Preferential isolation of DNA fragments associated with CpG islands

(Masahiko Shiraishi*†, Leonard S. Lerman‡, and Takao Sekiya*)

*Oncogene Division, National Cancer Center Research Institute, 1-1, Tsukiji 5-chome, Chuo-ku 104, Tokyo, Japan; and ‡Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139

Contributed by Leonard S. Lerman, January 17, 1995

ABSTRACT We describe a procedure for preferential isolation of DNA fragments with G+C-rich portions. Such fragments occur in known genes within or adjacent to CpG islands. Since about 56% of human genes are associated with CpG islands, isolation of these fragments permits detection and probing of many genes within much larger segments of DNA, such as cosmids or yeast artificial chromosomes, which have not been sequenced. Cloned DNA fragments digested with four restriction endonucleases were subjected to denaturing gradient gel electrophoresis. Long G+C-rich sections in fragments inhibit strand dissociation after the fragments reach retardation level in the gradient; such fragments are retained in the gel after most others disappear. Nucleotide sequences of the retained fragments show that about half of these fragments appear to be derived from CpG islands. Northern analysis indicated the presence of RNA complementary to most of the retained fragments. A heuristic approach to the relation between base sequence and the kinetics of strand dissociation of partly melted molecules appears to account for retention and nonretention. The expectation that CpG island fragments will be enriched among fragments retained in a denaturing gradient is supported by rate estimates based on melting theory applied to known sequences. This method, designated SPM for segregation of partly melted molecules, is expected to provide a means for convenient and efficient isolation of genes from unsequenced DNA.

Detection of anonymous genes within a large region of unsequenced DNA can be difficult. After chromosomal loci for human genes of unknown sequence are determined by low resolution methods, such as linkage analysis and deletion mapping, it is appropriate to examine yeast artificial chromosome, P1, or cosmid clones encompassing the region of interest for candidate genes. There are three major approaches toward recognition of genes: exon trapping (1, 2), cDNA selection (3, 4), and the identification of CpG islands. A CpG island is a relatively short stretch of a G+C-rich region (up to about 2 kb) in which the frequency of nonmethylated CpG dinucleotides is substantially higher than elsewhere in genomic DNA (5). Almost all housekeeping genes and many genes expressed in specific tissues (tissue-specific genes) have a CpG island at the 5′ end (6, 7). Since CpG islands are closely associated with genes, their identification is a useful step toward the isolation of genes (8).

CpG islands are usually identified as sequences containing clusters of recognition sites for certain restriction endonucleases; the clusters themselves are spaced at long distances, typically tens of kilobases. This pattern of site distribution was first recognized with Hpa II (CCGG) (5), which produced Hpa II tiny fragments from the sequences within islands, hence the earlier terminology, HTF islands. Rare cutters, such as BstHII (GCGCGC), Eag I (CGGCCG), and Sac II (CCGCCG), are particularly useful for island identification (8, 9). Detection of a single rare-cutter site is often sufficient to identify a gene sequence (10), and the presence of a BsrUI site (CGCG) is suggestive (11).

Denaturing gradient gel electrophoresis (DGGE) provides a means for separating DNA molecules on the basis of local variation in base composition within the DNA fragments. It is effective where part of the molecule is relatively dense in G+C pairs. The separation depends on the pronounced drop in electrophoretic mobility in polyacrylamide gels that occurs when part of a DNA molecule melts, forming a structure that is partly helical and partly random chain (12). The low residual mobility restricts migration into more strongly denaturing regions of the gradient and results in a relatively stable band pattern. The same conditions that effect retardation also tend to increase the rate of complete dissociation of the strands. However, the level of denaturant at which a molecule is retarded depends on the least stable portion of the molecule, while the rate of strand dissociation is controlled largely by the stability of the remaining helix. On prolonged exposure to the electric field, the band of retarded molecules will fade away through dissociation, unless the stability of the unmelted portion is much higher than that of the melted part.

The low-stability portion affects retardation at a denaturant concentration in the gradient determined by the sequence of that portion alone. Although equilibrium continues to favor helicity over strand separation in the remaining G+C-rich portion of the molecule, continuous electrophoretic transport of the small amount of dissociated strands away from the band will result in its slow disappearance. It will be convenient to define thermally stable regions (TSRs) as sections of sequence of sufficient length and G+C density to substantially restrain dissociation at an equivalent temperature below 80°C (equivalent temperature includes the effect of denaturants). Molecules carrying such regions and retarded at lower equivalent temperature will persist in the gel after others disappear. Since partial melting of molecules of this length is fully and rapidly reversible, retained fragments can be recovered from bands as fully helical molecules.

