J Mol Evol (1998) 46:45–53

JOURNAL OF MOLECULAR EVOLUTION

Compositional Properties of Homologous Coding Sequences from Plants

Nicolas Carels, Pascal Hatey, Kamel Jabbari, Giorgio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

Received: 24 June 1997 / Accepted: 20 August 1997

Abstract. In this work, we investigated (1) the compositional distributions of all available nuclear coding sequences (and of their three codon positions) of six dicots and four Gramineae; this considerably expanded our knowledge about the differences previously seen between these two groups of plants; (2) the compositional correlations of homologous genes from dicots and from Gramineae, as well as from both groups; all correlations were characterized by very good coefficients, with slopes close to unity in the former two cases and very high in the last; (3) the compositional transition that accompanied the emergence of Gramineae from an ancestral monocot; (4) the compositional correlations between exons and introns, which were very good in Gramineae, but only poor to good in dicots; and (5) the compositional profiles of homologous genes from angiosperms, which were characterized by a series of peaks (exons) and valleys (introns) separated by 15–20% GC. The conservative and transitional modes of compositional evolution in plant genes and their general implications are discussed.

Key words: Genes — Genomes — Angiosperms

Introduction

A compositional approach to the study of plant genomes (Salinas et al. 1988) showed (1) that the nuclear genomes of angiosperms are characterized by a compositional compartmentalization and (2) that the nuclear genomes of Gramineae exhibit compositional patterns that are different from those of the dicots investigated in that they have a higher average GC level (GC is the molar fraction of guanine + cytosine in DNA) and cover a wider GC range; the compositional distributions of coding sequences paralleled those of DNA molecules in that it is narrow, symmetrical, and centered around 46% GC in the dicots studied, whereas it is broad and asymmetrical, with the majority of sequences between 60% and 70% GC in the case of Gramineae; in both cases, introns exhibited remarkably lower GC levels compared to exons from the same genes.

Subsequent work (Matassi et al. 1989) demonstrated that the compositional differences found in exons and introns of dicots and Gramineae were also found when homologous genes were examined. Even if the gene sample available at that time was very small, this seemed to rule out the unlikely possibility that the differences previously seen were due to differences in the gene samples under comparison. Moreover, it was shown that the compositional DNA patterns of the genomes of at least one dicot (*Oenothera*) and of some monocots (e.g., *Allium*) overlap with those of Gramineae and of the dicots previously studied, respectively.

When a number of genes were localized in compositional fractions of plant genomes, it was observed (Montero et al. 1990) that the few genes tested were found in DNA fragments (50–100 kb in size) only covering a narrow GC range, 2.3–2.6%, in the case of pea, maize, and wheat. More recently, detailed investigations showed that in maize (Carels et al. 1995) and in other Gramineae (Barakat et al. 1997) most protein-encoding genes are present in DNA fragments only covering a

Abbreviations: GC, molar fraction of guanine + cytosine in DNA

Correspondence to: G. Bernardi

0.8–1.6% GC range and corresponding to 12–24% of the genomes, according to the species under consideration. The genome compartments, where the vast majority of genes were found, were scattered over all chromosomes and were collectively called the gene space, some seed storage protein genes and ribosomal genes occupying, however, compositional compartments characterized by lower and higher GC levels, respectively. The narrow compositional range of the gene space of maize which comprises genes (exons + introns) ranging from 40 to 75% GC was accounted for (Barakat et al. 1997) by the fact that genes and/or gene clusters are embedded in retroposons (SanMiguel et al. 1996) having a very narrow GC range.

In the present work, we have compared the compositional properties of coding sequences (and of their codon positions) from six dicots and four Gramineae. The former comprised *Arabidopsis thaliana* and *Glycine max* (soybean) from Brassicaceae, *Pisum sativum* (pea) from Fabaceae, *Nicotiana tabacum* (tobacco), *Lycopersicon esculentum* (tomato), and *Solanum tuberosum* (potato) from Solanaceae; the latter *Zea mays* (maize), *Oryza sativa* (rice), *Triticum aestivum* (wheat) and *Hordeum vulgare* (barley) from Poaceae (Gramineae). We have then studied (1) the compositional correlations of homologous genes from dicots and from Gramineae, as well as from both groups, the latter revealing a compositional transition; (2) the compositional correlations between exons and introns of both dicots and Gramineae; and (3) the compositional profiles of homologous genes from angiosperms.

Materials and Methods

Coding sequences from the angiosperms represented by the largest numbers of sequences (but excluding seed storage protein genes) were extracted from GenBank Release 94.0 (04/96) using the program ACNUC (Gouy et al. 1985). The orthology of homologous gene pairs was confirmed or rejected largely on the basis of the sequence descriptions of the gene family members, and redundancies were recognized by their descriptions, sequence lengths, and base compositions in the three codon positions. As an alternative method of selecting orthologous genes, we used the FASTA program (Pearson and Lipman 1988) from the GCG package (Devereux et al. 1984) to search for homologous pairs among the amino acid sequences. Pairs of coding sequences which were used if they were similar over at least 75% of their length and for which the corresponding amino acid sequences were at least 40% identical (Doolittle 1987) were used. For a few comparisons, pairs of sequences with amino acid identities as low as 30% were also used to insure a sufficiently large sample. A risk inherent to this method is that of occasionally retaining homologous sequences that are paralogous rather than orthologous, which could lead in some cases to an underestimate of correlation coefficients. Genes coding for seed storage proteins were not included in the FASTA comparison and were treated separately. We also rejected redundant genes that could be detected on the basis of their sequence length identity and/or their GC percentage similarity in each codon position. The program ANALSEQ (Gautier and Jacobzone 1989) was used to determine the base compositions of coding sequences and of their first, second, and third codon positions. Orthogonal regressions were calculated as described by Zoubak et al. (1996); they minimize the sums of squares of the orthogonal distances rather than that of the vertical distances, and provide, therefore, a more appropriate representation of a scatterplot than the linear regression (see D'Onofrio et al. 1991; Harvey and Pagel 1991; Clay et al. 1996).

