ELSEVIER

Gene 205 (1997) 119–124

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

Methylation patterns in the isochores of vertebrate genomes

Simone Cacciò [a,1], Kamel Jabbari [a], Giorgio Matassi [a,2], Fanny Guermonprez [b], Jean Desgrès [b], Giorgio Bernardi [a,*]

[a] *Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu 75005 Paris, France*
[b] *Laboratoire de Biochimie Médicale, Faculté de Médecine et Centre Hospitalier Universitaire, 7 Bd Jeanne d'Arc, 21033 Dijon, France*

Abstract

5-Methylcytosine (5mC) levels were determined in compositional DNA fractions corresponding to different isochore families from the genomes of *Xenopus*, chicken, mouse and human, four vertebrates which show different isochore patterns. The results obtained indicate that: (i) positive correlations exist between the 5mC levels and the GC levels of isochores within any given genome; and (ii) DNA from *Xenopus* isochore families is twice as methylated as DNA from the isochores having the same GC levels from mouse, human and chicken. Moreover, the positive correlations holding between CpG levels and the GC_3 levels of coding sequences of warm-blooded vertebrates were shown to comprise two regions with a border at approx. 75% GC_3. The correlation corresponding to the higher region (which comprises only very rare high GC_3 values in the case of *Xenopus*) has a higher slope than that corresponding to the lower GC_3 values, a phenomenon due in all likelihood, to the increasing contribution of CpG islands. Finally, the observed/expected CpG ratio is higher in *Xenopus* than in warm-blooded vertebrates. © 1997 Elsevier Science B.V.

Keywords: CpG dinucleotides; Vertebrate evolution

1. Introduction

Vertebrate genomes are mosaics of isochores, namely of long, compositionally homogeneous DNA segments, that belong to a small number of families having different GC levels and different gene densities (see Bernardi, 1995, for a recent review). In the present work, we have studied the methylation levels within four individual vertebrate genomes which are characterized by different isochore patterns.

The *Xenopus* genome shows a low GC level and the typical isochore pattern of a cold-blooded vertebrate, namely a pattern that is characterized by a very narrow compositional distribution of DNA molecules (Thiery et al., 1976; Bernardi and Bernardi, 1990a,b).

The chicken genome, like the genomes of all warm-blooded vertebrates, is composed of both GC-poor isochores (the 'paleogenome', which corresponds to the bulk of the genome of cold-blooded vertebrates) and GC-rich isochores, which represent the 'neogenome' (Bernardi, 1989), namely the regions of the warm-blooded genome which have undergone a compositional transition toward GC enrichment. The chicken isochore pattern is characterized by a very broad compositional distribution of DNA molecules, with a relatively large proportion of GC-rich isochores which attain very high GC levels (Cortadas et al., 1979; Olofsson and Bernardi, 1983; Kadi et al., 1993). This pattern is common to all the avian species studied so far (Kadi et al., 1993; Mouchiroud and Bernardi, 1993).

As far as mammals are concerned, both the human and the mouse genomes were studied. While the former is representative of the general mammalian isochore pattern, which is shared by the majority of mammalian orders (Sabeur et al., 1993), the latter shows a relatively narrow distribution of DNA molecules in which both very GC-poor and very GC-rich isochores are scarce or absent (Salinas et al., 1986; Zerial et al., 1986; Mouchiroud et al., 1988; Mouchiroud and Bernardi, 1993).

Other features of the genome organization are known

* Corresponding author. Tel. +33 1 43 295824; Fax: +33 1 44 277977.
[1] Present address: Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Roma, Italy.
[2] Present address: Inserm U76, 6 rue Alexandre Cabanel, 75015 Paris, France.

Abbreviations: 5mC, 5-methylcytosine; bp, base pair; GC, molar fraction of guanine + cytosine in DNA; HPLC, high-performance liquid chromatography.

PII S0378-1119(97)00560-X

in at least some of the species studied. These are the distribution of repetitive and single-copy DNA sequences (Meunier-Rotival et al., 1982; Soriano et al., 1983; Olofsson and Bernardi, 1983), of CpG islands (Aïssani and Bernardi, 1991a,b) and of the isochore family richest in genes among the isochore families of the genomes under consideration (Cacciò et al., 1994) and the correlations between isochores and chromosomal bands of the human and mouse genomes (Saccone et al., 1992, 1993, 1996, 1997). In addition, the four vertebrate species selected for this study have relatively large sequence sets in data banks.

