OVER 80% OF *NotI* SITES ARE ASSOCIATED WITH CpG RICH ISLANDS IN THE SEQUENCED HUMAN DNA

Jun KUSUDA, Makoto HIRATA, Norihiko YOSHIZAKI,
Yosuke KAMEOKA, Ichiro TAKAHASHI, and Katsuyuki HASHIMOTO

*Department of Virology and Rickettsiology, National Institute of Health,
Shinagawa-ku, Tokyo 141*

Summary Clones containing DNA segments linked to *NotI* sites are not only useful for ordering the *NotI* fragments fractionated by pulsed field gel, but also valuable in the search of unknown genes, because they often contain the CpG rich islands and genes related with them. To know the probability of association of *NotI* sites with CpG rich islands, we screened 5,188 sequences accumulated in DNA data base for the presence of *NotI* site and examined the distribution of CpGs around them. The sequential calculation of G+C content and frequency of CpG occurrence at each nucleotide position identified the CpG rich domains close to *NotI* sites in 77 sequences, which corresponds to 84% of total number of candidate sequences. This frequency is consistent well with the prediction that 89% of *NotI* sites in mammalian genome are likely to be present in CpG rich islands and would stress the importance of cloning of *NotI* linking sequences for direct isolation of desired genes. Furthermore, 63 islands newly identified in this study should provide a clue for understanding the transcriptional regulation of associated genes.

Key Words *NotI* linking sequence, CpG rich island, DNA data base, human DNA

INTRODUCTION

In vertebrate genome, dinucleotide CpG occurred at about one-fifth of the expected frequency calculated from G and C content (Losse *et al.*, 1961; Bird, 1980). This rarity was thought to be a consequence of methylation of cytosine residue and deamination; between 60–90% of CpGs in total genome are methylated at the 5' position on cytosine and it has recently become clear that in these modified genome,

Received September 25, 1990; Accepted October 5, 1990.

deficiency of CpG is caused by the failure of DNA repair mechanisms to recognize the deamination of 5'm-cytosine to give thymidine (Coulondre *et al.*, 1978; Wang *et al.*, 1982). Moreover, the distribution of remaining CpGs is spatially non-uniform; approximate 1% of vertebrate genome consists of a fraction that is very rich in CpG and that accounts for about 15% of all CpGs (Cooper *et al.*, 1983; Bird, 1986). Unlike heavily methylated bulk DNA, such CpG rich islands are usually unmethylated domains which contain about 50% of all nonmethylated CpGs. On average, these islands occur every 100 kb in mammalian genomes (Bird *et al.*, 1985; Brown and Bird, 1986) and in many cases appear to coincide with transcribed regions (Lindsay and Bird, 1987; Estivill *et al.*, 1987).

Owing to its high level of CpG methylation and rarity of CpG, vertebrate bulk DNA is poorly cleaved with methylation sensitive restriction endonucleases that contain one or more CpGs in their recognition sequences (CpG enzymes). In contrast, the CpG islands are expected to be abundant with the sites of these enzymes. Assuming that CpG islands in chromosome have average length of 1,000 bp, make up 1% of total DNA, have an average base composition of 65% G+C compared to 40% of bulk DNA and show no deficiency of CpG in the island region, Lindsay and Bird (1987) predicted that 74% of sites for CpG enzymes recognizing 6 bases (such as *SacII*, *BssHII* and *EagI*) and 89% of sites for those recognizing 8 bases (such as *NotI*) should be located in CpG islands. This estimation have led us to the development of strategies that allow specific cloning of potential CpG islands and hence of genes using the sites of CpG enzymes as indicator (Smith *et al.*, 1987; Frischauf, 1989; Kusuda *et al.*, 1989).

Especially, cloning of DNA segments linked to *NotI* sites is considered to a key step of chromosomal walking toward isolation of unknown genes, such as hereditary disease genes, because these clones are valuable in ordering the *NotI* fragments of on average 1,000 kb for long-range physical mapping, and are found to contain frequently the CpG rich islands and many genes neighboring them (Estivill *et al.*, 1987; Pohl *et al.*, 1988).

To know the probability of association of *NotI* sites with CpG rich islands should help us in finding efficiently the CpG rich islands and following genes in *NotI* linking sequences. According to this notion, we searched human DNA sequences compiled in Data base for *NotI* site and examined how frequently CpG islands could be identified around them.

METHODS

DNA sequences were obtained from the Gen-Bank (R-62). The ratio observed/expected (Obs/Exp) CpG was calculated as follows:

$$\text{Obs/Exp CpG} = \frac{\text{Number of CpG}}{\text{Number of C} \times \text{Number of G}} \times N.$$

Where N is the total number of nucleotides in the sequence being analyzed.

A moving average value for % G + C and for Obs/Exp CpG was calculated for each sequence, using a 100 bp window (N-100) moving across the sequence at 1 bp interval.

