ELSEVIER

Gene 224 (1998) 123–128

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families

Kamel Jabbari, Giorgio Bernardi *

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

Received 2 July 1998; received in revised form 1 September 1998; accepted 2 September 1998; Received by T. Gojobori

Abstract

A computer analysis of 946 human DNA sequences larger than 50 kb and representing about 118 Mb of DNA has led to the following observations. (i) Positive correlations hold between CpG levels and the GC levels of isochores and coding sequences, as expected from previous results. (ii) The correlation between CpG levels and the GC levels of pseudogenes is characterized by lower CpG values (at comparable GC levels) and by a lower slope compared with the correlation with coding sequences; this finding suggests that an extensive methylation followed by deamination has taken place on CpG doublets from inactive genes leading to a further CpG shortage. (iii) The frequency of CpG islands in long human sequences increases with increasing GC and almost parallels gene frequency. (iv) The frequency of Alu sequences also increases with increasing GC, but attains a maximum in H2 isochores, in agreement with previous experimental data. (v) The ratio 5mC/CpG (namely, the methylation level over available sites) decreases with increasing GC levels of isochores. This decrease is due only to a small extent to the increase of (unmethylated) CpG islands in GC-rich isochores, and takes place in spite of the increase of strongly methylated Alu sequences in GC-rich isochores; this stresses the much lower relative methylation (5mC/CpG) of single-copy sequences located in GC-rich isochores relative to those located in GC-poor isochores. (vi) CpG levels of Alus and CpG islands are positively correlated with the GC levels of the long sequences in which they are located. (vii) The CpG levels of both Alus and CpG islands increase with their GC levels. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Human genome; Repeated sequences

1. Introduction

The availability of an increasing number of long sequences belonging to different isochore families in databanks, now makes it possible to investigate some properties of isochores, the long, compositionally homogenous DNA segments making up vertebrate genomes, at the level of primary structure, and to compare these properties with those observed previously on compositional fractions of DNA (for a review, see Bernardi, 1995).

In the present work we have taken advantage of about 118 Mb of human sequences longer than 50 kb to investigate the following points: (i) the correlations between CpG levels and the GC levels (GC is the molar fraction of guanine+cytosine in DNA) of isochores, coding sequences and pseudogenes; (ii) the correlation between the 5-methylcytosine/CpG ratio (5mC values being obtained from Cacciò et al., 1997) and the GC levels of isochores; (iii) the correlations between the frequencies of CpG islands and Alu sequences, respectively, and the GC levels of isochores; (iv) the correlations between GC and CpG levels of Alus or CpG islands and the GC levels of the isochores containing them; and (v) the correlations between the CpG levels of Alus or CpG islands and the GC levels of these sequences.

2. Materials and methods

2.1. Databank sequences analysis

946 sequences longer than 50 kb were collected from GenBank (Release 115; February 1998) using the ACNUC retrieval system (Gouy et al., 1985). The program ANALSEQ (Gautier and Jacobzone, 1989) was used to determine the base composition and doublet frequencies of 6682 non-redundant human coding

* Corresponding author. Tel: +33 1 4427 7972; Fax: +33 1 4427 7977; e-mail: bernardi@citi2.fr

0378-1119/98/$ – see front matter © 1998 Elsevier Science B.V. All rights reserved.
PII: S0378-1119(98)00474-0

sequences; this dataset was obtained by using HOVERGEN (Duret et al., 1994), a database in which genes from vertebrates are grouped into homologous families on the basis of amino-acid sequence similarity. Pseudogenes were extracted from HOVERGEN, using 'pseudo' as a keyword to search the database; the small size of the resulting data set allowed its manual cleaning. 118 Mb (118 665 550 bp) of DNA were selected and assigned, according to their average GC levels, to four isochore families, L (i.e., L1+L2), H1, H2, and H3. The three isochore boundaries were taken as 41%, 46%, 53% GC (Zoubak et al., 1996); 37% GC was taken as the boundary L1 and L2.

Coding sequences were assigned to isochore families on the basis of their GC_3 values (GC_3 is the average GC level of third codon positions) and of the correlation between GC_3 level of human genes and the GC level of isochores in which they are embedded (Clay et al., 1996; Zoubak et al., 1996).