The identification of homologous gene using FASTA appears to give reliable results, even if in some cases an increased scatter in the data may be due to the unavoidable inclusion of a few paralogous genes among the orthologous genes targeted by this method. Its advantages are that it is fully automatable; that it is independent of the gene descriptions in the nucleotide sequence databases, which are not always reliable; and that it yields a larger sample of homologous gene pairs, allowing reliable conclusions to be drawn already for species with a small sequence database.

Detailed information on the sequences used is available upon request.

Results

Compositional Distributions of Coding Sequences

Table 1 shows the compositional features of the nuclear coding sequences of the plants studied. They quantify the higher GC levels and larger spread of the coding sequences of Gramineae compared to the dicots investigated. Similar differences could be seen in GC_1, GC_2, and especially in GC_3 values. Figures 1 and 2 show the compositional distributions of first, second, and third codon positions and coding sequences from the genomes of the six dicots and of the four Gramineae investigated here, respectively. Figure 2 also shows the compositional distributions of *Arabidopsis* genes for the sake of comparison.

GC levels of coding sequences in dicots are centered in the 44–47% range, while those of Gramineae are centered in the 59–61% GC range. In addition, the coding sequences of dicots show a narrow compositional distribution whereas those of Gramineae show a broad, multimodal, and asymmetrical distribution.

The frequency distribution of GC_2 and GC_1 values for dicots have mean values of 40–42% GC and 51–53% GC, respectively, whereas the corresponding mean values in the case of the Gramineae are 44–47% and 57–60%, respectively. The frequency distribution of GC_3 levels is broader than those of GC_1, GC_2, and GC_S, both in dicots, where mean GC_3 values range from 39% to 47%, and in Gramineae, where they range from 73% to 78%. It is known from previous work (Salinas et al. 1988) that the compositional distribution of zein genes is centered on low GC levels.

Correlations Between Orthologous Genes

The differences in the compositional distributions of dicots and Gramineae mentioned in the preceding section are not due to differences in the gene samples (1) because the results of Figs. 3 and 4 show that very good compositional correlations hold between orthologous coding se-

Table 1. Compositional features of coding sequences in some plant species[a]

Families	Species	*N*	$\overline{GC}_1$	σ_1	$\overline{GC}_2$	σ_2	$\overline{GC}_3$	σ_3	$\overline{GC}_s$	σ_s
Gramineae										
	Z. mays	257	59.7	7.1	45.9	8.5	75.3	17.2	60.3	8.1
	T. aestivum	122	57.2	6.9	46.9	9.2	74.9	18.6	59.7	8.7
	H. vulgare	184	59.1	6.6	42.2	8.2	77.7	16.3	61.4	7.1
	O. sativa	208	58.1	5.8	44.2	8.4	73.1	19.0	58.5	8.6
	Average		58.7	6.6	44.7	8.5	75.2	17.7	59.9	8.1
Brassicaceae										
	A. thaliana	797	52.6	4.8	41.2	6.1	45.4	6.9	46.4	3.4
	G. max	208	53.2	5.3	40.9	5.9	46.5	9.8	46.9	4.3
	Average		52.7	4.9	41.1	6.1	45.6	7.5	46.5	3.6
Fabaceae										
	P. sativum	202	51.5	5.3	41.2	5.5	39.2	7.0	44.0	3.3
Solanaceae										
	N. tabacum	243	51.6	6.1	42.1	7.2	40.4	6.9	44.7	3.9
	L. esculentum	235	51.7	6.7	43.8	10.4	39.8	8.2	44.9	5.1
	S. tuberosum	211	51.4	5.2	40.3	5.6	39.0	7.0	43.6	3.9
	Average		51.6	6.0	42.1	7.8	39.8	7.4	44.4	4.3

[a] *N* is the number of sequences investigated, $\overline{GC}_1$, $\overline{GC}_2$, $\overline{GC}_3$, $\overline{GC}_s$ are the average GC levels of first, second, and third codon positions and of the whole coding sequences, respectively. σ_1, σ_2, σ_3, σ_s are the corresponding standard deviations

quences and their first, second, and third codon positions from Solanaceae and Fabaceae (Figs. 3, 4); (2) because similar very good compositional correlations are found in comparisons concerning genes from different Gramineae, maize, rice, wheat, and barley (Fig. 5); and (3) because in both sets of data the regression lines coincide with the diagonal, indicating an essential identity of values for the genes under comparison; GC_3 values from pea genes tend, however, to be systematically lower compared to the corresponding values of *Arabidopsis*, whereas a number of GC_3 values from soybean genes are higher than those of *Arabidopsis* (Fig. 4).