For this study, we adopted a criteria for CpG rich island established by Gardiner-Garden and Frommer (1987). They have defined the island as stretches of DNA where both moving average of % G + C was greater than 50, and the moving average of Obs/Exp was greater than 0.6.

In order to unambiguously correlate the sequence examined to the already reported results elsewhere, locus names of Gen-Bank and definitions described in *Nucleic Acids Res. Supplement* (Schmidtke and Kooper, 1990) were used, which also enable us to distinguish different genes belonging to a multigene family. To designate sequences filed after issue of this Supplement, we mainly cited key words in the short directory of Gen-Bank. Analysis was carried out using software available on DNASIS (Hitachi Software Engineering Co., Ltd.) and some graphics and statistical programs developed by ourself.

RESULTS AND DISCUSSION

When screening 5,188 human DNA sequences filed in Gen-Bank for *NotI* sites, 170 sequences were picked up to have the site. Among them, there are some families consisting of identical or very similar sequences. In order to avoid the analytical bias, we used the largest sequence of them for further examination. Ninety-six sequences selected in such a way are found to encode genes transcribed by polymerase II with its flanking regions, except of four segments from the ribosomal RNA gene cluster. It is known that a proposed function of CpG rich islands is to regulate the transcription of their associated genes through methylation. As the transcriptional system by polymerase I is uncommon, the sequences related to ribosomal RNA genes were omitted for the analysis. Moreover, resulted entries included several sequences bearing two or more *NotI* sites. In that case, *NotI* site separated by more than 1 kb from adjoining sites was analyzed as independent site, since CpG rich islands are on average 1 kb in length (Bird *et al.*, 1985). Finally, 96 *NotI* sites resided in 92 sequences were satisfied these categories and summarized in Table. The sequences in this list were on average 3.1 kb in length and 37 sequences were originated from cDNAs. It should be noted that the sequences with % G + C over 50 and Obs/Exp over 0.6 are predominant, which leads to the high average of G + C content (59%) and CpG Obs/Exp (0.62) for entire length (see column of total sequence in Table). This suggests that the sequences linked to *NotI* sites are probably derived from the CpG rich islands or regions including them. However their average length are slightly longer than expected length of 1 kb for CpG islands. Therefore, to define the detailed regions rich in CpG of their sequences, we observed the regional deviation of G + C content and CpG occurrence for each sequence. As the functional significance of CpG

rich islands is unclear, a precise definition of the sequences requirements for a CpG rich island is not possible. In our survey of CpG islands, regions with moving average of % G+C over 50 and Obs/Exp CpG over 0.6 have been classed as CpG rich regions. Further, CpG rich regions over 200 bp in length are unlikely occurred by chance alone, so they are labeled as CpG islands (Gardiner-Garden and Frommer, 1987). Figure 1 shows a typical plot of these values along the sequence, including both position of CpG and GpC and sites of CpG enzymes

Fig. 1. Analysis of the distribution of % G+C and Obs/Exp CpG in human cytochrome *c* gene. (a) The distribution of % G+C and Obs/Exp CpG. Moving average of % G+C and Obs/Exp CpG along cytochrome *c* gene was calculated as described in METHOD. The value for % G+C are plotted as thick continuous line and the value for Obs/Exp CpG are plotted as thin continuous line. Regions under the thin continuous line was shadowed to accentuate the change in Obs/Exp CpG. (b) Transcriptional map of cytochrome *c* gene. Exons are represented by boxes. The start site of transcription is marked by a filled triangle and the stop site by open triangle. (c) The distribution of CpG and GpC. The position of CpG and GpC in the sequence is indicated by a vertical line. (d) Restriction map of cytochrome *c* gene showing sites for CpG enzymes.

to facilitate the identification of CpG rich regions. In an example of cytochrome c gene (Suzuki *et al.*, 1989), we found that the regions over 0.6 of Obs/Exp CpG are located both far upstream and just surrounding the transcriptional start site. However, the G+C content of the former region was under 50%, that is critical level to define the CpG rich islands, so there was judged as non-CpG rich island. While the later region with 76% G+C is classed as CpG rich island.

The successive calculation of these ratio at 1 bp interval for each entry identified CpG rich domains around NotI sites in 77 NotI-including sequences. Although several sequences have two NotI sites to be analyzed, there was no sequence with two islands and the CpG rich island was found around either site of them, leaving the case of myc gene in which a CpG rich domain spans both NotI sites. The sequences embracing CpG rich islands amounts to 84% of total number of candidate sequences and this frequency is consistent well with the prediction that 89% of NotI sites in mammalian genome are likely to be present in CpG islands (Lindsay and Bird, 1987). Regions qualified to be CpG rich are on average 0.93 kb in length and represented on average, G+C content of 70 and Obs/Exp CpG of 0.87. These values, much higher than the criteria used to identify them, indicate that the occurrence of CpGs in these regions is nearly close to the frequency expected from G+C content. In contrast, the remaining regions gave an average Obs/Exp CpG of 0.39, even though its average G+C content is still high (52%), suggesting that in the sequences flanking by CpG rich regions, the CpG is suppressed at the level close to those of bulk DNA and a high G+C content is not sufficient to prevent the CpG depletion. Each CpG rich island was also analyzed in terms of the location relative to the transcriptional unit of the associated genes. As shown in Table, CpG islands were found to associate with the 5' ends of all house keeping genes and many tissue-specific genes, and with the 3' ends of some tissue-specific genes.