The observed CpG frequency and the CpG frequency expected from the base composition was assessed for each long sequence. CpG islands are defined here as sequences longer than 500 bp, with an average GC level higher than 50% and a frequency of CpG greater than 60% of the statistically expected value (see Gardiner-Garden and Frommer, 1987; Larsen et al., 1992). CpG islands were determined by moving window CpG density plots (200 bp window size), using the GCG window program (Devereux et al., 1984). Alu sequences were identified using the program RepeatMasker (Jurka et al., 1996; Smit and Green, 1996).

3. Results and discussion

3.1. Correlations between CpG and GC levels of isochores and coding sequences

As a preliminary remark, it should be noted that the long human genomic sequences (>50 kb) were well distributed among isochore families, as judged by comparing (Fig. 1) the histogram of GC levels with the CsCl profile of human DNA (>50 kb).

The results obtained on CpG doublets from long sequences, assigned to four isochore families L (i.e., L1+L2), H1, H2, and H3 on the basis of their GC levels (see Section 2), showed that CpG frequencies exhibit (Fig. 2A) a linear increase with GC of isochore families. A similar increase was observed also when plotting CpG against GC of coding sequences, as assigned to isochore families on the basis of the criteria indicated in Section 2.

The linear increase of CpG with increasing GC of long sequences and coding sequences confirms and considerably extends previous findings (Aïssani and Bernardi, 1991a,b) which concerned coding sequences, introns and flanking sequences. The shifts to the top right of the coding-sequence points relative to the points of long DNA sequences (Fig. 2A) is due to the higher GC and CpG levels of coding sequences, respectively, compared to non-coding sequences. The plots of observed/expected CpG ratios (Fig. 2B) exhibited by

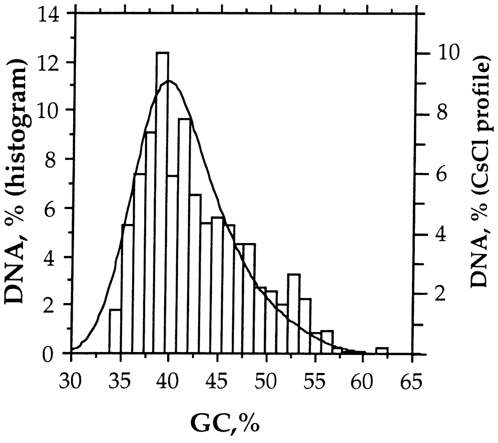

Fig. 1. The relative amounts of long (>50 kb) human DNA sequences in 2.5% GC (histogram) are compared with the CsCl profile of human DNA (>50 kb).

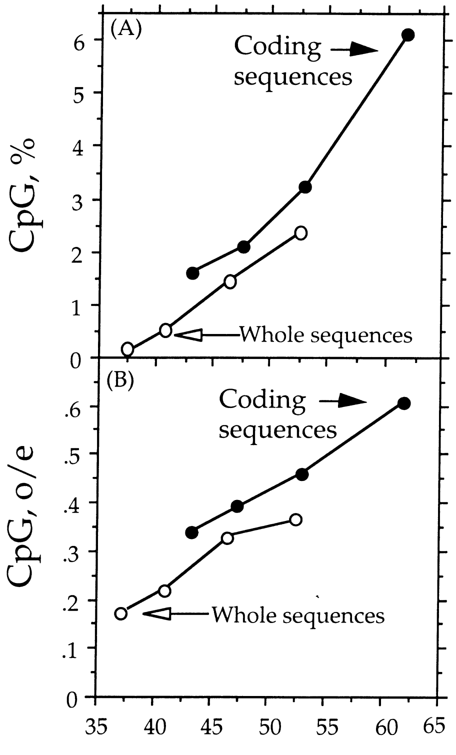

Fig. 2. Plots of CpG (A) and CpG o/e (B) versus GC of 946 long (>50 kb) human genomic DNA sequences or of the coding sequences divided into four classes corresponding to isochore families L1+L2, H1, H2, and H3 (see Section 2).

the long sequences and by coding sequences were similar to those of Fig. 2A.

3.2. Correlation between CpG and GC levels of pseudogenes

Fig. 3 compares the CpG vs GC plots of coding sequences (A) and pseudogenes (B). While the first plot shows a positive correlation with an increasing slope, as expected from the data of Fig. 2A, the second one is characterized by comparatively lower values of CpG for corresponding GC levels. The CpG levels of some pseudogenes, for T-cell binding protein (DB protein), dihydrofolate reductase, calmodulin and zeta globin, are higher than average.