The data of Fig. 5 also show the comparison of homologous coding sequences of *Arabidopsis* (as a representative of dicots) and maize (as a representative of Gramineae). While the regression line for the GC_2 comparison is close to the diagonal (with, however, a number of values slightly above it), the slope of the orthogonal regression line of GC_3 plots is around 4. This steep regression line shows that GC_3 values of maize genes are consistently higher than those of *Arabidopsis* genes and that increasingly GC-richer *Arabidopsis* genes correspond to points increasingly farther above the diagonal in maize. The regression line for the GC_S comparison has a slope of 2.6, a value intermediate between the slopes of GC_3 (4.3), GC_1 (1.6), and GC_2 (1.0). Results similar to those reported for maize were found for other Gramineae (not shown).

Correlations Between Exons and Introns

As shown in Fig. 6, no compositional correlation holds between exons and introns of *Arabidopsis*, both GC sets being comprised in a very narrow range, which is 15% lower in the case of introns than in that of exons. Other dicots exhibit, however, either weak or significant correlations. (The latter case is exemplified by the genes of pea). In the case of maize, introns and exons show a strong compositional correlation, the values for introns being about 20% lower than in that of exons.

Compositional Profiles of Plant Genes

Figure 7 shows a typical GC plot for an *Arabidopsis* gene and its maize homolog, which is characterized by an alternating series of peaks and valleys corresponding to the GC levels of exons and introns, respectively. Expectedly, intron positions are very largely conserved and intron valleys are deeper, on the average, in maize compared to *Arabidopsis*. Moreover, several maize introns are absent in *Arabidopsis*. Flanking sequences show 5′ and 3′ GC levels which are closer to those of introns than to those of exons.

Discussion

The Compositional Conservation of Orthologous Sequences of Dicots and Gramineae

The orthologous coding sequences of Fabaceae (pea) and Solanaceae (tobacco, tomato, potato) are compositionally conserved in all three codon positons (see Fig. 3). This also apply to the orthologous coding sequences from Gramineae (see Fig. 5). In other words, nucleotide changes were not accompanied by compositional changes in the genomes of these three families of angiosperms over the millions of years separating the species compared. (The separation between maize and wheat was estimated at 50–70 million years ago; Wolfe et al.

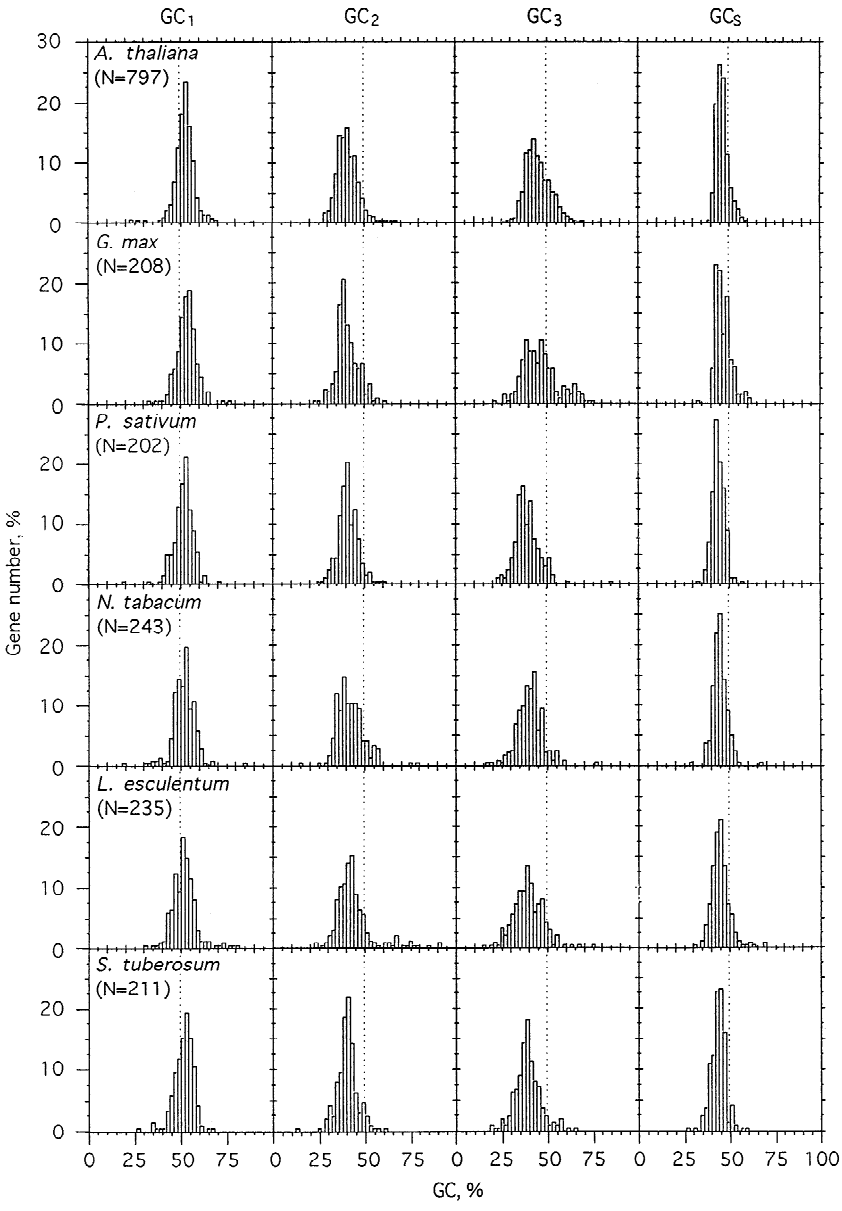

Fig. 1. Distributions of coding sequences from dicots according to the GC_1, GC_2, GC_3, and GC_S levels. GC_1, GC_2, GC_3, and GC_S are the GC levels of first, second, and third codon positions and of the whole coding sequences, respectively.

