Vol. 213, No. 1, 1995 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
August 4, 1995 Pages 147-153

LOOP-SIZE SPACINGS BETWEEN CGCG CLUSTERS IN LONG SEGMENTS OF HUMAN DNA

N. AVRIL[1], P. DESCHAVANNE[1], M. BELLIS[2] and J. FILIPSKI[1,*]

[1]Laboratoire de Mutagénèse Institut J. Monod. 2, Place Jussieu Tour 43, 75251 Paris, France

[2] Institut de Biologie, UPR 9008 CNRS - U249 INSERM, 4 Bd Henri IV 34000, Montpellier, France

Received June 19, 1995

The CGCG tetranucleotides are clustered inside the CpG islands in the genomes of vertebrates. In order to study the distribution of the islands in the human chromosome we have mapped the loci sensitive to the CGCG specific restriction nuclease, in a 1.5 Mb long DNA segment cloned as Yeast Artificial Chromosome (YAC). The sites most sensitive to Bsh 1236 I nuclease show chromosomal loop-size spacing. This result, as well as the result of nucleotide sequence analysis of long genomic segments, suggests that the CGCG are organised in clusters (not always undermethylated) which are coincident with GC peaks on the sine wave-like curve representing DNA composition along the mammalian chromosome.

The non random distribution of CpG dinucleotides in vertebrate genomes has been discovered by DNA sequence studies (1) and by restriction enzyme analysis of mouse genomic DNA. Bird *and coll.* (2) have demonstrated that a fraction of mouse DNA could be converted to low molecular weight (about 10^2 bps) fragments by digestion with Hpa II (which cuts within CCGG tetranucleotide providing that the second cytosine is not methylated). The finding that as much as 1% of DNA is present as such a fraction came as a surprise : the CpG dinucleotides are rare in the genomes of vertebrates and about 70% are methylated (3). The discovery of these, so called, "Hpa II Tiny Fragments" (HTF), indicated that 1°- a considerable proportion of CpG dinucleotides is clustered, and 2°- the CpG dinucleotides in these clusters remain unmethylated. The DNA regions from which the HTF were excised were named "CpG islands". This unusual sequence organisation suggested that the CpG islands have a special function in genomes. It appeared that they are located at the 5' end of some genes frequently overlapping their regulatory sequences and first exons. The genes accompanied by the islands usually replicate during early period of the "S" phase of the cell cycle, are GC-rich, and are located within chromosomal R bands (4). The islands are believed to be involved in regulation of transcriptional activity of accompanying genes. They may constitute a preferred site of

*To whom correspondence should be addressed. Fax: (33 1) 44 27 57 16.

0006-291X/95 $12.00

interaction of DNA binding proteins with regulatory sequence elements. The CpG dinucleotides in the islands are protected against methylation and the chromatin in the islands resembles the chromatin of transcribed genes (5).

Here we wanted to determine the long range distribution of CpG islands in the human genome taking advantage of the finding that the CGCG tetranucleotides are frequent within the islands (in the range of 5 such sites or more per island, 6) while they are rare in the island-free genomic regions (for example in the 70 kb long β-globin gene cluster there are only five isolated CGCG sites). To do so, we mapped the loci sensitive towards the restriction nuclease recognising these tetranucleotides in a 1.5 Mb long YAC carrying segment of DNA extracted from human chromosome 21. In this YAC we have found that the CGCG sites are organised in relatively regularly distributed clusters. Statistical analysis of a DNA sequence extracted from the data bank suggests that the clustered distribution of CGCG tetra nucleotides is related to a sine wave-like long-range pattern of the DNA composition.