These results support strongly that the NotI sites in human genome are surely concentrated in CpG rich islands neighboring transcribable regions and this would stress the importance of isolation of NotI linking clones for the search of unfound genes. Furthermore, 63 CpG islands newly identified in this study should provide a clue for understanding the transcriptional regulation of each gene and such a considerable number of sequences might aid to elucidate the localization of CpG islands along chromosomal banding (Gardiner *et al.*, 1990).

REFERENCES

- Anson, D.S., Blake, D.J., Winship, P.R., Birnbaum, D. and Brownlee, G.G. 1988. Nullisomic deletion of the mcf. 2 transforming gene in two haemophilia B patients. *EMBO J.*, **7**: 2795-2799.
- Bird, A.P. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**: 1499-1504.
- Bird, A.P., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. 1985. A fraction of the

- mouse genome that is derived from islands of nonmethylated, CpG rich-DNA. *Cell* **40**: 91-99.
- Bird, A.P. 1986. CpG-rich islands and the function of DNA methylation. *Nature* **231**: 209-213.
- Bird, A.P. 1989. Two classes of observed frequency for rare-cutter sites in CpG islands. *Nucleic Acids Res.* **17**: 9485.
- Brown, W.R.A. and Bird, A.P. 1986. Long-range restriction site mapping of mammalian genomic DNA. *Nature* **322**: 477-481.
- Cooper, D.N., Taggart, M.H. and Bird, A.P. 1983. Unmethylated domains in vertebrate DNA. *Nucleic Acids Res.* **11**: 647-658.
- Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**: 775-780.
- Estivill, X., Farrall, M., Scambler, P.J., Bell, G.M., Hawley, K.M.F., Lench, N.J., Bates, G.P., Kruyer, H.C., Frederick, P.A., Stanier, P., Watson, E.K., Williamson, R. and Wainwright, B.J. 1987. A candidate for the cystic fibrosis locus isolated by selection for methylation-free islands. *Nature* **326**: 840-845.
- Frischauf, A.M. 1989. Construction and use of linking libraries. *TECHNIQUE* **1**: 3-10.
- Gardiner-Garden, M. and Frommer, M. 1987. CpG island in vertebrate genomes. *J. Mol. Biol.* **196**: 261-282.
- Gardiner, K., Horisberger, M., Kraus, J., Tantravahi, U., Korenberg, J., Rao, V., Reddy, S. and Patterson, D. 1990. Analysis of human chromosome 21: correlation of physical and cytogenetic maps; gene and CpG island distributions. *EMBO J.* **9**: 25-34.
- Josse, J., Kaiser, A.D. and Kornberg, A. 1961. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *J. Biol. Chem.* **236**: 864-875.
- Kusuda, J., Kameoka, Y., Takahashi, I., Fujiwara, H. and Hashimoto, K. 1989. A simple method for screening of *NotI* linking clones. *Nucleic Acids Res.* **17**: 8890.
- Lindsay, S. and Bird, A.P. 1987. Use of restriction enzymes to detect potential gene sequences in mammalian DNA. *Nature* **327**: 336-338.
- Pohl, T.M., Zimmer, M., MacDonald, M.E., Bucan, M., Poustka, A., Volinia, S., Searle, S., Zehetner, G., Wasmuth, J.J., Gusella, J., Lehrach, H. and Frischauf, A.-M. 1988. Construction of *NotI* linking library and isolation of new markers close to the Huntington's disease gene. *Nucleic Acids Res.* **16**: 9185-9198.
- Schmidtke, J. and Cooper, D.N. 1990. A comprehensive list of cloned human DNA sequences. *Nucleic Acids Res.* **18** Supplement: 2413-2547.
- Smith, C.L., Lawrence, S.K., Gillespie, G.A., Cantor, C.R., Weissman, S.M. and Collins, F.C. 1987. Strategies for mapping and cloning macroregions of mammalian genomes. *Methods Enzymol.* **151**: 461-489.
- Suzuki, H., Hosokawa, Y., Nishikimi, M. and Ozawa, T. 1989. Structural organization of the human mitochondrial cytochrome *c*₁ gene. *J. Biol. Chem.* **262**: 1368-1374.
- Wang, R.Y.-H., Kuo, K.C., Gehrke, C.W., Huang, L.-H. and Ehrlich, M. 1982. Heat and alkali-induced deamination of 5-methylcytosine and cytosine residues in DNA. *Biochim. Biophys. Acta* **697**: 371-377.