The compositional organization and the expression of the *Arabidopsis* genome

Nicolas Carels, Giorgio Bernardi*

Laboratorio di Evoluzione Molecolare, Stazione Zoologica 'Anton Dohrn', Villa Comunale, 80121 Naples, Italy

Received 15 February 2000; received in revised form 31 March 2000

Edited by Matti Saraste

Abstract The base composition patterns of genes, coding sequences and gene expression levels were analyzed in the available long sequences (contigs) of *Arabidopsis*. Chromosome 5 was analyzed in detail and all chromosomes for which sequence data are now available show essentially the same large-scale compositional properties. Guanine+cytosine levels of genes and of their coding regions, as well as gene densities and expression levels, all show a marked tendency to be higher in the distal regions of *Arabidopsis* chromosomes.

© 2000 Federation of European Biochemical Societies.

Key words: Chromosome; Isochore; Expression; Plant; *Arabidopsis*

1. Introduction

The genomes of angiosperms are characterized by heterogeneity in base composition among regions of 50–150 kb [1–3]. Compositional heterogeneity is smaller in dicots than in Gramineae [1,2,4], yet it is still discernible in the compact dicot genome that has been studied most extensively, that of *Arabidopsis*. In the present study, we have investigated the variations in GC (guanine+cytosine) levels along the chromosomes of *Arabidopsis*. The GC-rich telomeric regions are found to be richer in (GC-rich) genes and to exhibit higher expression levels than do the GC-poor central regions. The genes in these regions have, on average, higher GC, fewer introns and higher expression levels compared to the central regions of the chromosomes. Our work, part of a Ph.D. thesis [5], has concentrated on chromosome 5, which is especially well annotated. The conclusions obtained for chromosomes 2 and 4 [6,7] are similar to ours and the main results presented here are, almost certainly, valid for all chromosomes.

2. Materials and methods

Sequences and annotations of chromosome 5 of *Arabidopsis thaliana* were retrieved from the KAOS data base (Kazusa *Arabidopsis* data Opening Site, Kazusa DNA Research Institute, Japan, http://www.kazusa.or.jp/arabi/; cf. [8]), AtDB data base (*Arabidopsis thaliana* Database, Stanford University, http://genome-www.stanford.edu/Arabidopsis/; now at TAIR, http://www.arabidopsis.org/) and GenBank (releases 111 and 113; 1999).

To confirm that the contig samples analyzed were unbiased, the distributions of their GC levels were compared to an *Arabidopsis* CsCl profile obtained previously [9].

GenBank sequences were extracted using the Infobiogen server (see http://www.infobiogen.fr) with the ACNUC/QUERY retrieval system [10]. Sequences with self-contained coding sequence (CDS) fields (features) were retained if they had a total CDS longer than 220 bp. ATG and stop codons indicated by the feature specifications were checked and the corresponding 'complete' CDSs were also checked to ensure that they contained an integral number of codons. Sequences without ATG or stop codons in the expected positions and pseudogenes were not included in the analysis. To allow automation, only introns interrupting the CDSs of genes were processed. Obvious gene redundancies were recognized and eliminated on the basis of identity of sequence size, equivalence of descriptions and similarity of base composition (within 2% GC) in the three codon positions. In the KAOS data base, only complete gene sequences with a clear definition were analyzed; sequences corresponding to 'hypothetical', 'unknown' and 'putative' proteins were excluded. Some of the annotated genes retained for analysis have not been experimentally confirmed and must therefore still be classified as potential genes, but the possible presence of occasional unexpressed genes in the sample should not systematically bias any of the results presented here; the alternative, namely retaining only experimentally confirmed genes, would result in unacceptably low sample sizes. In the case of contiguous repeated genes, only one copy was investigated. A base composition of sequences was calculated using ANALSEQ [11].

In the following, we only considered genes in which introns are in the range of 0–3.5 kb. This restriction eliminates very long genes with introns up to 100 kb, which are rare, yet can considerably lower the correlation coefficient of intron size versus genic GC level.