MATERIALS AND METHODS

Preparation and treatment of spheroplasts. Yeast were grown overnight with vigorous agitation in 100 ml of YP medium at 30°C to the cell density of about 10^8 cells/ml, harvested by centrifugation at 1600 g, resuspended in 3 ml of buffer containing 1M sorbitol, 50 mM Tris-HCl pH 7.8, 10 mM $MgCl_2$, 30 mM DTT and incubated 15 min at room temperature. After the incubation the cells were centrifuged as before, washed in 2 ml of a similar buffer which contains only 2 mM DTT and resuspended in 2 ml of this buffer supplemented with 0.3 ml 120 mg/ml glucanase (purchased from Novo Nordisk Ferment AG) and incubated at 30°C. The formation of spheroplasts was followed by removing periodically aliquots of 10 µl of the suspension and mixing them with 1 ml of distilled water. The spheroplasts as opposed to intact cells swell and burst when transferred from isotonic buffer to distilled water which results in a decrease in OD at 600 nm to about 10 % of its initial value. The resulting suspension of spheroplasts was washed twice with 4 ml of the same buffer without glucanase, centrifuged 5 min at 1600 g and resuspended in 400 µl of the same buffer. After a control of the final volume, the suspension was mixed with the same volume of liquid (45°), 2% LMP agarose (Beckman Instruments) prepared on the same buffer, and immediately poured into glass moulds and cooled on ice. Solidified slabs were cut into plugs 6mm x 5mm x 1mm containing about 2 µg of DNA each. The plugs were incubated overnight at 37°C in 3 ml of lysing buffer containing 0.5 M NaCl, 0.2 M EDTA, 0.125 M TRIS-HCl pH 8, 1% SDS and 50µl 20 mg/ml proteinase K (purchased from Boehringer Mannheim). The plugs were subsequently washed in several changes of the storage buffer containing 50 mM EDTA, 10 mM Tris-HCl pH 8 and 90 mM NaCl, saturated with chloroform.

Digestion of DNA in plugs.

The plugs were washed several times with appropriate buffer complemented with 0.5mM PMSF. Partial digestions were performed using serial dilutions of Bsh1236I (purchased from Eurogentec), one among several restriction nucleases recognising CGCG tetranucleotide. Using batch of enzyme having specific activity 12 units/µl the dilution were ranging from 1/1,000 to 1/50,000. The digestion was performed during 18 hours in the recommended buffer, and was stopped with EDTA. Digestion by NotI (specific activity 5 units/µl was performed overnight at 37°C using 10 units of enzyme in a volume of 250 µl for 6 plugs.

Pulsed field gel electrophoresis.

The fragments were separated on a 1% agarose gel at 200 V in 0.5 x TBE buffer at 10°C, in a Beckman GeneLine pulse field electrophoresis system in the same buffer as described (7). λ DNA digested by Hind III (1-23 kb) and λ DNA (New England Biolabs) ligated to form concatemers (<50-600 kb) were used as molecular weight markers. Two different conditions of separation were used depending on the required resolution Fractionation of the molecules ranging from 50 to 600 kb were done using pulse time 25 s for 25 h and the molecules 50 to 250 kb long were separated using 1 s pulses for 8 h followed by 10 s pulses for 10 h and 15 s pulses for 8 h. The gel slabs after coloration with ethidium

bromide were put on UV transilluminator and the pictures were taken using a Polaroid instant camera .

Hybridisation .

The DNA from gels was transferred to Hybond N^+ nylon membrane (Amersham) using alkaline buffer (8). The hybridisations were done using Rapid-hyb buffer (purchased from Amersham) under the conditions recommended by the manufacturer using probes labelled with α^{32}P-dATP and the Megaprime labelling system provided by Amersham. The filters were hybridised in an oven (Techne) at 65°C, were subsequently washed with 2 X SSPE, 0.2 % SDS and were autoradiographed using medical X-ray Fuji films at -80°C.

YAC and probes used in this study.

The Yeast Artificial Chromosome 881 D2 obtained from Centre d'Etudes du Polymorphisme Humain (CEPH) in Paris contains a 1.45 Mb long segment of human DNA extracted from the region 21q11.2 of human chromosomal complement. It contains two Not I restriction sites corresponding to chromosomal loci D21S13 and LL56 and hybridises to a series of Sequence Tagged Sites mapped in this region. Distance between the two Not I sites is equal within experimental error to the distance obtained by restriction enzyme digestion of genomic DNA (9) which suggests that this YAC did not undergo extensive rearrangements during cloning.

The YAC 316 C4 carrying 415 kb long DNA segment extracted from the CpG island -free region 11p15.5 carrying human β globin gene cluster has been used as a "negative control".