MATERIALS AND METHODS

Plasmid and Cosmid Clones. The plasmid clones used in this study are summarized in Table 1. Restriction endonucleases were applied sequentially following the manufacturer’s recommendation at each step. The human placenta genomic DNA library in a cosmid vector was obtained commercially (Stratagene).

Abbreviations: DGGE, denaturing gradient gel electrophoresis; SPM, segregation of partly melted molecules; TSR, thermally stable region; RHST, relative helix survival time.

†To whom reprint requests should be addressed.
‡The sequences reported in this paper have been deposited in the GenBank data base (accession nos. D31740–D31759).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.
DGGE. We followed published procedures (13, 14) with slight modifications. Briefly, 0.5 μg of digests of the inserted fragment of a plasmid clone or 1.0 μg of a total cosmid clone DNA digested with restriction enzymes was loaded on an 8%polyacrylamide gel containing a linear gradient of chemical denaturant in a bath maintained at 59.5°C. The denaturant concentration was 9% at the top and 90% at the bottom; 100% denaturant is 7 M urea/40% formamide in 1x TAE buffer (40 mM Tris/20 mM NaOAc/1 mM EDTA, pH 7.4). An electric field of 10 V/cm was applied for the time indicated; in the patterns shown in Figs. 1 and 3, the electric field had been applied 3 h before the first sample was added.

Cloning of Retained Fragments. Retained fragments were eluted out of the gel following published procedures (15). Restriction enzyme digests of pGEM-5Zf(+) (Promega; Sp I, Nde I, and Sp I/Nde I), pGEM-7Zf(+) (Promega; Eco RI and Eco RI/Sp I), and pUC18 (Taq KaRa; Eco RI/Nde I) were prepared to accommodate any restriction fragments having Mse I, Nla III, Bfa I, or Tsp509I ends. After dephosphorylation, these plasmids were ligated as a mixture together with recovered fragments.

Theoretical Calculations. Calculations on the distribution of thermal stability in DNA molecules, the progression of melting, and its relation to DGGE were carried out using programs based on the Poland–Fixman–Freire algorithms, similar to MELT, as described (16).

Northern and Southern Analyses. Human multiple tissue Northern blots (Clontech) were hybridized with 32P-labeled probes. Hybridization and washing were carried out by following the supplier’s recommendation. Southern analysis was carried out similarly as described (17).

Nucleotide Sequence Analysis. About 3 μg of plasmid DNA was used as a template for DNA sequencing analysis (AutoRead Sequencing Kit, Pharmacia). The reaction mixture was analyzed by an automated laser fluorescent DNA sequencer (Pharmacia).

RESULTS

Analysis of Known Genes. To study the possibility that fragments associated with CpG islands can be isolated preferentially by DGGE, we examined the properties of DNA fragments that include the entire length of each of several cloned genes known to carry CpG islands.

DNA fragments were digested with restriction endonucleases selected to preserve as much as possible the integrity of the CpG islands, while cutting frequently elsewhere in the gene. Preliminary experiments revealed that sequential digestion with four enzymes, Mse I (TtaAa), Tsp509I (AaT), Nla III (CAtG), and Bfa I (CAtA), yielded DNA fragments of appropriate length for analysis by DGGE. Since digestion with fewer enzymes gave additional retained fragments that could not be identified with CpG islands, all subsequent work was carried out after treatment with these four enzymes. Recognition sites of restriction endonucleases such as Sau3A1 (GATC) or Rsa I (GTAC) appear relatively frequently in some CpG islands, and digestion yielded island fragments too short to be retained selectively.

Each digest was electrophoresed through a denaturing gradient gel with a broad set of run times. With ECGF fragments, only one band with reduced mobility was retained after 12 h (Fig. 1A). This band was found to contain two fragments. Fragment I-1 (862 bp) was found to contain exon 1 and flanking sequence on both sides making up a large part of the CpG island, as indicated by rare-cutter sites (Fig. 2). The other fragment, fragment I-2 (990 bp), contained exons 8–10 (Fig. 2). The fragmentation pattern with rare-cutters suggests that fragment I-2, from the 3’ region, represents a second CpG island sequence.

Restriction enzyme digests of DNA fragments encompassing FOS, HRAS, MYC, MYCN, and 28S ribosomal RNA genes were similarly analyzed (Fig. 1B–F). Single fragments were recovered from FOS, HRAS, and MYC, which were derived from regions predicted to be CpG islands by rare-cutter sites. Two MYCN fragments were retained, one of which was derived from a predicted CpG island (fragment 6). In the case of the 28S ribosomal RNA gene, at least five fragments were retained, reflecting a high overall G+C-density.

