review), shortage and small sizes of introns might be viewed as advantageous features for genes that are transcribed in a constitutive or at least in an extensive way.

In the case of GC-poor genes, which are largely tissue specific in vertebrates (Bickmore and Craig 1997), the abundance and size of introns in these genes would be favorable for alternative splicing, an important mechanism of expression regulation of tissue-specific genes (Bell *et al.* 1998).

The phenomenon of intron depletion described here, which accompanies the increase in GC content of genes, was already reported in vertebrates (Duret *et al.* 1995), but it appears to be more general. In fact, the phenomenon is more striking in plants than in vertebrates: a comparison of the present data with those for vertebrates shows that in plant genes the contrast between intron lengths in GC-poor and GC-rich genes is at least 15–20% larger than that in vertebrates.

We thank O. Clay for technical help and useful discussions.

LITERATURE CITED

Barakat, A., G. Matassi and G. Bernardi, 1998 Distribution of genes in the genome of Arabidopsis thaliana and its implications for the genome organization of plants. Proc. Natl. Acad. Sci. USA **95:** 10044–10049.

Bell, M. V., A. E. Cowper, M. P. Lefranc, J. I. Bell and G. R. Screaton, 1998 Influence of intron length on alternative splicing of CD44. Mol. Cell. Biol. **18:** 5930–5941.

Bennett, M. D., and J. B. Smith, 1991 Nuclear DNA amounts in angiosperms. Philos. Trans. R. Soc. Lond. B Biol. Sci. **334:** 309–345.

Bernardi, G., 1995 The human genome: organization and evolutionary history. Annu. Rev. Genet. **29:** 445–476.

Bernardi, G., and G. Bernardi, 1990 Compositional patterns in the nuclear genome of cold-blooded vertebrates. J. Mol. Evol. **31:** 265–281.

Bickmore, W., and J. Craig, 1997 *Chromosome Bands: Patterns in the Genome.* Springer, New York.

Brenner, S., G. Elgar, R. Sandford, A. Macrae, B. Venkatesh *et al.*, 1993 Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature **366:** 265–268.

Carels, N., A. Barakat and G. Bernardi, 1995 The gene distribu-

tion of the maize genome. Proc. Natl. Acad. Sci. USA **92:** 11057–11060.

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

Cavalier-Smith, T., 1985 Eukaryote gene numbers, non-coding DNA and genome size, pp. 69–103 in *The Evolution of Genome Size*, edited by T. Cavalier-Smith. Wiley, London.

Chiapello, H., F. Lisacek, M. Caboche and A. Hénaut, 1998 Codon usage and gene function are related in sequences of Arabidopsis thaliana. Gene **209:** GC1–GC38.

Duret, L., D. Mouchiroud and C. Gautier, 1995 Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. **40:** 308–317.

Gautier, C., and M. Jacobzone, 1989 <http,//biom3.univ-lyon1.fr,8080/doclogi/docanals/manuel.html>, Publication interne, UMR CNRS 5558 Biometrie, Genetique et Biologie des Populations, Universite Claude Bernard, Lyon I, France.

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

Hughes, A. L., and M. K. Hughes, 1995 Small genomes for better flyers. Nature **377:** 391.

Mason, P. J., D. J. Stevens, L. Luzzatto, S. Brenner and S. Aparicio, 1995 Genomic structure and sequence of the Fugu rubripes glucose-6-phosphate dehydrogenase gene (G6PD). Genomics **26:** 587–591.

Matassi, G., L. M. Montero, J. Salinas and G. Bernardi, 1989 The isochores organisation and compositional distribution of homologous coding sequences in the nuclear genome of plants. Nucleic Acids Res. **17:** 5273–5290.

Pearson, W. R., T. Wood, Z. Zhang and W. Miller, 1997 Comparison of DNA sequences with protein sequences. Genomics **46:** 24–36.

Pizon, V., G. Cuny and G. Bernardi, 1984 Nucleotide sequence organization in the very small genome of a tetraodontid fish, Arothron diadematus. Evol. J. Biochem. **140:** 25–30.

Salinas, J., G. Matassi, L. M. Montero and G. Bernardi, 1988 Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucleic Acids Res. **16:** 4269–4285.

Student, 1908 The probable error of a mean. Biometrika **6:** 1–25.

Student, 1925 New tables for testing the significance of observations. Metron **5:** 105–120.

Villard, L., F. Tassone, T. Crnogorac-Jurcevic, K. Clancy and K. Gardiner, 1998 Analysis of pufferfish homologues of the AT-rich human APP gene. Gene **210:** 17–24.

Communicating editor: S. Yokoyama