The statistical significance of the differences of means was assessed using the test of Student [12,13], at the 5% significance level (one-tailed test).

The GC level distributions for all chromosomes and, independently, for chromosome 5, were represented by histograms with an interval size of 0.25% GC. The histograms were first fitted using the program MacCurveFit (1.0.8) and were found to be close to a Gaussian curve with a mean at 35.4% GC and standard deviation 1.56% GC. In the case of chromosome 5, the subsequent fitting to a sum of two Gaussian components, shown in Fig. 3, was based on the partitioning of this chromosome (see text) and was adjusted manually while monitoring the sum of squared distances. In this fit, which matched the single Gaussian fit and which was very similar to the CsCl profile of *Arabidopsis*, the GC-poor and GC-rich components of chromosome 5 were characterized, respectively, by relative amplitudes of 9 and 14.9, means at 34.4 and 36.1% GC and standard deviations of 1.2 and 1.3% GC.

The gene expression level was estimated as described in [14]. 945 CDSs, in a total of 252 contigs, extracted from the KAOS database were compared to a data set of 25 547 ESTs of *Arabidopsis* (GenBank Release 111) using BLASTN [15]. The same procedure was then applied to compare the 3610 genes in 508 contigs from all chromosomes of *Arabidopsis* with a total of 45 743 ESTs (GenBank Release 113). Each BLASTN alignment showing at least 95% identity over at least 100 bp was counted as a sequence match. The number of EST matches was then assumed, as in [14], to statistically reflect the expression level of the genes.

To confirm that the main results obtained for chromosomes 5 and 3 (also represented in the KAOS/AtDB databases) were valid also for chromosomes 2 and 4, the recently published plots of GC level, predicted gene density and EST frequencies for chromosomes 2 [6] and 4 [7] were compared with the corresponding plots for chromosomes 5 and 3. One plot not given in these publications, the GC plot of chromosome 4, was obtained at window sizes 50 and 100 kb from

*Corresponding author. Fax: (39)-81-7641 355.
E-mail: bernardi@alpha.szn.it

0014-5793/00/$20.00 © 2000 Federation of European Biochemical Societies. All rights reserved.
PII: S0014-5793(00)01476-9

the complete sequences (short arm, long arm; MIPS ftp site, ftp://warthog.mips.biochem.mpg.de/pub/cress/chrIV/ESSAseq/). The GC plot for chromosome 2, shown in [6] at window size 10 kb, was obtained also at window sizes 50 and 100 kb where the large-scale compositional trends are much more clearly visible from the complete sequence (ftp://ftp.tigr.org/pub/data/a_thaliana/chromosomeII/).

3. Results and discussion

A compositional map of chromosome 5 was obtained by plotting the average base compositions (GC levels) of the long DNA sequences from the KAOS database (which have a sharp mode at 85 kb) against their approximate positions (in 2 Mb intervals), according to the KAOS physical map. Fig. 1 shows the average GC levels of the contigs in each 2 Mb interval along chromosome 5 (bottom plot), together with (from bottom to top) the average GC levels of genes, third codon positions (GC_3) and CDSs in the same intervals. This compositional map was then refined by using information on the order of these contigs given in the AtDB database. The resulting plot (not shown) reveals a succession of fluctuations in GC level that are large, with respect to the total variability in the genome, but that still show a persistent, gradual increase in average GC level from the pericentromeric regions towards the telomeres. Despite the high variability (~6% GC) among the base compositions of contigs in each 2 Mb interval (Table 1), GC levels are significantly higher in the distal regions (0–8 Mb on the short arm and 18–26 Mb on the long arm), where their average is 36.3% GC, than in the central region (8–18 Mb), where their average is 34.6% GC.