1989; see also Crane et al. 1995; Laroche et al. 1995). This is especially remarkable in the case of Gramineae since many GC_3 values are in the 80–100% range.

The Compositional Transition of Orthologous Sequences of Dicots and Gramineae

In contrast with the results just discussed, plots of orthologous sequences of maize and *Arabidopsis* indicate compositional changes which essentially affect GC_1 and GC_3. The conservation of GC_3 values in orthologous coding sequences of Gramineae and of dicots, respectively (see above), suggests that the compositional transition between orthologous sequences of Gramineae and dicots took place between the common ancestor of Gramineae and the corresponding dicot sister group. In fact, the existence of monocots, like onion, asparagus, *Scindapsus*, and *Typha* (which belong to four different families), that show a compositional DNA pattern (i.e., a CsCl profile) centered on GC values that are even lower than those of most dicots studied so far (Matassi et al. 1989), suggests that monocots exist which have lower GC_3 values in genes orthologous to the GC_3-richest genes of maize. Indeed, GC_3 values of plant genes are correlated with the modal buoyant densities of plant DNAs (Montero et al. 1990). The few genes which could be tested confirm this view. For example, since onion, a monocot which is not a Graminea, is 40% GC on the average in DNA and since its few available coding sequences are comprised between 40% and 55% GC_S and 35% and 70% GC_3, we can conclude that the compositional transition is specific to Gramineae and not to

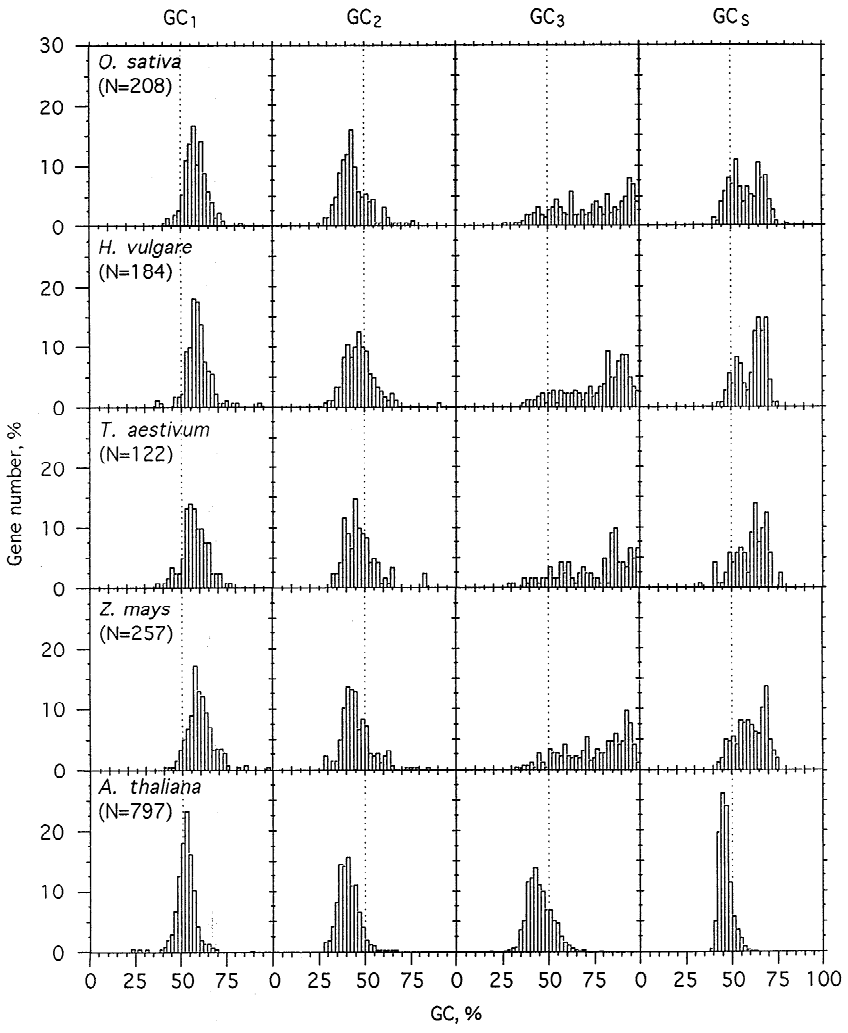

Fig. 2. Distributions of coding sequences from Gramineae according to the GC_1, GC_2, GC_3, and GC_S levels. The *Arabidopsis* distribution is shown in the *bottom frame* for the sake of comparison. Other indications as in Fig. 1.

