Identification and Characterization of the Potential Promoter Regions of 1031 Kinds of Human Genes

Yutaka Suzuki,[1,2,3,9] Tatsuhiko Tsunoda,[2,3] Jun Sese,[4] Hirotoshi Taira,[5] Junko Mizushima-Sugano,[1,2] Hiroko Hata,[1] Toshio Ota,[6] Takao Isogai,[6] Toshihiro Tanaka,[2] Yusuke Nakamura,[2] Akira Suyama,[7] Yoshiyuki Sakaki,[2,3] Shinichi Morishita,[4] Kousaku Okubo,[8] and Sumio Sugano[1,2]

[1]Department of Virology and [2]Human Genome Center, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo 108–8639, Japan; [3]Genome Science Center, Institute of Physical and Chemical Research (RIKEN); Wakoshi, Saitama 351–0106, Japan; [4]Department of Complexity Science and Engineering Graduate School of Frontier Science, University of Tokyo, Bunkyo-ku, Tokyo 113–0033, Japan; [5]Intelligent Communication Laboratory, Nippon Telegraph and Telephone Communication Science Laboratories, Seika-cho, Soraku-gun, Kyoto 619–0237, Japan; [6]Helix Research Institute, Kisarazushi, Chiba 292–0812, Japan; [7]Department of Life Sciences, University of Tokyo, Meguro-ku, Tokyo 153–0041, Japan; [8]The Institute of Molecular and Cell Biology, Osaka University, Suita-shi, Osaka 565–0871, Japan

To understand the mechanism of transcriptional regulation, it is essential to identify and characterize the promoter, which is located proximal to the mRNA start site. To identify the promoters from the large volumes of genomic sequences, we used mRNA start sites determined by a large-scale sequencing of the cDNA libraries constructed by the "oligo-capping" method. We aligned the mRNA start sites with the genomic sequences and retrieved adjacent sequences as potential promoter regions (PPRs) for 1031 genes. The PPR sequences were searched to determine the frequencies of major promoter elements. Among 1031 PPRs, 329 (32%) contained TATA boxes, 872 (85%) contained initiators, 999 (97%) contained GC box, and 663 (64%) contained CAAT box. Furthermore, 493 (48%) PPRs were located in CpG islands. This frequency of CpG islands was reduced in $TATA^+/Inr^+$ PPRs and in the PPRs of ubiquitously expressed genes. In the PPRs of the *CGM2* gene, the *DRA* gene, and the *TM30pl* genes, which showed highly colon specific expression patterns, the consensus sequences of E boxes were commonly observed. The PPRs were also useful for exploring promoter SNPs.

[The nucleotide sequences described in this paper have been deposited in the DDBJ, EMBL, and GenBank data libraries under accession nos. AU098358–AU100608.]

To understand the mechanism of transcriptional regulation, it is indispensable to identify and characterize the promoter. The promoter is usually located just proximal to or overlapping the transcription initiation site and contains several sequence motifs with which transcription factors (TFs) interact in a sequence-specific manner. When recruited, these TFs serve as molecular switches, which turn the transcription of the gene on or off. The combinations of the TF-binding motifs in promoters vary depending on the gene, so that an appropriate subset of genes can be expressed according to tissue types or developmental stages (Mitchell and Tjian 1989; Novina and Roy 1996). Among many TF-binding motifs, TATA box and initiator (Inr) are considered to be especially important because only these motifs are directly recognized by the general transcription factors (Roeder 1996; Smale 1997). GC box and CAAT box are also thought to be important promoter elements besides TATA box and Inr.

Whether the promoter is located in CpG islands or not is also important for transcriptional regulation. CpG islands are defined as dispersed regions of DNA with a high frequency of CpG dinucleotide relative to the bulk genome (Gardiner-Garden and Frommer 1987; Larsen et al. 1992). When CpG islands remain unmethylated, TF-binding sites can be recognized by TF. In contrast, when methylated, the presence of 5-methylcytosine in CpG islands interferes with the binding of TFs and thus suppresses transcription (Cross and Bird 1995; Costello et al 2000).

Despite the important roles of the promoters, the number of genes whose promoters have been identified is limited. In the Eukaryotic Promoter Database (EPD; Rel. 62; http://www.epd.isb-sib.ch; Perier et al. 2000), which accumulates previously-characterized promoter sequences, only 273 human promoters have

[9]Corresponding author.
E-MAIL ysuzuki@ims.u-tokyo.ac.jp; FAX 81 3 5449 5416.
Article published on-line before print: *Genome Res.,* 10.1101/gr.164001. Article and publication are at www.genome.org/cgi/doi/10.1101/gr.164001.

been registered. This may be due to the fact that the exact mRNA start sites have not been identified for most of the genes. The conventional methods used to identify the mRNA start site, such as S1 mapping, primer extension, or 5′ RACE (Berk and Sharp 1977; McKnight and Kingsbury 1982; Schaefer 1995) are technically difficult and often lead to the inaccurate identification of the mRNA start sites.