Moreover, it can be seen that the four plots in Fig. 1 largely parallel each other. As expected from this parallelism, correlations were found between the per-contig average GC levels of genes, of their CDSs or of their third codon positions and the GC level of the contigs in which the genes were located (e.g. $R = 0.52$ for the genes versus contigs correlation, shown in Fig. 2). Consistently lower correlation coefficients were found for the correlations between GC levels of individual genes, of their CDSs or of their third codon positions (GC_3) and the GC level of the contigs ($R = 0.22$, 0.20 and 0.15, respectively, for chromosome 5 contig sets with 768 genes analyzed). All of the average GC levels are higher in the distal regions (0–8 and 18–26 Mb intervals) than in the central region (8–18 Mb intervals), these differences being significant for all plots of Fig. 1 except for the GC_3 plot, as revealed by the test of Student. For the complete set of contigs (1349 genes analyzed), we found that the correlation coefficients were all between $R = 0.21$ (for CDSs) and $R = 0.24$ (for GC_3), that of the genes being again at $R = 0.22$.

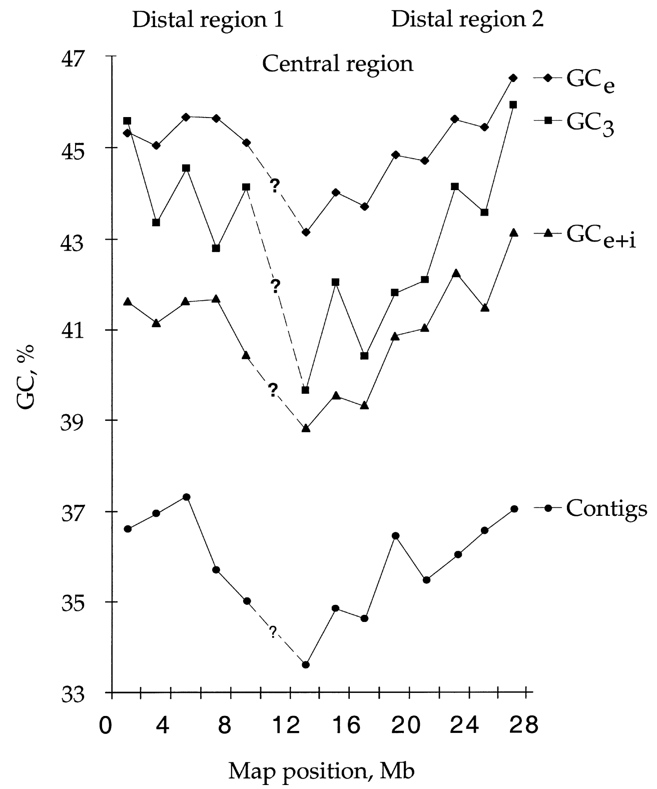

Fig. 1. Compositional profiles of contigs, genes (i.e. exons+introns, GC_{e+i}), CDSs (GC_e) and third codon positions (GC_3). Each point corresponds to the average value within a 2 Mb segment of chromosome 5. The question mark indicates the centromeric region of chromosome 5, for which there are presently no data available.

As can be deduced from Figs. 1 and 2 and from the lower average GC level of genes (exons+introns) in the central region of chromosome 5, this latter region tends to be associated with the GC-poorer genes, which have average GC levels (taken over 2 Mb intervals) below 41%. The same observation can be generalized to the whole genome of *Arabidopsis*, since Fig. 2 refers to contigs and genes from the complete set of