3.3. CpG islands and Alu sequences in human isochores

CpG islands were found to represent 0.34%, 0.89%, 1.70%, 3.19%, and 5.47% of the long sequences belonging to the L1, L2, H1, H2, and H3 isochore families, respectively (see Fig. 4A). This histogram is similar to that (Fig. 4B) corresponding to the gene distribution in the same isochore families (Zoubak et al., 1996). The data of Fig. 4A and B show that density of CpG islands increases by a factor of 15–16 between L1 and H3. This ratio is close to the ratio, 17, of the gene concentrations in (L1 + L2) and H3 (Zoubak et al., 1996), the difference arising, in all likelihood, from the presence of CpG island-negative genes in L1 isochores.

Alu repeats were found to account for 4.32%, 6.36%, 12.98%, 19.45%, and 16.30% of the long sequences located in the L1, L2, H1, H2, and H3 isochore families, respectively (Fig. 4C). Interestingly, about 23% of all CpG dinucleotides of the human genome were found to be present in Alu repeats, with only minor differences in the different isochore families. The contribution of

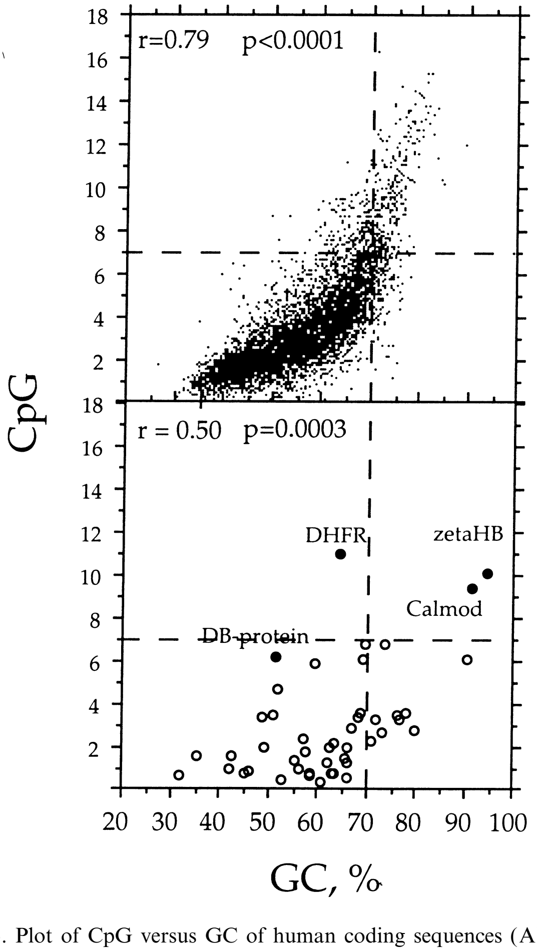

Fig. 3. Plot of CpG versus GC of human coding sequences (A) and pseudogenes (B); solid circles correspond to some pseudogenes exhibiting relatively high CpG levels (see text). Broken lines at 7% CpG and 70% GC are shown to provide a reference. Note that a plot of CpG vs GC_3 concerning the same sample of coding sequences was published by Cacciò et al. (1997).