2. Materials and methods

2.1. DNA sources and nucleoside analysis

DNA was prepared as previously described (Cuny et al., 1981) from placenta (man), liver (mouse), and blood (chicken and *Xenopus*). DNA preparations were fractionated in Cs_2SO_4/BAMD preparative gradients, as described (Cortadas et al., 1977, 1979; Macaya et al., 1978). BAMD is 3,6-bis(acetato-mercuri-methyl-dioxane). A ligand/nucleotide molar ratio (r_f) of 0.14 was used in the case of man, mouse and chicken, whereas an r_f of 0.10 was used in the case of *Xenopus*. The analytical procedure for the quantitative analyses of nucleosides in DNA is described in the preceding paper (Jabbari et al., 1997).

2.2. Data bank analyses

Sequences from GenBank (Release 99; 15 February 1997) and HOVERGEN (Duret et al., 1994) were processed using the ACNUC retrieval system (Gouy et al., 1985); the program ANALSEQ (Gautier and Jacobzon, 1989) was used to determine the base composition and doublet frequencies of coding sequences. Four non-redundant data sets were obtained, which comprise 888, 952, 3902 and 6657 coding sequences from *Xenopus*, chicken, mouse and man, respectively.

3. Results

3.1. Methylation patterns in isochores from vertebrate genomes

Fig. 1 displays the 5mC and the GC levels of compositional DNA fractions obtained from the four vertebrate species analyzed, as well as the relative amounts of each DNA fraction within each genome. These results can be summarized as follows:

(1) DNA fractions from the *Xenopus* genome cover a narrow GC range, from 36.4 to 43.5%, and methylation levels range from 1.14 to 1.40%, if the high methylation levels observed in the last two fractions (1.65 and 2.02%), which are due to a satellite DNA component (Thiery et al., 1976; Macaya et al., 1978), are neglected.
(2) As far as chicken is concerned, DNA fractions cover a large GC range, the GC level varying from 36.5 to 57.2%, while methylation levels range from 0.37 to 0.71%, neglecting again the very high methylation levels of the last three fractions (1.23–1.88%), which are due, as in the case of *Xenopus*, to satellite DNAs (Cortadas et al., 1979; Kadi et al., 1993).
(3) In the case of the mouse genome, GC levels of isochores range from 35.5 to 50.2%, and methylation levels range from 0.54 to 0.86%. The higher methylation level observed in fractions 3 and 4 is caused by a satellite DNA, which is known to be hypermethylated, to the extent of 2.4–3.1% 5mC (Feinstein et al., 1985).
(4) Finally, in the case of man, the GC level of isochores varied from 37.2 to 50% and methylation levels covered a 0.4–1.1% range. The relatively low GC level of the last fraction, which contains DNA from the H3 isochore family, is due to an AT-rich satellite. Indeed, the GC level of the main-band DNA of the last fraction, as estimated from its buoyant density using analytical centrifugation and, therefore independently of the satellite, is 54% (Saccone et al., 1996), and its methylation level, as measured on fractions free of the satellite is 1.1%, well on the regression line.

Positive, statistically highly significant correlations hold between methylation and GC levels of DNA fractions (see Fig. 1) for all the species investigated. Interestingly, the slopes cover a 3-fold range, the lowest and the highest values being observed for the mouse and chicken genomes, respectively. While the difference between the slopes of human and chicken fractions is doubtful, that of human and mouse DNA fractions is clear.

3.2. An analysis of CpG distribution in coding sequences

Correlations between CpG levels and GC levels have already been reported for coding sequences of vertebrates (Bernardi et al., 1985; Bernardi, 1985; Aïssani and Bernardi, 1991a,b), for coding sequences of viruses of warm-blooded vertebrates (Bernardi and Bernardi, 1986) and for coding sequences of plants (Montero et al., 1990). Here, we have re-examined this point by using the currently available coding sequences of *Xenopus*, chicken, mouse and human. The results are shown in Fig. 2.