Table 1
Compositional relationships along chromosome 5 of *Arabidopsis*

Map position (Mb)	Contigs (GC%)	S.D.	Sample size	Genes (GC%)	S.D.	GC_3 (%)	S.D.	Coding sequences (GC%)	S.D.	Sample size
0–2	36.59	(1.23)	10	41.61	(4.03)	45.58	(6.99)	45.33	(2.84)	49
2–4	36.94	(1.23)	7	41.15	(3.88)	43.36	(6.66)	45.05	(2.79)	49
4–6	37.33	(1.10)	8	41.61	(3.88)	44.54	(6.75)	45.67	(2.89)	54
6–8	35.71	(1.42)	8	41.67	(4.41)	42.79	(7.71)	45.65	(3.24)	33
8–10	34.99	(1.07)	17	40.43	(3.68)	44.12	(6.66)	45.11	(2.84)	53
12–14	33.60	(1.23)	25	38.79	(4.28)	39.65	(8.14)	43.11	(3.36)	26
14–16	34.85	(1.05)	35	39.58	(3.77)	42.05	(6.42)	44.00	(2.90)	108
16–18	34.62	(1.20)	30	39.32	(3.81)	40.42	(6.27)	43.69	(2.82)	77
18–20	36.44	(0.86)	31	40.86	(3.69)	41.79	(7.02)	44.82	(2.84)	36
20–22	35.47	(1.14)	35	41.03	(3.71)	42.09	(6.41)	44.69	(2.82)	118
22–24	36.00	(1.10)	29	42.26	(3.61)	44.12	(6.19)	45.61	(2.80)	101
24–26	36.55	(1.29)	35	41.45	(3.70)	43.55	(6.18)	45.42	(2.82)	135
26–28	37.02	(1.88)	4	43.12	(4.45)	45.90	(5.97)	46.53	(2.60)	16

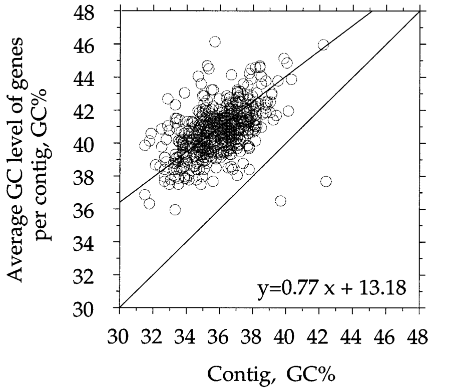

Fig. 2. Relationship between the GC level of contigs and the average GC level of their genes. The orthogonal regression line ($R = 0.52$) and the main diagonal of slope 1 are shown.

chromosomes and since GC-poor contigs are preferentially observed in the central regions of chromosomes 2 [6], 3 (data not shown), 4 [7] and 5 (see also Section 2 and [5]). However, as predicted by the absence of a strong correlation between the GC levels of genes and contigs ($R = 0.22$), a substantial proportion of GC-poor genes were found in the distal regions of chromosomes (or, as defined below, in the GC-rich component) and vice versa. In angiosperms, we have previously found that the GC-richer genes of a given species have fewer and shorter introns than do the GC-poorer genes [16]. The difference in intron number and size could correspond to some advantage for constitutive or, at least, for extensive expression of genes (see [16], for a discussion of the two classes of genes).

The GC distribution of the chromosome 5 contigs, which is similar to that of the long (> 50 kb) sequences from all five chromosomes, can be conveniently decomposed into two