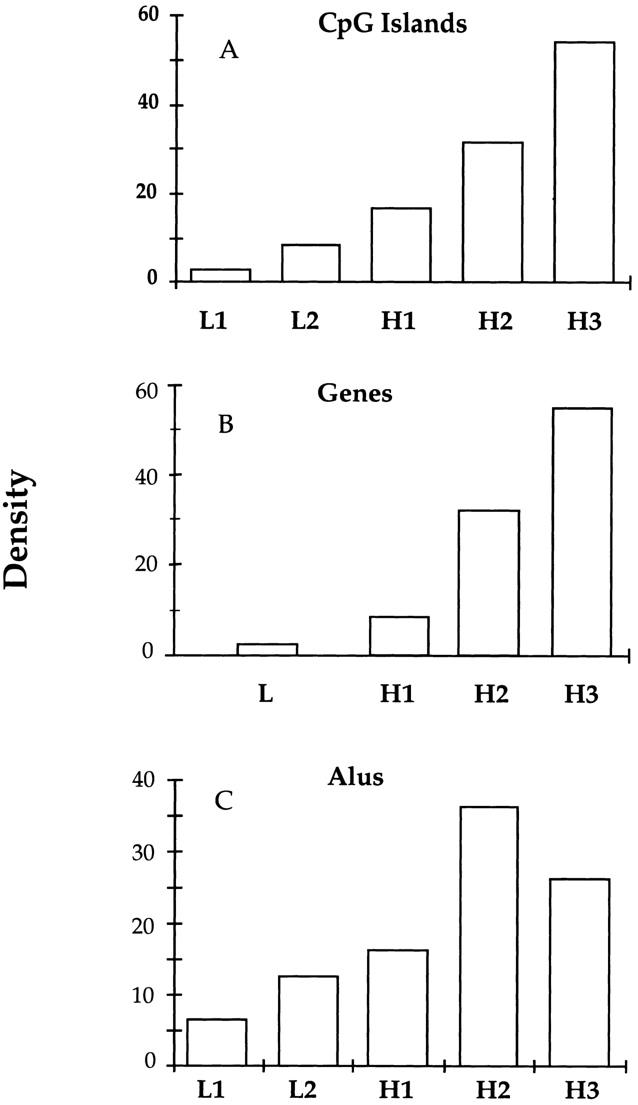

Fig. 4. Density of CpG islands, genes (from Zoubak et al., 1996) and Alu sequences in isochore families. Relative numbers of sequences over relative amounts of isochore families are presented in the histograms.

Alu repeats to the methylation pattern of the human genome is of interest, because Alus are known to represent approx. 10% of the human genome (Kochanek et al., 1993; 12%, in our analysis) and to account for up to 30% of the total methylation level (Hellmann-Blumberg et al., 1993). Since the frequency of Alu repeats increases by a factor of almost 5 from L1 to H2 in the sequences analyzed (see Fig. 4C), in agreement with previous experimental data (Zerial et al., 1986), the gradual increase in the methylation level of human isochores is due in part to the uneven distribution of these repeats, i.e., to their concentration in the GC-rich part of the human genome.

In connection with the results on Alu sequences, it should be mentioned that the experimental results of Zerial et al. (1986) are confirmed here on the basis of the analysis of long human sequences (see Fig. 4B). Incidentally, these results were claimed to be contradicted by an analysis of the distribution of Alu sequences in introns, which showed approximately equal levels in L1 + L2, H1 + H2, and H3 isochore families (Duret et al., 1995). This discrepancy is due to the fact that in the first study the frequencies of Alu sequences were measured in isochores, whereas in the second they were measured only in introns and not in intergenic DNA. There is, therefore, no contradiction between the two sets of data, and the supposed argument against the preferential integration of Alu repeats into compositionally matching isochores is not supported by these findings. In fact, the composition of Alu sequences does tend to match that of the isochores in which they are embedded (see below). In contrast, what the data of Duret et al. (1995) do show is that Alu sequences tend to avoid the introns of the GC-richest genes.

3.4. *Correlations between 5mc/CpG and GC levels of long sequences and coding sequences*

The 5mC/CpG ratio was calculated using the 5mC values determined in Cacciò et al. (1997) for compositional fractions of placental human DNA and the CpG frequencies observed in the long sequences. This ratio decreased with increasing GC (Fig. 5). Since this result is influenced by the CpG islands, which are known to be unmethylated, and by the higher methylation of Alu sequences in somatic cells, the contribution of these elements to the long human sequences was investigated.

The contribution of CpG islands to the observed 5mC/CpG ratio in the long DNA sequences was assessed and the ratio was recalculated excluding CpG islands, since they are unmethylated. However, after correction for CpG islands, the ratio 5mC/CpG still decreased with increasing GC.

The contribution of Alu sequences to the observed 5mC/CpG ratio was also assessed and the ratio was recalculated excluding Alu sequences, since they are hypermethylated. As expected, the ratio 5mC/CpG further decreased with increasing GC. Fig. 5 also shows the results obtained after correction for the contribution of CpG islands and Alu sequences.