Fragments that were retained through an initial 12-h run were retained after an additional 12-h run with the exception of fragment 4 from MYC.

Analysis of Plasmid Clones Containing Anonymous DNA Fragments. Six large fragments from plasmid clones containing anonymous DNA fragments were subjected to the same four enzymes and DGGE procedure as above. All restriction fragments of the six digests except one disappeared from the gel. One band from plasmid clone pHHH208 was retained after 12 h (Fig. 3A, fragment 12, 863 bp). Its nucleotide sequence indicated the presence of a G+C-rich region. A Northern blot showed the presence of a 1.9-kb RNA complementary to at least a part of fragment 12 in human peripheral blood leukocyte and spleen (Fig. 3B), suggesting that fragment 12 is associated with a gene expressed in hematopoietic tissue.

Analysis of Cosmid Clones. Out of 31 cosmid clones randomly selected from a human genomic library, 29 clones gave at least one retained band (Fig. 4), with an average of about three retained fragments per cosmid clone. Nineteen fragments from ten cosmid clones were subcloned. Southern analysis, probing with total human male DNA, revealed that 11 of the 19 subclones were derived from single-copy sequences.

![Fig. 1. Parallel DGGE of complete restriction enzyme digests of plasmid fragments containing the full length of the following genes: ECGF (A), FOS (B), HRAS (C), MYC (D), MYCN (E), and 28S ribosomal RNA genes (F). The digested fragments were electrophoresed for 6 h, lane 1; 7.5 h, lane 2; 9 h, lane 3; 10.5 h, lane 4; or 12 h, lane 5 adjusted by delayed loading. Recovered fragments are indicated by arrows and numbers.](image-url)
DNA sequencing showed that nine of the single-copy fragments carried G+C-rich regions (TSR in Table 2), and one TSR-carrying fragment was derived from a highly repetitive sequence. All single-copy fragments hybridized with human DNA (data not shown).

Northern hybridization was carried out in a search of transcribed sequences to test the hypothesis that retained fragments are associated with expressed genes (Fig. 5). A 3.6-kb RNA was detected in all tissues examined when fragment R27-2 was used as the probe. This fragment is considered to be a part of a housekeeping gene. In contrast, a 7.2-kb RNA was detected only in skeletal muscle and heart when fragment R1-2 was used as the probe. Probing with fragment R20-1 detected a 1.8-kb RNA in colon and 2.0-, 3.6-, and 5.6-kb RNAs in prostate. These fragments are considered to be a part of tissue-specific genes. Of the nine single-copy G+C-rich fragments, six contained transcribed sequences (Table 2).

Three fragments (R1-3, R23-1, and R35-2) containing G+C-rich regions failed to detect complementary RNA, although homologous DNA fragments were detected in monkey, rat, mouse, hamster, and dog by Southern analysis (data not shown). This suggests that the fragments represent conserved but unexpressed genes or that exons are not present in these fragments.

**DISCUSSION**

We describe a means to collect restriction fragments that contain at least part of a CpG island from large, uncharacter-
ized DNA molecules. The system depends on both extensive fragmentation at restriction sites that are infrequent in CpG islands and expected relative durability during DGGE of surviving fragments in which part of the sequence is locally dense in G+C. Other fragments are expected to be dissipated from the sharp, retarded bands characteristic of DGGE through the more or less rapid dissociation of their strands.

Application of the procedure to five genes coding for known proteins resulted in 16 retarded bands in denaturing gradient gels (Table 3). Of these, seven were retained after a 12-h electric field exposure, and only one of the seven failed to contain a restriction site characteristic of CpG islands.

To examine the generality of these observations and to predict application to other sequences, we have attempted to relate the observed properties to formal melting theory. Retention of a molecule in a band at its retardation level will depend on the rate of strand dissociation at that equivalent temperature. Lacking a rigorous theory concerning the dissociation rate at temperatures lower than those sufficient to melt the entire molecule, we have followed a heuristic assumption that a quantity resembling an activation pertains to the rate of separation and unzipping of each remaining helical base pair. That quantity is taken to be proportional to the difference between the temperature of the retardation level, \( t_{\text{ret}} \), and the \( t_m(i) \) of each residually helical pair, \( i \), as given by the melting map (Fig. 6). The RHST required for separation will be the individual time values, as defined by each activation energy summed over all remaining helical pairs. This sum can be expected to be small for decreasing fragments and large for retained fragments. RHST is then

\[
\text{RHST} = \sum \left[ \exp\left( t_m(i) - t_{\text{ret}} / q \right) - 1 \right]
\]  

![Fig. 5. Representative results of Northern blot analysis of human tissues probed with some of the recovered fragments.](image)
where \( q \) is an arbitrary constant that is uniform for all fragments; and the summation extends only over pairs calculated by MELT to be helical at the retardation level. For simplicity in the present context, we have allowed MELT to define the retardation level, \( T_{ret} \), at the temperature required to melt about 121 bp, reducing mobility in the gel to about 20% of its initial value (16).