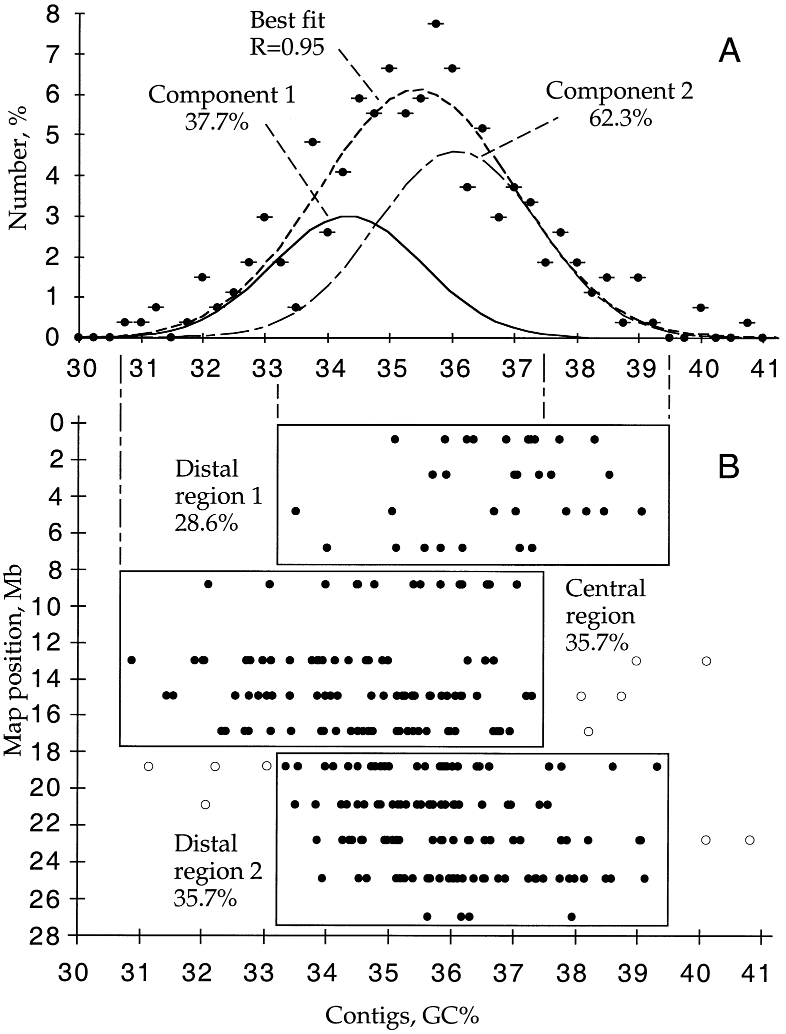

Fig. 3. Relationships between compositional distribution of contigs, Gaussian components, sum of components and compositional map along chromosome 5. The compositional distribution of contigs in (A) is represented by the closed circles with horizontal bars. The dashed line shows the best fit ($R = 0.95$) to the compositional distribution of contigs by a sum of the two Gaussian curves, component 1 (continuous line) and component 2 (dash-dot line). The percentages indicated for components 1 and 2 represent their relative contributions to the area under the best fit curve. (B) represents the GC level distribution of the contigs along chromosome 5, in intervals of 2 Mb, using physical map positions. The inner rectangle corresponds to the central region of the chromosome, the top and bottom rectangles to the distal parts of the chromosome. The percentages shown for these rectangles represent the proportions of chromosome 5 that they cover according to the physical map. The GC averages, standard deviations and proportions are essentially the same for the components as for the corresponding chromosomal regions (central region and sum of the two distal regions). Open circles represent the few contigs outside of the rectangles. They were included in the calculation of standard deviation and average GC level of central and distal regions.