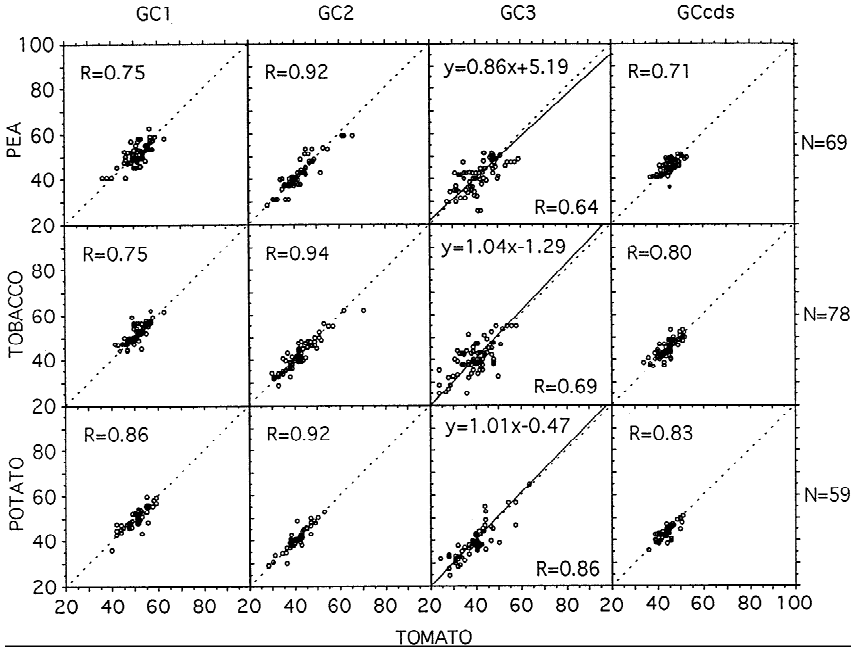

Fig. 3. GC_1, GC_2, GC_3, and GC_S of pea, tobacco, and potato genes (*ordinate*) are plotted against the corresponding values of their homologs from tomato (*abscissa*). Orthogonal regression lines (*solid*), diagonals (*dashed*), and correlation coefficients (*R*) are shown. Other indications as in Fig. 1.

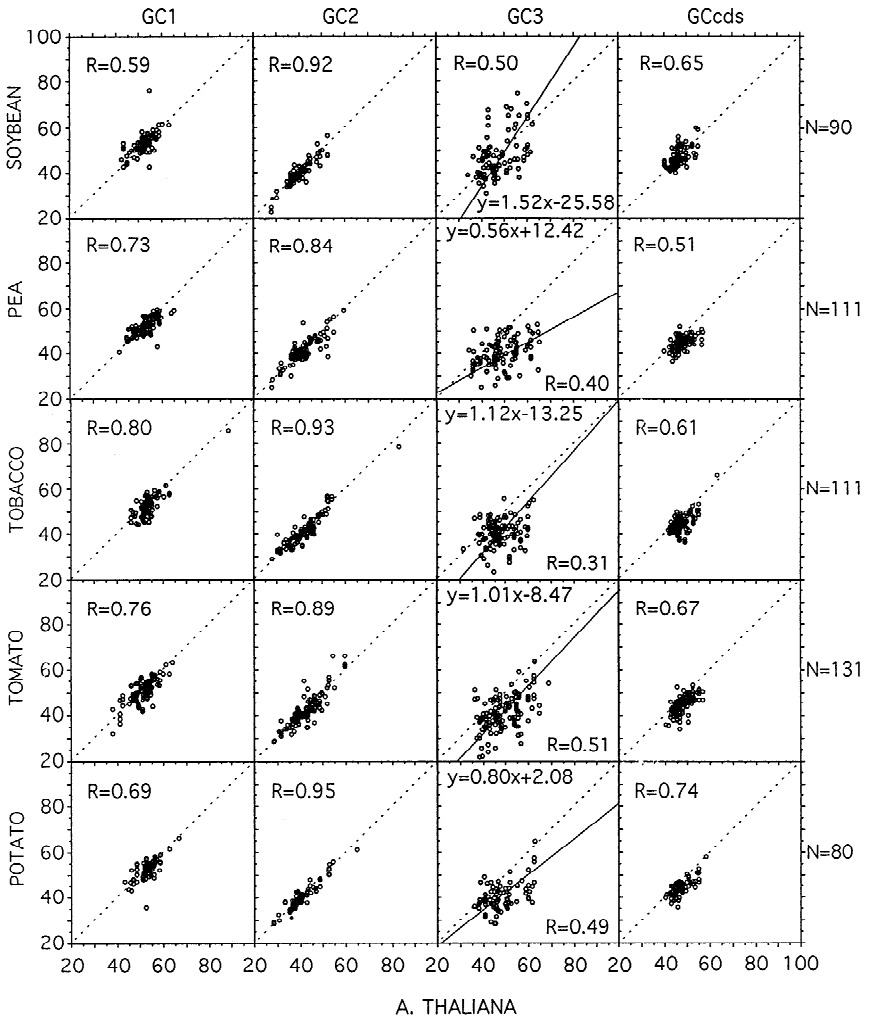

Fig. 4. GC_1, GC_2, GC_3, and GC_S of pea, tobacco, tomato, and potato genes (*ordinate*) are plotted against the corresponding values of their homologs from *Arabidopsis* (*abscissa*). Other indications as in Fig. 3.

monocots in general. Therefore, the compositional pattern of extant Gramineae is the result of a compositional transition that took place at the time of their emergence from monocots at the Upper Cretaceous, i.e., ~70 MYA (Crepet and Feldman 1991).