In order to map the clusters of CGCG tetranucleotides in the studied YAC, the following probes were used in the indirect end-labelling experiments. The position of the probes are shown on Fig 2.

"C" is the Pst I - Pvu II fragment of pBR 322. It hybridises to the YAC vector arm which contains both telomere and centromere.

"T" probe is the Bam HI - Pvu II fragment of pBR322, which hybridises to the opposite YAC vector arm.

Probe "L" is the Not I-EcoR I low molecular weight fragment (3 kb) of the plasmid pGSM 21 E carrying DNA sequences excised by EcoR I from the chromosomal locus D21S13 containing Not I site.

"H" is the NotI - EcoRI high molecular weight (6 kb) fragment of the same plasmid.

DNA sequence analysis.

The DNA sequence of the 70 Kb long genomic DNA containing the genes coding the human growth hormone GH1 and GH2 and chorionic somatomammotropin CS 1, 2 and 5 was extracted from the GeneBank and analysed using a retrieval package installed on a central computer in Paris (Centre Interuniversitaire de Traitement d' Information) (10).

RESULTS AND DISCUSSION

The CpG islands are usually mapped in genomes with the help of a battery of rare-cutter restriction enzymes like Not I (recognition sequence : GCGGCCGC), SacII (CCGCGG), Eag I (CGGCCG) and others having usually multiple CG doublets within a six- or eight-nucleotide long recognition sequence. However, not every island contains one of these rare sites. In this case the mapping would produce a "false negative". Presence of one of these sites outside the island would produce a "false positive". The work described here has been undertaken with the aim of developing a complementary procedure which may help to overcome these problems. We used for the mapping the restriction enzyme Bsh1236 I recognising CGCG tetranucleotide which are clustered within the islands. The enzyme concentration was low so as the digestion was only partial but still the chances were relatively high that the enzyme would cut one of the recognition sites present within a cluster. Given that the resolution of the electrophoresis used throughout the experiments reported here was usually not higher than 5 kb, a set of fragments extending from one end of the chromosome to a cluster of CGCG sites produced a single band on the autoradiogram. On the other hand the probability that the enzyme would cut the DNA

in the isolated CGCG site were low under the condition of partial digestion. Thus the unclustered CGCG tetranucleotides produced only a weak background.

For our study we have chosen the YAC 881 D2 constructed from DNA extracted from the 21q11.2 region of human chromosomal complement, from one of the chromosomal R bands. These bands are carrying mostly GC-rich genes and are rich in CpG islands (3). The studied YAC carries two markers identified as Not I linking clones : LL56 and D21S13 (11). This latter locus and which is coincident with an identified CpG island is tightly linked to Alzheimer disease (12).

The result of a typical mapping experiment is shown on Fig 1. Pairs of the agarose blocks, one containing (in addition to the set of the yeast endogenous chromosomes) the YAC 881 D2 and the second the YAC 316 C4, were treated with the restriction enzyme. The highest enzyme concentration corresponds to the lanes 1 and 2 and the lowest to the lanes 5 and 6. The lanes 7 and 8 correspond to the reaction mixtures containing no enzyme. The YAC 316 C4 carries β-globin gene cluster. It was included in the experiment in order to enable us to follow