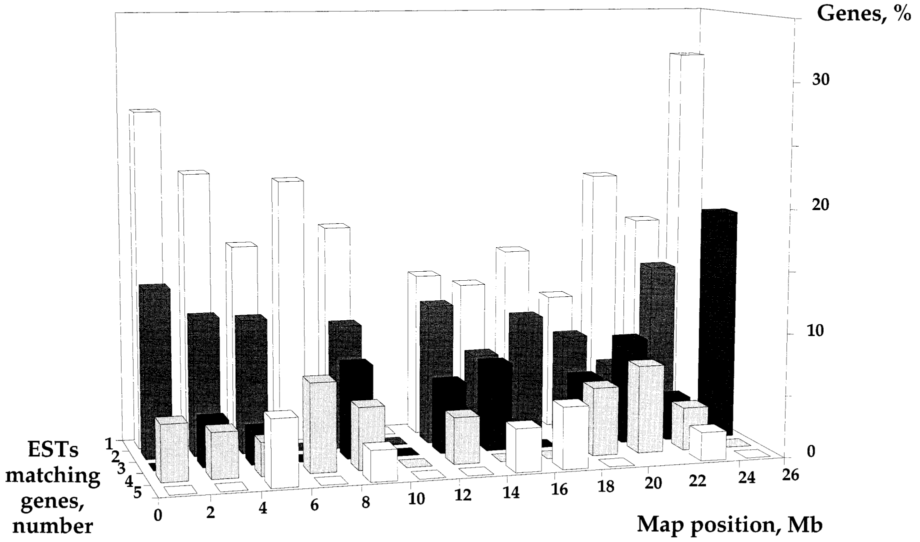

Fig. 4. Profile of gene expression along chromosome 5. The bar heights show the percentage of the genes in each 2 Mb interval that were matched by 1, 2, 3, 4 or 5 ESTs.

Gaussian components, namely a GC-poorer and a GC-richer component (Fig. 3A). This decomposition corresponds to the partitioning of chromosome 5 into central (GC-poorer) and distal (GC-richer) regions (Fig. 3B). The two components obtained for chromosome 5 (Fig. 3A) are characterized by average GC contents (34.4 and 36.1% GC) and standard deviations (GC-poor: 1.2 and GC-rich: 1.3% GC) that are very similar to those of the contigs of distal and central regions to which they correspond (Fig. 3B). Similarly, the ratio of the components' areas (GC-poor: 37.7%; GC-rich: 62.3%) was very close to the ratio of the lengths of the corresponding chromosomal regions. The sum of the two Gaussian components was itself very nearly Gaussian (because of the nearly symmetric GC distribution of the DNA in *Arabidopsis*) and showed a good fit to the GC distribution of the contigs in chromosome 5 ($R = 0.95$) as well as to the experimentally obtained GC distribution of the entire genome's DNA fragments in a CsCl gradient [9]. This decomposition of the *Arabidopsis* DNA of chromosome 5 into two overlapping components allows a good correspondence between GC levels and chromosomal position. Indeed, the compositional differences between the central and distal regions are not clear-cut and borders could be localized only within 2 Mb, so that a partitioning of the GC axis into two segments would be inappropriate.

In order to determine if one of the two components of chromosome 5, or of other chromosomes, is associated with higher gene expression, we estimated expression levels in the entire set of *Arabidopsis* contigs by EST matching, as described by Duret and Mouchiroud [14]. High expression levels were much more common in the contigs having GC levels at or above 34.5% (data not shown), i.e. in the contigs corresponding to the GC-richer of the two components in the case of chromosome 5 (see Fig. 3). This prompted us to examine gene expression levels along this chromosome in more detail. We found that genes are indeed expressed at higher levels in the more distal regions. The distribution of expression levels along the (2 Mb intervals of) chromosome 5, up to a level of 5 ESTs per gene (cf. Section 2), is shown in Fig. 4. The few genes with higher expression levels (> 5 ESTs per gene; not shown) followed the same general shape of this distribution, i.e. an upward curve toward the telomeres. The expression level was found higher on the long arm than on the short one (data not shown). In contrast, the minimum of gene expression was found in the central part of the chromosome. The only exception was an extreme outlier representing three copies of the ribulose biphosphate carboxylase small chain gene, which were located in a contig (MXI10) in the central (GC-poor) region even though they were very GC_3-rich (60–63%) and each of which matched between 60 and 100 ESTs.

Finally, gene density was analyzed. When the gene density of each contig was measured by the proportion covered by genes, the full set of *Arabidopsis* contigs exhibited a correlation between gene density and GC level. This is shown in Fig. 5, for the plot of average gene density in each 0.25% GC interval versus the midpoint of the GC interval. While the animal species investigated so far, including human [17,18], fugu [19,20] and *Drosophila* [21], have shown a dramatic in-