At present, however, we can make a detailed comparison of the compositional pattern of Gramineae only with that of dicots (and more specifically with *Arabidopsis*) because a large set of homologous coding sequences is only available for the latter.

It should be stressed that the GC_3 plot of maize vs *A. thaliana* is very strongly reminiscent of that previously found for man vs *Xenopus* (see Bernardi et al. 1997), the slope of the observed regression line being even higher (4.3 vs 2.7). In both cases, the GC_3-poorest genes are practically identical in GC_3, whereas the differences becomes increasingly larger for GC_3-rich genes.

It should be noticed that, next to the ''major transition'' between Gramineae and dicots just discussed, some ''minor transitions'' were found among dicots. For instance, a number of orthologous genes from soybean and pea show GC_3 values that do not match compositionally those of *Arabidopsis*. However, in the following, we will limit our discussion to the major transition undergone by Gramineae.

Two Modes of Compositional Evolution in Plant Genomes

The points made above show that, as in the case of vertebrate genomes (Bernardi et al. 1988), two compositional modes of evolution can be distinguished in the genomes of Gramineae: a transitional (or shifting) mode in which GC_3 changes took place in the ancestor of Gramineae, and a conservative one in which the GC_3 compositional patterns did not undergo any further changes in spite of the accumulation of point mutations. This conservative mode is also predominant in the dicots studied (except for some cases exhibiting minor shifts; see above).

As far as the two modes of evolution are concerned, two different viewpoints were put forward in the analogous case of vertebrates (see also Bernardi et al. 1997). The first one (Bernardi and Bernardi 1986; Bernardi et al. 1988) was that a selective advantage favored GC in-

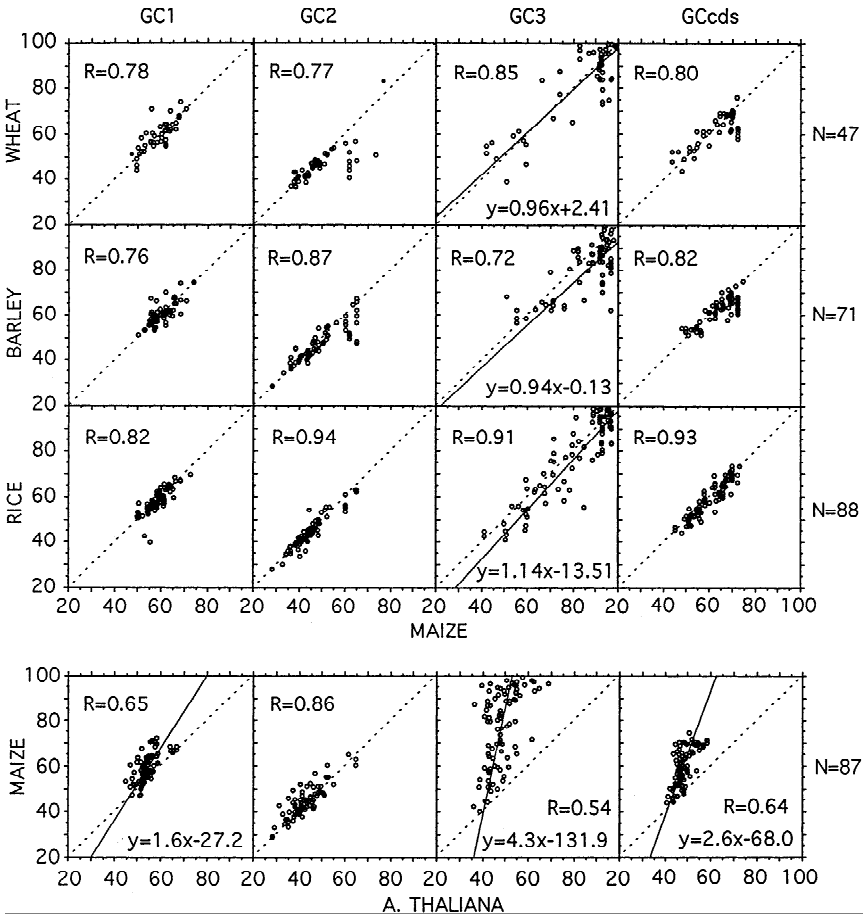

Fig. 5. GC_1, GC_2, GC_3, and GC_S values of wheat, barley, and rice genes (ordinate) are plotted against the corresponding values of their homologs from maize genes (*abscissa*). In the *bottom frame*, GC_1, GC_2, GC_3, and GC_S of maize genes (*ordinate*) are plotted against the corresponding values of their homologs from *Arabidopsis* genes (*abscissa*). Other indications as in Fig. 3.