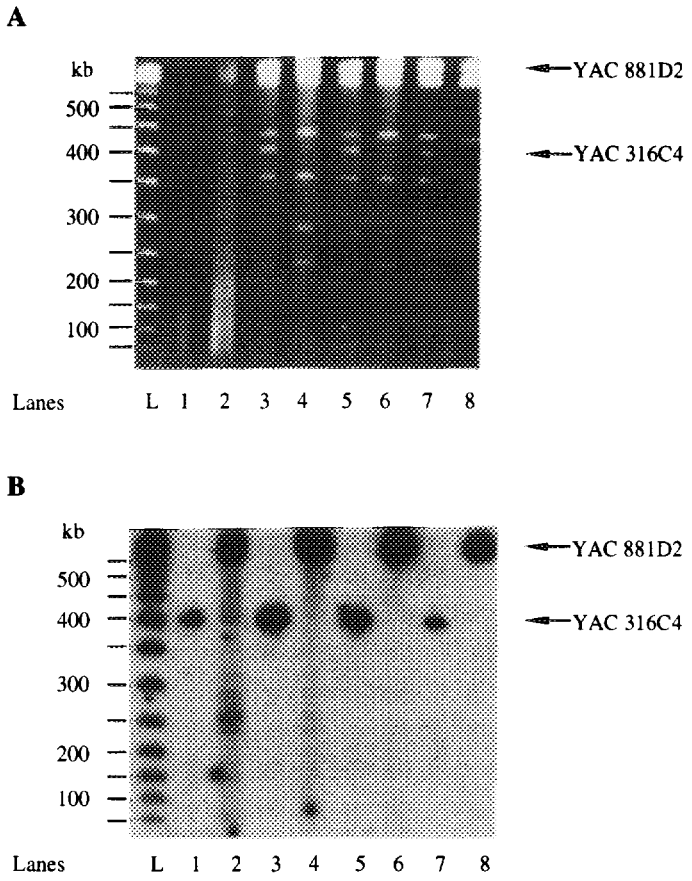

Fig. 1. Indirect end labelling of Bsh1236 I restriction site clusters in the YAC 881 D 2 (lanes 2, 4, 6,) and YAC 316 C4 (lanes 1, 3, 5) : Lanes 1 and 2, 3 and 4 and 5 and 6 contained the products of digestion by the enzyme diluted 10^3, $2x10^4$ and $5x10^4$ times,respectively. Lanes 7 and 8 contained no enzyme, lane L : molecular weight standard (λ ladder) A. EtBr staining, B. Hybridisation pattern using probe "T".

the digestion of isolated CGCG tetranucleotides. The blocks containing the digestion products were subsequently embedded in a single gel slab (in the even and odd lanes respectively) and electrophoresed in a pulsed field. Fig 1A shows the gel after the electrophoresis and ethidium bromide staining. Blotting and hybridisation with the ^{32}P labelled telomeric probe T produced a series of bands in the autoradiogramm (Fig. 1B). In the lanes 8 and 7 (no enzyme) and 6 and 7 (the lowest concentration of the enzyme) only the bands corresponding to the intact artificial chromosomes 881 D2 and 316 C4 respectively are visible. In lane 4 containing the YAC 881 D2 a series of bands corresponding to the loci most sensitive to the employed restriction enzyme becomes apparent while lane 3 containing the products of partial digestion of the YAC 316 C4 shows no bands. In lane 2, corresponding to the most extensive digestion of the YAC 881 D2, the bands are "stronger" while in lane 1 containing the YAC 316 C4 incubated in the same tube only weak bands are visible even after longer exposure compensating for differences in the DNA concentration in the blocks.

Fig 2 shows the map of the studied YAC established on the basis of the experiments similar to the one presented in Fig 1. The loci LL56 and the D21S13 are the sites the most sensitive to the restriction nuclease employed here indicating the presence of multiple CGCG