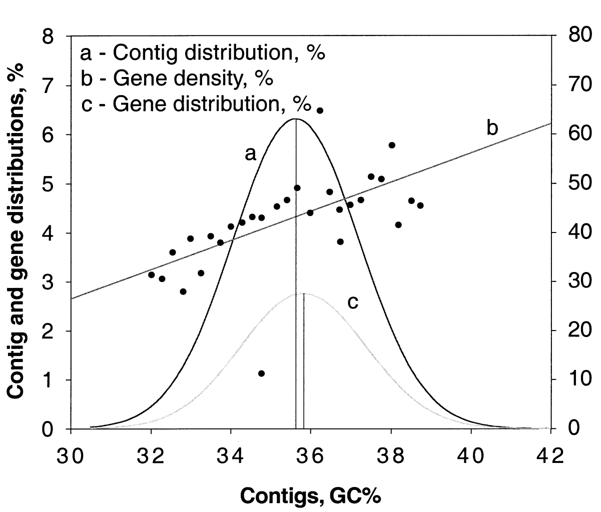

Fig. 5. Profiles of gene density (closed circles, regression line) and of contig (black curve) and calculated gene distributions (gray curve) in the *Arabidopsis* genome.

crease of gene density with GC level that is highly non-linear (or only piecewise linear), the corresponding relation in *Arabidopsis* is approximately linear ($y = 3.22x - 71.35$). Although it cannot be completely ruled out that this linear relationship may be partly an artefact of gene recognition protocols or algorithms used for annotating the contigs (cf. also Section 2), there is as yet no indication of a significant GC bias in these algorithms, especially since the GC range in *Arabidopsis* is not very wide. Based on an average gene length of 2.1 kb [16] and the data shown in Fig. 5, we could calculate that the gene density ranges from 1 gene per 6 kb in the GC-poor contigs to 1 gene per 3.8 kb in the GC-rich contigs. Therefore, the gene density range is about 10 times lower in *Arabidopsis* than in the case of the human genome.

The good correlation coefficient ($R = 0.62$, $P < 0.001$) supporting the linear relation between gene density and GC level of contigs prompted us to estimate the gene distribution in the *Arabidopsis* genome, by multiplying the best fit of the contigs' GC distribution (which was also close to the CsCl profile of *Arabidopsis*; see above and Fig. 3) with the equation. A comparison of this calculated gene distribution with the contig (DNA) distribution (Fig. 5) shows that the mode of the gene distribution is slightly shifted. The superposition of these two curves offers another way of visualizing and quantifying the modest, but significant, tendency of genes to occur preferentially in GC-richer regions of the *Arabidopsis* genome.

The main trends reported here for chromosome 5, namely the higher GC levels, gene densities and expression levels in the distal regions, were confirmed for chromosome 3 (which is also represented in the KAOS/AtDB databases) and for the recently sequenced chromosomes 2 and 4 [6,7]. In particular, both the long and the short arms of chromosomes 2 and 4 show the persistent rise in GC level towards the telomeres that is exhibited by chromosomes 5 and 3, as was confirmed by chromosomal maps of 50 and 100 kb segments along the chromosomes (see Section 2). In some cases (GC content in chromosome 2, gene and EST densities in chromosome 4) the tendencies are only faintly visible in the published plots of [6,7], due to the scales and window/interval sizes used in those studies, although differences in gene-prediction criteria may also be partly responsible for some quantitative differences.