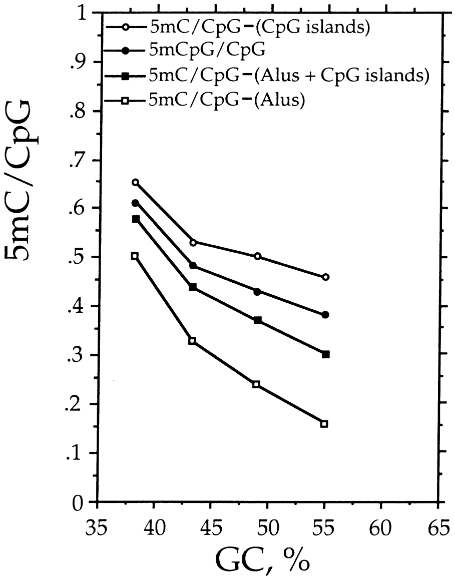

Fig. 5. The ratios of 5mC, as estimated from the data of Cacciò et al. (1997), to CpG of long sequences (solid circles), or of long sequences minus CpG islands (open circles), minus Alus (open squares) and minus Alus and CpG islands (open triangles) are plotted against GC of long sequences.

The results of Figs. 2 and 3A lead to two straightforward conclusions: (i) at comparable GC level, CpG levels are higher in coding than in non-coding sequences; and (ii) at high GC levels, the slope of the plot CpG vs GC of coding sequences is higher compared to that at low GC levels.

Finding (i) may be due to the fact that CpG tends to be less methylated and subsequently deaminated in coding versus non-coding sequences. Finding (ii) may be due to an increasing contribution of CpG islands covering the 5′ ends or covering coding sequences; this stresses the fact that coding sequences undergo a less severe loss of CpGs, being less methylated than non-coding sequences.

The results obtained with pseudogenes (Fig. 3B) suggest that, upon loss of function, CpG doublets undergo methylation, deamination of 5mC to T and a selective loss of CpG. Pseudogenes showing higher CpG values might correspond to sequences that have been inactivated more recently. More detailed studies on this point are in progress and will be reported elsewhere.

3.5. *Correlation between GC and CpG levels in Alus or CpG islands and the GC level in the long sequences*

Figs. 6 and 7 show that significant positive correlations exist between the GC and CpG levels of Alu

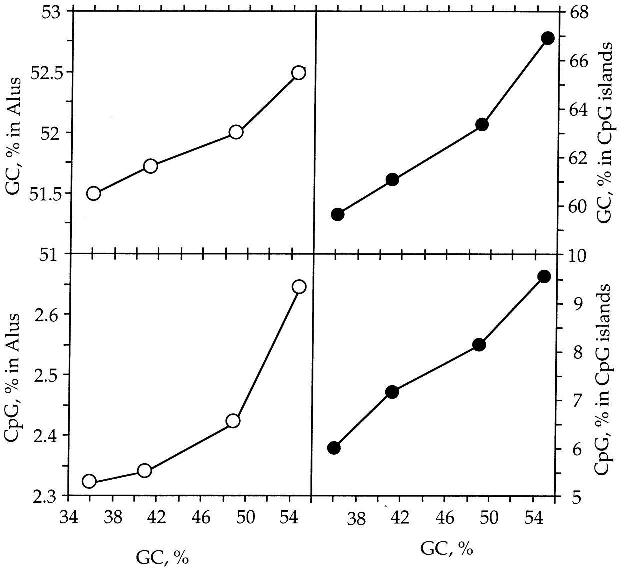

Fig. 6. Correlation of GC and CpG levels in Alus or CpG islands and GC levels of long sequences in which they are located.