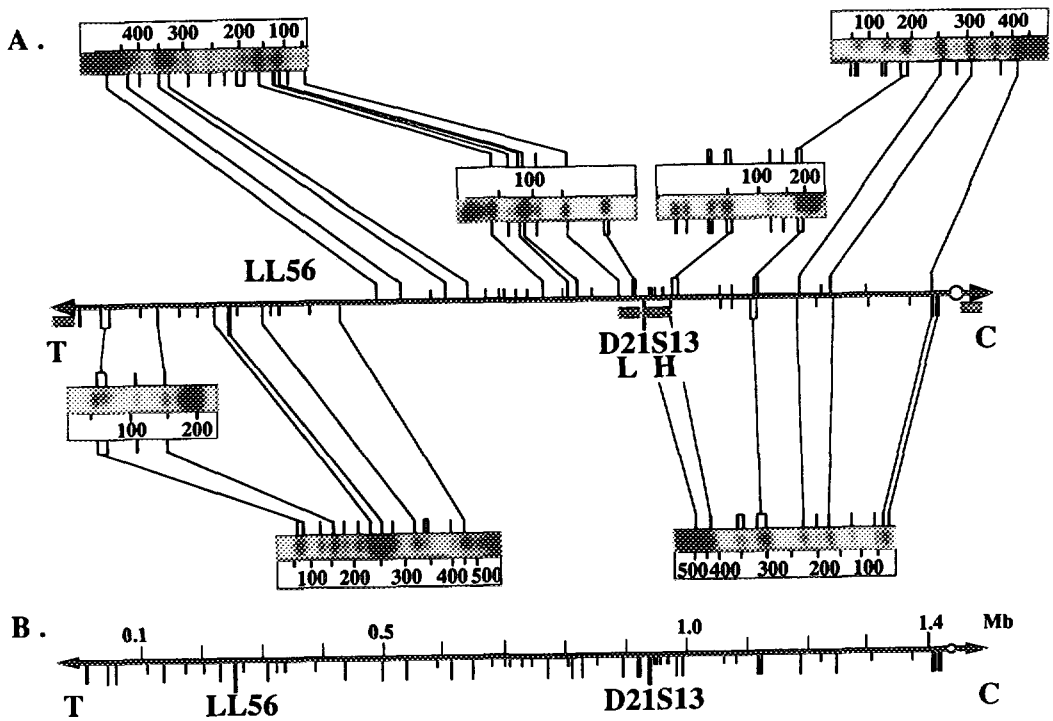

Fig. 2. The map of the CGCG clusters in the YAC 881 D 2 : A. Positions of the bands seen on the autoradiogramms obtained in the experiments similar to the one shown on Fig 1 were marked on the schematic representation of the YAC. Positions of the probes used in the mapping are shown as hatched rectangles labelled T, C, L and H,respectively. Two loci corresponding to Not I sites localized on the chromosome 21 as Sequence Tagged Sites (STS) are labelled D21S13 and LL56,respectively. B. The map of the YAC. Numbers correspond to the map position in Mb starting from the T end of the YAC. Long and short marks correspond to the "strong" and "weak" bands on the autoradiagramms shown in the panel "A".The longest marks correspond to the map positions of the STS.

tetranucleotides. This suggest that both are localised within CpG islands, the high number of CGCG sites suggesting that DNA in these loci is unmethylated and thus protected against the loss of CpG dinucleotides. Presence of several other relatively "strong" bands suggests that there may be more CpG islands in this chromosomal region. We consider however, that the most interesting finding of this study is the demonstration of the regularity with which the CGCG clusters are distributed. In particular, in the 400 kb long segment located at the T end of the studied YAC (fig 1 lane 2) one can distinguish 13 bands (including one double band) regularly spaced with an average distance of 30.1 kb between each other. The 500 kb long C terminal region of this YAC shows somewhat lower density of the CGCG clusters, in the range of 1 for every 50 kb. The differences in the frequency of the CGCG clusters suggests that the two extremities of this YAC are built of different isochores (the long stretches of DNA showing relatively homogeneous average base composition 13).

Interestingly a similar regularity in the distribution of CG-rich loci in DNA has been found by mapping human 11p13 locus with the rare cutter restriction enzymes (14). Only some of the mapped loci were unmethylated fitting the definition of CpG islands. The authors called the studied locus the "CpG island archipelago". Our result suggests that this organisation is not an exception.

In order to get a closer look at the genomic distribution of the CGCG sites we have analysed a long, moderately GC-rich DNA sequence carrying a cluster of human growth hormone and somatomammotropine genes extracted from the GeneBank. No CpG islands were identified there. The sequence has been scanned using 10 kb wide window and 1kb steps. Surprisingly, the composition of the analysed fragment of DNA shows sinewave-like pattern (Fig 3). The two clusters of the CGCG tetra nucleotides found there are roughly coincident with two regions of the highest GC content (53 % GC). The distance between the first cluster (7 sites distributed in two packs in the region between 4 and 14 kb on the map) and the second