An increase in CpG (and GpC; not shown) with increasing GC_3 of coding sequences was observed in all cases. A tendency towards increasing slopes in the

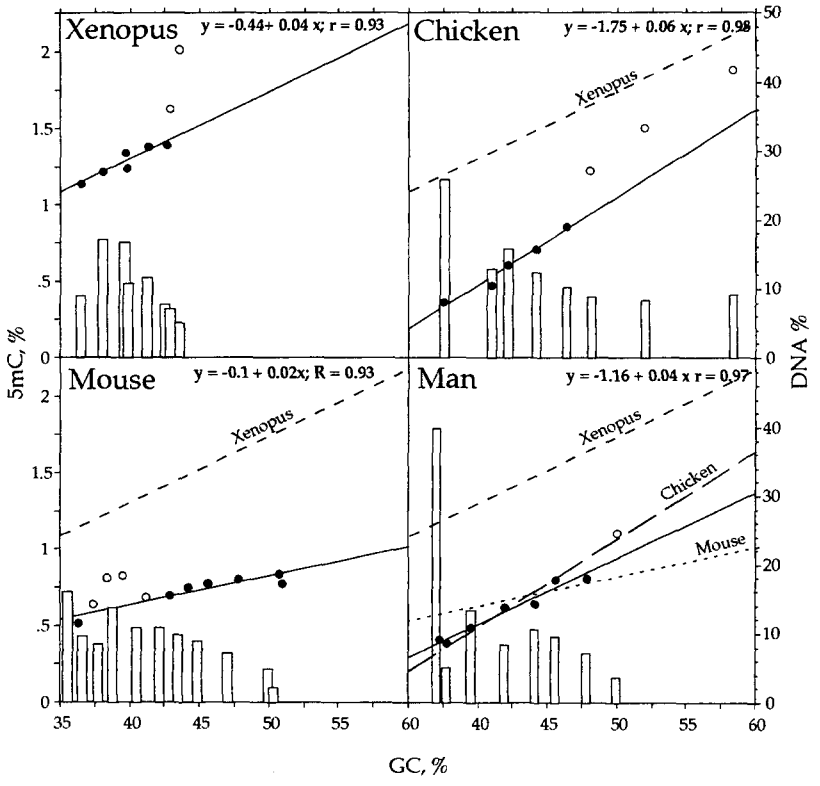

Fig. 1. Plot of 5mC vs GC for compositional DNA fractions from *Xenopus*, chicken, mouse and man (left-hand scale). The histograms indicate the relative amounts of DNA in the fractions (right-hand scale). Circles indicate 5mC levels. Least square lines through the points (and their equations) are shown; only filled circles were used in drawing the lines. The *Xenopus* (broken) line is also shown in the other diagrams for comparison. Likewise, the mouse and chicken lines are shown in the human diagram.

highest GC_3 range is evident in all warm-blooded vertebrates studied here, whereas this trend is barely visible in *Xenopus*, in which case the GC_3 range does not reach high values.

Interestingly, the observed/expected CpG ratios of *Xenopus* coding sequences are higher than those of coding sequences (in the same GC_3 range) from warm-blooded vertebrates. Fig. 3 shows the results of the man/*Xenopus* comparison.

Finally, the percentage of CpG doublets in positions (1,2), (2,3), and (3,1) of coding sequences was analyzed (Table 1). The 3,1 positions were predominant in all cases, showing a slight increase from *Xenopus* (41%) to chicken (45%). This trend is related to GC-richness of the coding sequences analyzed, as shown by the fact that the corresponding CpG o/e ratios were essentially the same in all species.

4. Discussion

The results shown in Fig. 1 can be summarized as follows:

(1) The 2–3-fold higher methylation level in *Xenopus* isochores, compared with that of the corresponding isochores (L1 and L2) of each of the warm-blooded vertebrates analyzed, is consistent with the results obtained at the whole genome level (see Jabbari et al., 1997). This finding shows that the difference is not due to the presence or absence of satellite DNAs, nor to the different amounts and qualities of interspersed repeated sequences present in the genomes of the three warm-blooded vertebrates (see below).