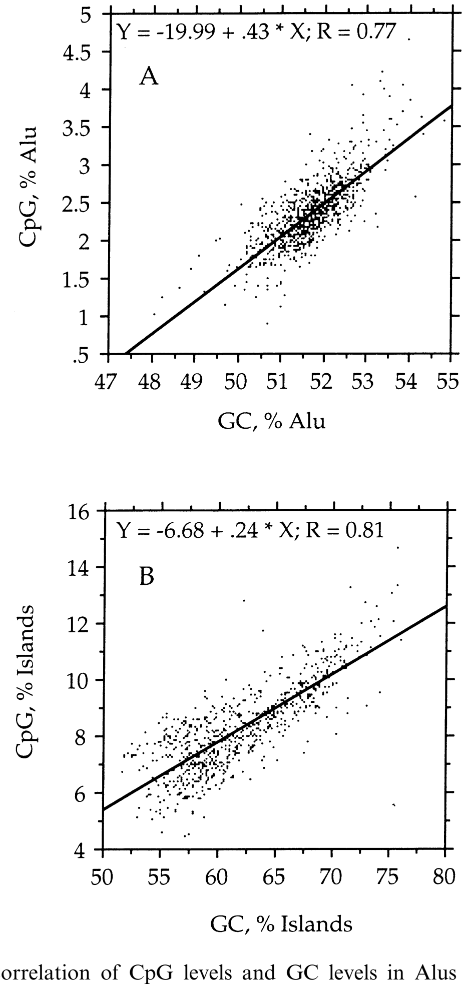

Fig. 7. Correlation of CpG levels and GC levels in Alus and CpG islands.

sequences or CpG islands and the GC levels of the long sequences containing them, on one hand, and between CpG levels in Alu sequences or CpG islands and their GC levels.

While the latter result is somehow expected, the former is of great interest because it indicates that both Alu sequences and CpG islands, which are thought to play a regulatory role in transcription and replication (see Chu et al., 1998,Delgado et al., 1998 Federico et al., 1998) undergo compostional constraints, as do regulatory sequences of retroviruses, the long terminal repeats (Zoubak et al., 1992; see also Rynditch et al., 1998, for a recent review).

Acknowledgements

We thank Oliver Clay and Philippe Dessen for their interest and advice.

References

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

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

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

Cacciò, S., Jabbari, K., Matassi, G., Garmonprez, F., Desgrès, J., Bernardi, G., 1997. Methylation patterns in the isochores of vertebrate genomes. Gene 205, 119–124.

Chu, W.M., Ballard, R., Carpick, B.W., Williams, B.R., Schmid, C.W., 1998. Potential Alu function: regulation of the acivity of double-stranded. Mol. Cell. biol. 18, 58–68.

Clay, O., Caccio', S., Zoubak, S., Mouchiroud, D., Bernardi, G., 1996. Human coding and non-coding DNA: compositional correlations. Mol. Phylogen. Evol. 5, 2–12.

Delgado, S., Gomez, M., Bird, A., Antequera, F., 1998. Initition of DNA replication at CpG islands in mammalian chromosomes. EMBO J. 17, 2426–2435.

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

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

Duret, L., Mouchiroud, D., Gouy, M., 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. 40 (3), 308–317.

Federico, C., Saccone, S., Bernardi, G., 1998. The gene-richest bands of human chromosome replicate at the onset of the S-phase. Cytogenet. Cell. Genet. 80: 83–88.

Gardiner-Garden, M., Frommer, M., 1987. CpG islands in vertebrate genomes. J. Mol. Biol. 20 (196), 261–282.

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

Gouy, M., Gautier, C., Attimonelli, N., Lanave, C., Di Ppaola, G., 1985. ACNUC — Portable retrieval system for nucleic acid sequence database: logical and physical design and usage. Cabios 1, 167–172.

Hellmann-Blumberg, U., Hintz, M.F., Gatewood, J.M., Schmid, C.W.,

1993. Developmental differences in methylation of human Alu repeats. Mol. Cell. Biol. 13, 4523–4530.

Jurka, J., Klonowski, P., Pelton, P., 1996. CENSOR — a problem for identification and elimination of repetitive elements from DNA sequences. Comput. Chem. 20, 119

Kochanek, S., Renz, D., Doerfler, W., 1993. DNA methylation in the Alu sequences of diploid and haploid primary human cells. EMBO J 12, 1141–1150.

Larsen, F., Gundersen, G., Lopez, R., Prydz, H., 1992. CpG islands as gene markers in the human genome. Genomics 13, 1095–1107.

Rynditch, A.V., Zoubak, S., Tsyba, L., Tryapitsina-Guley, N. Bernardi, G., 1998. The regional integration of retroviral sequences into the mosaic genomes of mammals. Gene. In press (ms. 11571).

Smit, A.F.A., Green, P., 1996. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html.

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.

Zoubak, S., Rynditch, A., Bernardi, G., 1992. Compositional bimodality and evolution of retroviral genomes. Gene 119, 207–213.

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