Application of Eq. 1 to fragments of sequenced genes is shown in Table 3, for which a \( q \) value of 3.185 was used. All retained fragments show RHST values of 0.93 \( \times 10^3 \) or greater, while dissipated ones have values of 0.80 \( \times 10^3 \) or less. The mean for retained fragments is 2.81 \( \times 10^3 \); for nonretained fragments, the mean is 0.37 \( \times 10^3 \). Discrimination is not determined simply by the length-of the fragment.

The results of applying Eq. 1 to the sequences of anonymous retarded fragments derived from cosmids are shown in Table 2, where it can be seen that RHST values of only 3 of 19 lie below 0.80 \( \times 10^3 \). The exceptions are unusual in the distribution of thermal stability in the molecule in having a very short, lowest melting domain. Other formulations for a retention criterion based on sequence were less successful.

Within CpG islands there are sequences of nearly uniformly high stability that are at the edge of dissociation as partial melting begins (for example, E1 in ECGF). They will not be retained and their RHST value will be low. However, calculation indicates that a few fairly stable fragments may be found away from the regions of generally high G+C-density.

It is convenient to define a TSR as a segment with a calculated \( T_{mel} \) greater than 80°C extended over at least 10% of the length of molecules longer than about 300 bp. Calculation of the melting maps of a total of 22 genes of known sequence showed that the TSRs were found predominantly in the predicted CpG islands and rarely in bulk DNA, suggesting that retarded fragments carrying a TSR are usually associated with CpG islands. In Table 2, we note that most fragments containing BstUI sites also carry a TSR. It is clear from Table 2 that the TSR is not a predictor of gradient retention.

A method depending on binding to a specific nuclear protein (11) has been reported for selective isolation of CpG islands from genomic DNA. However, in cloned DNA fragments, where information of methylation status in the original genome is lost, the present method, segregation of partly melted molecules (SPM), permits preferential isolation of putative CpG islands. Detection of transcribed sequences or sequence conservation across species shows that all recovered single-copy fragments with a TSR lie within or overlap gene sequences.

Assuming that the average length of the insert in a cosmid clone is about 40 kb and that each fragment corresponds to a different CpG island, the data suggest a mean of one CpG island per 40-kb DNA. According to a recent estimate based on the methylation status of Hpa II sites (18), it was suggested that there are about 45,000 CpG islands per haploid genome in humans, or an average of one CpG island per 65 kb. Although the two calculations are derived from different measurements, they give roughly similar values.

SPM should be useful in locating genes for the following reasons: (i) SPM is easily carried out, and the procedures can be finished in a few days. (ii) SPM is independent of rare-cutter sites, and the laborious steps in mapping rare-cutter sites are obviated. (iii) DNA fragments from CpG islands are isolated en bloc, not as rare-cutter sites, and most of the recovered single-copy fragments are effective as probes in Northern hybridization. (iv) SPM is applicable both to cloned or amplified DNA fragments, as well as to genomic DNA, since it does not depend on methylation. (v) It should be possible to recognize a large number of island fragments within one lane of a gel, because retardation levels are determined by the sequences with the lowest melting temperature.

In summary, about half of a large set of anonymous DNA fragments retained in a denaturing gradient meet most criteria for inferring localization of CpG islands: the presence of high G+C-density, correspondence with a region of rare-cutter sites, the presence of BstUI sites, proximity to transcribed sequences, and cross-species conservation. Theoretical analysis of the properties of DNA relevant to denaturing gradient separations and stability extends its applicability to CpG island detection.

We thank K. Hagiwara, I. M. Verma, Y. Taya, and F. Alt for providing plasmid clones used in this study. Several plasmid clones were obtained from the American Type Culture Collection (Rockville) or the Japanese Cancer Research Resources Bank (Tokyo). This work was supported in part by a Grant-in-Aid from the Ministry of Health and Welfare for a Comprehensive 10-Year Strategy for Cancer Control, Japan (to T.S.), and by a grant from the National Institutes of Health (HG00345 to L.S.L.).