(2) While all genomes show highly significant, positive correlations between 5mC and GC levels, the slopes of the regression lines between 5mC and GC levels are different in different species, and thus represent features characteristic of the genomes, as was previously observed for plants (Matassi et al., 1992). The slope differences are, however, minor differences compared with the large differences in methylation level between *Xenopus* and warm-blooded vertebrates.

In order to understand the 2-fold difference in slope exhibited by mouse as compared with human DNA fractions (see Fig. 1), the following should be recalled. Reassociation kinetic experiments, which assessed the distribution of the fold-back, fast, intermediate and slow reassociating components in the isochore families of the human and mouse genomes, show remarkable differ-

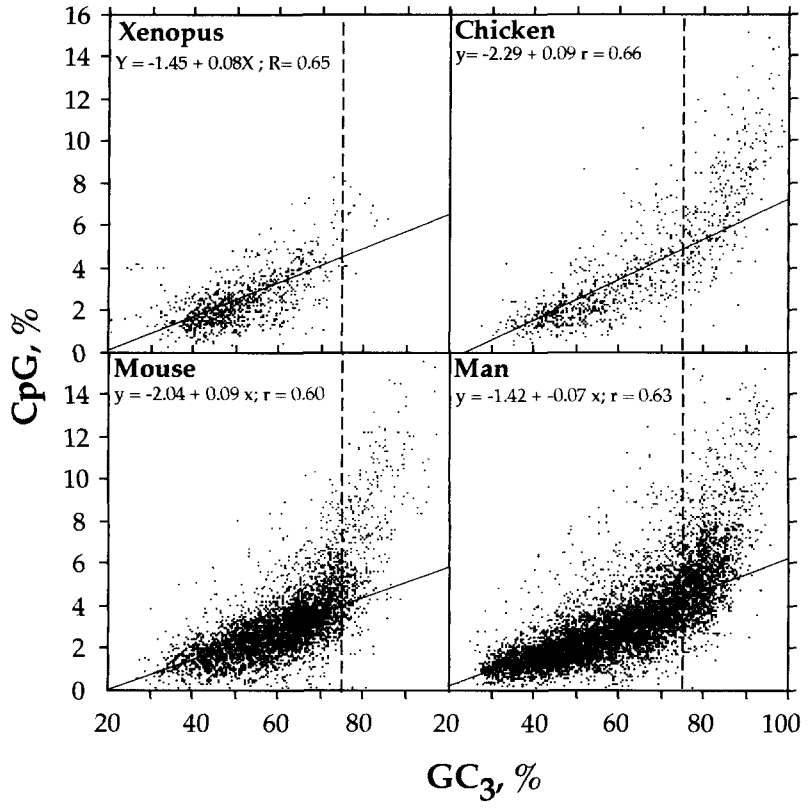

Fig. 2. Plots of CpG vs GC_3 of coding sequences from *Xenopus*, chicken, mouse and man. The equations for the line through the points (up to 75%, GC_3) and the correlation coefficients are shown.

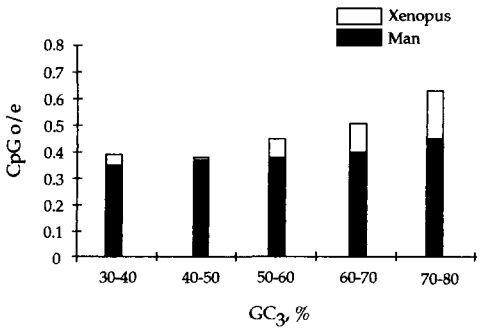

Fig. 3. Observed/expected CpG ratio for coding sequences of *Xenopus* and man. Black bars correspond to results obtained with human sequences belonging in 5 GC_3 intervals. White bars correspond to the excess CpG o/e values of *Xenopus* coding sequences in the same GC_3 intervals.

ences (Soriano et al., 1981). Indeed, while the total amount of repeated sequences of all classes is very close in the two genomes (41% in mouse and 38% in man), their relative amounts decrease from 44% in L1 to 9% in H2 for mouse, whereas it increases from 29% in L1 to 54% in H2 for man. These differences might account for the differences in the methylation slopes, even if the base compositions of denatured and of reassociated DNA were the same within experimental error over the whole *cot* range (cot is the product of the initial DNA concentration and the reassociation time). It should be stressed, however, that while the differences in methylation slopes can be explained by the different distribution of repeated DNA, this factor cannot explain the large difference between the *Xenopus* methylation and that of warm-blooded vertebrates.