The only tendency of chromosomes 2 and 4 that was not visible in the chromosome 5 data, presumably due to the gap in the available sequences (see Fig. 3; 10–12 Mb according to the KAOS map, 11–13 Mb according to [22]), was a series of large fluctuations, attaining high GC levels, in a centromeric region covering about 2–3 Mb. The corresponding region has been mapped in detail, but not yet sequenced, in chromosome 5 [22]. This region is characterized by a high density of retrotransposons, 180-bp repeats and other repetitive DNA, an absence of genes and, in the case of the sequenced chromosomes 2 and 4, by GC levels that show strong fluctuations [6,7,21]. In chromosome 5, two clusters of 5 S-rDNA (containing over 200 copies and about 30 copies, respectively), which are GC-rich (cf. also the CsCl profile in [9]), have been mapped near the boundaries of this region. It is of interest that in chromosome 2 an (apparently very recent) insertion of about 700 kb of mitochondrial DNA has been reported, in this same centromeric region [6]. The region also seems to be difficult to map: two groups have produced conflicting physical maps for it in the case of chromosome 5 [22,23]. It corresponds to only a small fraction of the chromosomes studied in detail until now (about 1/10 for chromosome 5 and about 1/6 for the smaller chromosomes 2 and 4). In chromosomes 2 and 4, the region is markedly different from the rest of the chromosome, devoid of genes, and its GC-rich segments lie beyond the GC-richer of the two Gaussian components (at 37–41% GC), so that it could interfere at most marginally with the results presented above. To avoid misunderstandings, however, it should be pointed out here that the compositional trends described above, including the two Gaussian components, are properties of the two arms of the chromosomes and not of the centromeric region between them. Indeed, the clear-cut compositional anomaly that characterizes the latter region could be used to operationally divide *Arabidopsis* chromosomes into short arm, centromeric region and long arm.

Acknowledgements: We gratefully thank O. Clay for his careful revision of the manuscript.

References

[1] Salinas, J., Matassi, G., Montero, L.M. and Bernardi, G. (1988) Nucleic Acids Res. 16, 4269–4285.
[2] Matassi, G., Montero, L.M., Salinas, J. and Bernardi, G. (1989) Nucleic Acids Res. 17, 5273–5290.
[3] Montero, L.M., Salinas, J., Matassi, G. and Bernardi, G. (1990) Nucleic Acids Res. 18, 1859–1867.
[4] Carels, N., Hatey, P., Jabbari, K. and Bernardi, G. (1998) J. Mol. Evol. 46, 45–53.
[5] Carels, N. (1999) The Organization and Evolution of Angiosperm Genomes. Ph.D. thesis, Université de Paris 6, December 14, 1999.
[6] Lin, X. and Kaul, S. (1999) Nature 402, 761–768.
[7] Mayer, K. and Schüller, C. (1999) Nature 402, 769–777.
[8] Kaneko, T., Katoh, T., Sato, S., Nakamura, Y., Asamizu, E., Kotani, H., Miyajima, N. and Tabata, S. (1999) DNA Res. 6, 183–195.
[9] Barakat, A., Matassi, G. and Bernardi, G. (1998) Proc. Natl. Acad. Sci. USA 95, 10044–10049.
[10] Gouy, M., Gautier, C., Attimonelli, N., Lanave, C. and Di Paola, G. (1985) CABIOS 1, 167–172.
[11] Gautier, C. and Jacobzone, M. (1989) UMR CNRS 5558 Biometrie, Genetique et Biologie des Populations, Universite Claude Bernard, Lyon I, France.
[12] Student (1908) Biometrika 6, 1–25.
[13] Student (1925) Metron 5, 105–120.
[14] Duret, L. and Mouchiroud, D. (1999) Proc. Natl. Acad. Sci. USA 96, 4482–4487.
[15] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[16] Carels, N. and Bernardi, G. (2000) Genetics, in press.
[17] Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, Gautier, C. and Bernardi, G. (1991) Gene 40, 181–187.
[18] Zoubak, S., Clay, O. and Bernardi, G. (1996) Gene 174, 95–102.
[19] Bernardi, G. (1995) Annu. Rev. Genet. 29, 445–476.
[20] Tassone, F., Villard, L., Clancy, K. and Gardiner, K. (1999) Gene 226, 211–223.
[21] Jabbari, K. and Bernardi, G. (2000) Gene, in press.
[22] Kotani, H., Hosouchi, T. and Tsuruoka, H. (1999) DNA Res. 6, 381–386.
[23] Tutois, S., Cloix, C. and Cuvillier, C. et al. (1999) Chromatogr. Res. 7, 143–156.