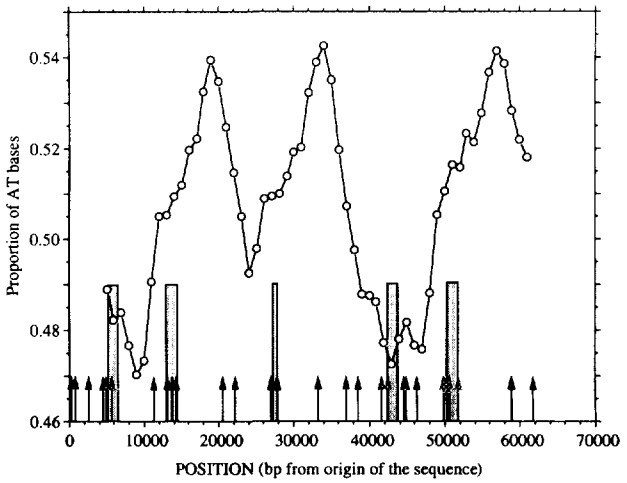

Fig. 3. CG content and CGCG restriction sites in the GC-rich segment of the DNA carrying the human GH1, GH2 and CS1, 3 and 5 genes : Hatched rectangles represent coding regions and arrows indicate the CGCG tetranucleotides found in this gene cluster.

one (8 sites in the region between 41 and 52kb on the map) is about 33 kb. Several other examples of such a clustering of CGCG tetranucleotides in GC-rich peaks in long stretches of DNA has been found in the data bank (Deschavanne *et al.* in preparation). The CpG/GpC ratio in some of these clusters but not in all were in the range expected for CpG islands.

Interestingly, *S. cerevisiae* chromosomes also show compositional oscillations with periodicities of about 50 to 100 kb. (15-17). It has been speculated that they might reflect the distribution of replication origins, chromosomal fibre folding, attachment to the matrix or distribution of structural elements involved in the homology search that precedes synapsis in the early meiotic prophase.

The results of mapping and the computer analysis presented here demonstrate that the composition of human DNA also shows long range compositional oscillations, with the particularity that some of the GC-rich peaks in human genome have a character of CpG islands.

ACKNOWLEDGMENT

This project has been supported by the grant 520634 from GREG to J.F.

REFERENCES

1. Tykocinski N.L., and Max, E.E. (1984) 12, 4385-4396
2. Bird, A., Taggart, M., Frommer, M., Miller, O.J. and Macleod, D. (1985) Cell 40, 91-99
3. Bird, A.P.(1987) TIG 3, 342- 347
4. Goldman, M.A., Holmquist, G.P., Gray, M.C., Caston, L.A. and Nag, A. (1984) Science, 224, 686-692
5 Tazi J. and Bird A. (1990) Cell 60, 909-920
6. Filipski,J., Salinas, J. and Rodier, F. (1987) DNA 6, 109-118
7. Gardiner K. and Patterson D. (1989) Electrophoresis 10, 296-302
8. Koetsier P., Scherr J. and Doerfler W. (1993) Biotechniques.15, 260-262
9. Ichikawa,H., Hosoda,F., Arai,Y., Shimizu,K.,Ohira,M. and Ohki,M. (1993) Nature Genetics 4, 361-365
10. Dessen, P., Fondrat, C., Valencien, C. and Mugnier, C. (1990) Cabios 6, 355-356
11. Ichikawa, H., Shimizu, K., Saito, A. Wang, D, Oliva, R., Kobayashi, H., Kaneko, Y., Miyoshi, H., Smith, C.L., Cantor, C.R. and Ohki, M. (1992) Proc Natl Acad Sci USA 89, 23-27
12. Stinissen, P., Van Hul, W., Van Camp, G., Backhovens, H., Wehnert, A., Vanderberghe, A. and Van Broeckhoven C. (1990) Genomics 7, 119-122
13. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985). Science 228, 953-958
14. Bonetta, L., Kuehn, S.E., Huang, A., Law, D.J., Kalikin, L.M., Koi, M., Reeve, A.E., Brownstein, B.H., Yeger, H., Williams, B.R.G. and Feinberg, A.P. (1990) Science 250, 994-997
15. Dujon, B. *et al.* (1994) Complete DNA sequence of yeast chromosome XI. Nature, 369, 371-378
16. Feldmann, H. *et al.* (1994) The EMBO Journal 13, 5795-5809
17. Sharp P.M. and Lloyd, A.T. (1993) Nucleic Acids Res. 21, 179-183