Table 1
CpG doublets as distributed in different codon positions

	Xenopus	std	Mouse	std	Man	std	Chicken	std
CpG 1,2	2.8	1.7	3.2	2.1	3.4	2.1	3.5	2.0
CpG 2,3	1.8	1.5	2.6	2.3	2.7	2.2	3.6	2.9
CpG 3,1	3.2	2.0	4.4	3.0	4.6	3.6	6.0	4.7
CpG 1,2 o/e	0.7	0.4	0.7	0.3	0.7	0.3	0.8	0.3
CpG 2,3 o/e	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.2
CpG 3,1 o/e	0.4	0.2	0.4	0.2	0.4	0.2	0.5	0.3

On the other hand, the chicken genome (Olofsson and Bernardi, 1983) is characterized by a lower proportion of repeated sequences (16%), the range being 13% in L1 to 30% in H2. These much lower values, compared with human DNA fractions, show again that the relative amounts of repeated sequences are not the only factor playing a role in the methylation level, a large part of methylation being due to 5mC present in slow-reassociating fractions.

The results of Fig. 2 indicate two major differences between the CpG levels of coding sequences of *Xenopus*, as opposed to those of warm-blooded vertebrates. The first difference is that coding sequences from warm-blooded vertebrates reach higher GC_3 values compared with coding sequences of *Xenopus*. Only very few genes having a GC_3 level higher than 75% exist in *Xenopus*, whereas the relative amounts of genes in mouse, man and chicken are not only much higher than in *Xenopus* but also increasingly so in this order. The second difference is that the slope of the CpG vs GC_3 plot of warm-blooded vertebrates becomes steeper at higher GC_3 levels. This latter difference is due in all likelihood to CpG islands covering increasingly larger 5′ regions of the GC-rich coding sequences (Aïssani and Bernardi, 1991a; Aïssani and Bernardi, 1991b).

Another difference between the *Xenopus* data and those from warm-blooded vertebrates is that the former show a higher observed/expected CpG ratio than the latter (Fig. 3), which is in agreement with overall results from the genomes from cold and warm-blooded vertebrates (Jabbari et al., 1997).

References

Aïssani, B., Bernardi, G., 1991a. CpG islands features and distribution in the genome of vertebrates. Gene 106, 173–183.

Aïssani, B., Bernardi, G., 1991b. CpG islands, genes and isochores in the genome of vertebrates. Gene 106, 185–195.

Bernardi, G., 1985. The organization of the vertebrate genome and the problem of the CpG shortage. In: Cantoni, G.L. & Razin, A. (Eds), Chemistry, Biochemistry and Biology of DNA Methylation. Alan Liss, New York, pp. 3–10.

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

Bernardi, G., 1989. The isochore organization of the human genome. Annu. Rev. Genet. 23, 637–661.

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

Bernardi, G., Bernardi, G., 1990a. Compositional patterns in the nuclear genomes of cold-blooded vertebrates. J. Mol. Evol. 31, 265–281.

Bernardi, G., Bernardi, G., 1990b. Compositional transitions in the nuclear genomes of cold-blooded vertebrates. J. Mol. Evol. 4, 282–293.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.

Cacciò, S., Perani, P., Saccone, S., Kadi, F., Bernardi, G., 1994. Single-copy sequence homology among the GC-richest isochores of the genomes from warm-blooded vertebrates. J. Mol. Evol. 39, 331–339.

Cortadas, J., Macaya, G., Bernardi, G., 1977. An analysis of the bovine genome by density gradient centrifugation fractionation in Cs2SO4/3,6 bis (acetato-mercurimethyl) dioxane density gradient. Eur. J. Biochem. 76, 13–19.

Cortadas, J., Olofsson, B., Meunier-Rotival, M., Macaya, G., Bernardi, G., 1979. The DNA components of the chicken genome. Eur. J. Biochem. 99, 179–186.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G., 1981. The major components of the mouse and human genomes' preparation, basic properties and compositional heterogeneity. Eur. J. Biochem. 111, 227–233.