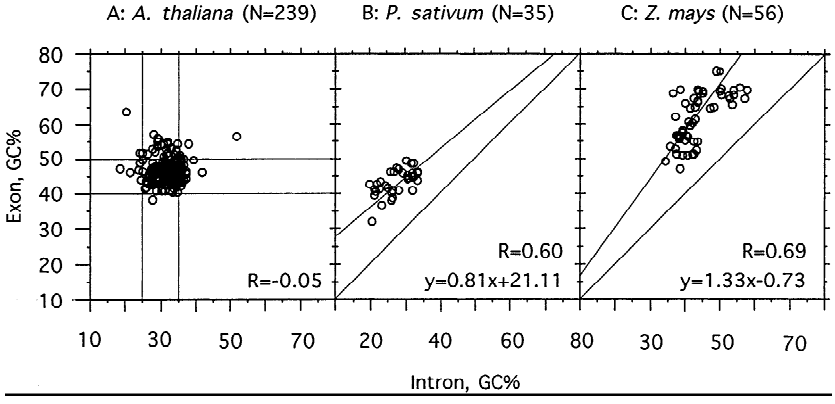

Fig. 6. $\overline{GC}$ levels of introns are plotted against $\overline{GC}$ levels of corresponding exons for *Arabidopsis* **(A),** pea **(B),** and maize **(C).** Other indications as in Fig. 3.

creases in the synonymous positions of most genes. Once these high values had been reached, the selective advantage favored compositionally conservative changes in these GC-rich synonymous positions. The second one was that repair (Filipski 1987), replication (Wolfe et al. 1989), or recombination (Eyre-Walker 1990) biases were the cause of the changes and of the subsequent conservation.

In order to discriminate between these two viewpoints, one should consider that the biases underlying the second explanation are due to mutations in the genes coding for the repair, replication, or recombination machineries. This explanation is, then, contradicted in the case of the two separate transitions leading from cold- to warm-blooded vertebrates by the fact that the transitions only occurred in the two lines leading from reptiles to mammals and to birds, but in no other line. In other words, the mutational bias hypothesis is difficult to accept in view of the random character of mutations and of the fact that no other cold-blooded vertebrate ever developed a warm-blooded genome pattern (Bernardi et al. 1997). Likewise, in the transition observed between dicots and Gramineae, it is of interest that there is no evidence of any transition similar to that observed in

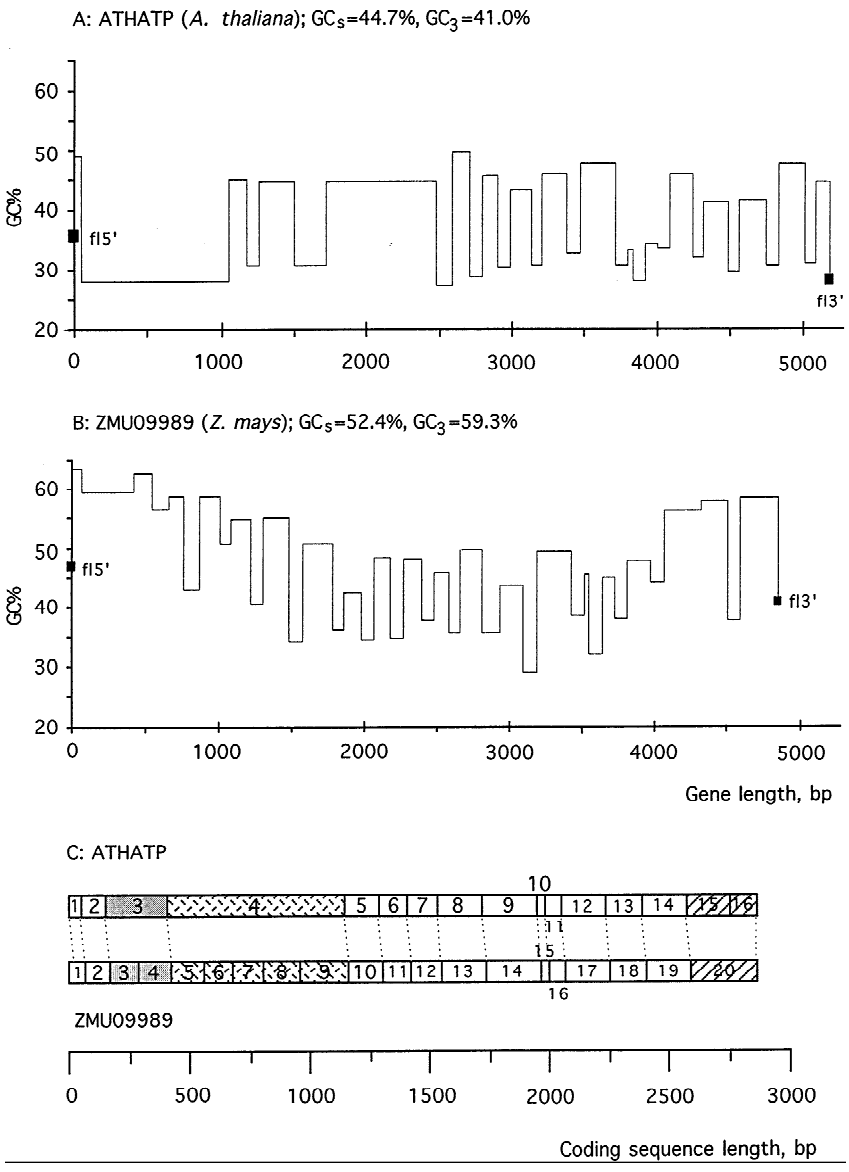

Fig. 7. Compositional profiles of the ATPase gene of *Arabidopsis* **(A),** ATHATP and of its maize homolog, FMU09989 **(B).** 5′ and 3′ flanking sequences are 213 and 260 bp long in the case of *Arabidopsis* and 1,445 and 2,995 bp in the case of maize, respectively. Exon correspondences and positions are shown in **C.**

Gramineae in either monocots or dicots (with the possible single exception of *Oenothera*). The mutational bias hypothesis seems, therefore, to be untenable also in the case of Gramineae.