Duret, L., Mouchiroud, D., Gouy, M., 1994. HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res. 22, 2360–2365.

Feinstein, S.I., Racaniello, V.R., Ehrlich, M., Gehrke, C.W., Miller, D.A., Miller, O.J., 1985. Pattern of undermethylation of the major satellite DNA of mouse sperm. Nucl. Acids Res. 13, 3969–3978.

Gautier C., Jacobzon M., 1989. <http://biom1.univ-lyon1.fr: 8080/doclogi/docanals/manuel.html>, Publication interne, UMR CNRS 5558 Biometrie, Génétique et Biologie des population, Université Claude Bernard-Lyon I, France.

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. CABIOS 1, 167–172.

Jabbari, K., Cacciò, S., Matassi, G., Garmonprez, F., Desgrès, J., Bernardi, G., 1997. Evolutionary changes in CpG and methylation levels in vertebrate genomes. Gene 205, 109–118.

Kadi, F., Mouchiroud, D., Sabeur, G., Bernardi, G., 1993. The compositional patterns of the avian genomes and their evolutionary implications. J. Mol. Evol. 37, 544–551.

Macaya, G., Cortadas, J., Bernardi, G., 1978. An analysis of the bovine genome by density gradient centrifugation. Eur. J. Biochem. 84, 179–188.

Matassi, G., Melis, R., Kuo, K.C., Macaya, G., Gehrke, C.W., Bernardi, G., 1992. Large-scale methylation patterns in the nuclear genomes of plants. Gene 122, 239–245.

Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F., Bernardi, G., 1982. Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. Proc. Natl. Acad. Sci. USA 79, 355–359.

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

Mouchiroud, D., Bernardi, G., 1993. Compositional properties of coding sequences and mammalian phylogeny. J. Mol. Evol. 37, 109–116.

Mouchiroud, D., Gautier, C., Bernardi, G., 1988. The compositional distribution of coding sequences and DNA molecules in human and murids. J. Mol. Evol. 27, 311–320.

Olofsson, B., Bernardi, G., 1983. Organization of nucleotide sequences in the chicken genome. Eur. J. Biochem. 130, 241–245.

Sabeur, G., Macaya, G., Kadi, F., Bernardi, G., 1993. The isochore patterns of mammalian genomes and their phylogenetic implications. J. Mol. Evol. 37, 93–108.

Saccone, S., Cacciò, S., Kusuda, J., Andreozzi, L., Bernardi, G., 1996. Identification of the gene-richest bands in human chromosomes. Gene 174, 85–94.

Saccone, S., Cacciò, S., Perani, P., Andreozzi, L., Rapisarda, A., Motta, S., Bernardi, G., 1997. Compositional mapping of mouse chromosomes and identification of gene-rich regions. Chromos. Res. 5, 293–300.

Saccone, S., De Sario, A., Della Valle, G., Bernardi, G., 1992. The highest gene concentrations in the human genome are in T-bands of metaphase chromosomes. Proc. Natl Acad. Sci. USA 89, 4913–4917.

Saccone, C., De Sario, A., Wiegant, J., Rap, A.K., Della Valle, G., Bernardi, G., 1993. Correlations between isochores and chromosomal bands in the human genome. Proc. Natl Acad. Sci. USA 90, 11929–11933.

Salinas, J., Zerial, M., Filipsky, J., Bernardi, G., 1986. Gene distribution and nucleotide sequence organization in the mouse genome. Eur. J. Biochem. 160, 469–478.

Soriano, P., Macaya, G., Bernardi, G., 1981. The major components of the mouse and human genomes: reassociation kinetics. Eur. J. Biochem. 115, 235–239.

Soriano, P., Macaya, G., Bernardi, G., 1983. The distribution of interspersed repeats is non-uniform and conserved in the mouse and human genomes. Proc. Natl Acad. Sci. USA 80, 1816–1820.

Thiery, J.P., Macaya, G., Bernardi, G., 1976. An analysis of eukaryotic genomes by density gradient centrifugation. J. Mol. Biol. 108, 219–235.

Zerial, M., Salinas, J., Filipsky, J., Bernardi, G., 1986. Gene distribution and nucleotide sequence organization in the human genome. Eur. J. Biochem. 160, 479–485.