Three additional, independent lines of evidence go in the same direction. (1) A detailed analysis of synonymous substitution in quartet codons of genes from four mammalian orders (Cacciò et al. 1995; Zoubak et al. 1995) has shown that, while GC_3-poor positions show the frequencies of substitution and the nucleotide compositions expected for a random nucleotide change process, this is not true for GC_3-rich positions, where the conservation found implies that the changes which must have occurred have been lost by a negative selection process. Two additional factors might, however, have helped compositional conservation. The first one, only a speculation at this time, is that the error rate of the replication machinery is lower in GC-rich regions compared to GC-poor regions of the genome. The second one is that highly transcribed genes are more efficiently repaired than rarely transcribed genes. This means that GC-rich genes, which comprise highly transcribed housekeeping genes, are better repaired than GC-poor genes, which comprise less transcribed tissue-specific genes. While these factors may reduce the load put on negative selection, obviously the latter is indispensable to maintain the extremely high GC_3 level of the majority of genes from both warm-blooded vertebrates and Gramineae. Although this detailed analysis has not been done in the case of Gramineae, the mere existence of a large number of genes from Gramineae having a 90–100% GC level in third positions indicates a negative selection. Indeed, the changes which must have occurred over the more than 50 MY of evolution of Gramineae are not seen because those genotypes were selected against. In contrast, the GC-poorest genes of Gramineae exhibit the same GC_3 values as their homologues from *Arabidopsis.* (2) A second line of evidence comes from very

recent work (Alvarez et al. 1997) on the synonymous substitutions in mammalian genes showing the existence of intragenic correlations with the nonsynonymous substitutions and with the base composition of synonymous positions. (3) A third line, which is specific to plant genomes, makes the strict repair/replication bias explanation impossible to hold. Indeed, introns and flanking sequences are considerably lower in GC (by about 20%) than exons (much more than in the case of vertebrates, where the difference is about 5% GC; Clay et al. 1996). This implies that the proposed mutational biases only affect third codon positions of individual exons, but not the neighboring introns. The conclusion that one can draw is, therefore, again that of natural selection on synonymous positions. Interestingly, this also seems to apply to the base composition of introns which are regularly, on the average, 20% lower than exons and 25% lower than third colon positions in Gramineae.

Acknowledgments. We thank Tomoko Ohta for valuable discussions and Oliver Clay for technical advice and critical reading of the manuscript.

References

Alvarez-Valin F, Jabbari K, Bernardi G (1997) Synonymous and nosynonymous substitutions in mammalian genes: intragenic correlations. J Mol Evol (in press)

Barakat A, Carels N, Bernardi G (1997) The distribution of genes in the genomes of Gramineae. Proc Natl Acad Sci USA 94:6857–6861

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

Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18

Bernardi G, Hughes S, Mouchiroud D (1997) The major compositional transitions in the vertebrate genome. J Mol Evol 44:S44–S51

Cacciò S, Zoubak S, D'Onofrio G, Bernardi G (1995) Nonrandom frequency patterns of synonymous substitutions in homologous mammalian genes. J Mol Evol 40:280–292

Carels N, Barakat A, Bernardi G (1995) The gene distribution of the maize genome. Proc Natl Acad Sci USA 92:11057–11060

Clay O, Cacciò S, Zoubak S, Mouchiroud D, Bernardi G (1996) Human coding and noncoding DNA: compositional correlations. Mol Phylogenet Evol 5:2–12

Crane PR, Friis EM, Pedersen KR (1995) The origin and early diversification of angiosperms. Nature 374:27–33

Crepet WL, Feldman GD (1991) The earliest remains of grasses in the fossil record. Am J Bot 78:1010–1014

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12: 387–395

Doolittle RF (1987) Of URFs and ORFs. A primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley, USA, p 103

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

Eyre-Walker A (1990) Recombination and mammalian genome evolution. Proc R Soc Lond B 252:237–243

Filipski J (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett 217:184–186

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

Gautier C, Jacobzone M (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

Harvey PH, Pagel M (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Laroche J, Li P, Bousquet J (1995) Mitochondrial DNA and monocot-dicot divergence time. Mol Biol Evol 12:1151–1156

Matassi G, Montero LM, Salinas J, Bernardi G (1989) The isochore organisation and compositionnal distribution of homologous coding sequences in the nuclear genome of plants. Nucleic Acids Res 17:5273–5290

Montero LM, Salinas J, Matassi G, Bernardi G (1990) Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Res 18:1859–1867

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

Salinas J, Matassi G, Montero LM, Bernardi G (1988) Compositionnal compartmentalization and compositionnal patterns in the nuclear genomes of plants. Nucleic Acids Res 16:4269–4285

SanMiguel PH, Tikhonov A, Jin Y-K, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen J (1996) Nested retrotransposons in the intergenic regions of the maize genome. Sciences 274:765–768

Wolfe KH, Gouy M, Yang Y, Sharp PM, Li WH (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86:6201–6205

Zoubak S, D'Onofrio G, Cacciò S, Bernardi G, Bernardi G (1995) Specific compositional patterns of synonymous positions in homologous mammalian genes. J Mol Evol 40:293–307

Zoubak S, Clay O, Bernardi G (1996) The gene distribution of the human genome. Gene 174:95–102