Previously, we developed a novel method to construct a full-length enriched and 5′-end enriched cDNA library (Maruyama and Sugano 1994; Suzuki et al. 1997). This "oligo-capping" method uses the cap structure of mRNA, which is the characteristic structure of the 5′ end of eukaryotic mRNAs. By three sequential enzyme reactions, the oligo-capping method replaces the cap structure of mRNA with synthetic oligoribonucleotide (Fig. 1). Using this 5′ oligoribonucleotide as a sequence tag, cDNAs that originally contained the cap structure are selectively cloned. This type of library (oligo-capped cDNA libraries) contained 50%–80% of the full-length cDNAs whose 5′ ends correspond to the mRNA start sites (Suzuki et al. 1997, 2000).

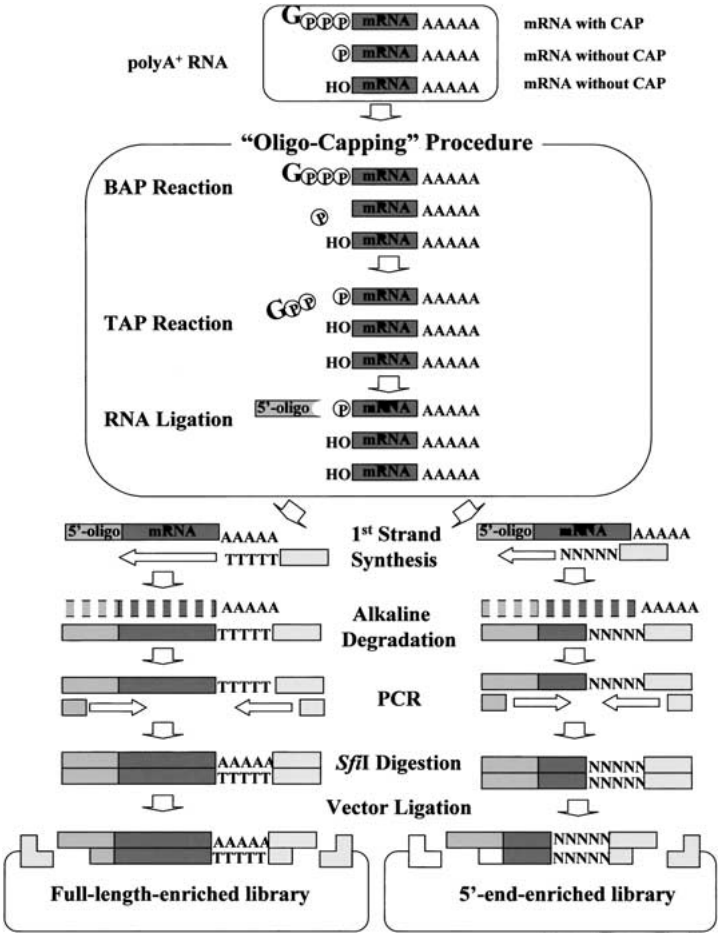

Figure 1 Schematic representation of the construction of oligo-capped cDNA libraries. The cap structure of the mRNA was replaced with the 5′ oligonucleotide by the oligo-capping method, which consists of three enzymatic reaction steps. Bacterial alkaline phosphatase (BAP) hydrolyzes the phosphate of the 5′ ends of truncated mRNAs whose cap structures have been degraded. Tobacco acid pyrophosphatase (TAP) removes the cap structure, leaving a phosphate at the 5′ end. T4 RNA ligase, which requires a phosphate at the 5′ end as its substrate, selectively ligates the 5′ oligonucleotide to the 5′ end that originally had the cap structure. Using oligo-capped mRNA, first-strand cDNA was synthesized with dT adapter primer. After alkaline degradation of the RNA, first-strand cDNA was amplified by PCR, digested with restriction enzyme *Sfi*I, and cloned into a plasmid vector. For further details of the procedure, see references (Suzuki et al. 1997, 2000). RNA and DNA molecules are represented by dark gray bars, the 5′ oligonucleotide by light gray boxes, and PCR primers by broken bars. (Gppp) Cap structure, (p) phosphate, (OH) hydroxyl.

The oligo-capped cDNA libraries are found to be good resources for identification of the mRNA start site for many genes. We have constructed oligo-capped cDNA libraries from 34 kinds of human tissues and cultured cells and sequenced the 5&prime ends of 100,000 clones from these cDNA libraries. By clustering the sequence data, we identified the mRNA start sites at least for 2251 genes. We aligned these transcriptional start sites onto the genomic sequences and retrieved adjacent sequences as the potential promoter regions (PPRs) for 1031 genes. Here we report the identification and characterization of our first 1031 PPRs.

RESULTS AND DISCUSSION

Identification of the PPRs of 1031 Kinds of Genes

We searched for the promoters using the mRNA start sites of 2251 kinds of genes identified by the oligo-capped method (Suzuki et al. 2000). The genomic sequences in Genbank were searched by BLASTN (Altschul et al. 1990) and aligned by CLUSTALW (Thompson et al. 1999). The genomic sequences were obtained from GenBank on February 8, 2000, when draft and finished sequences altogether had covered ~60% of the entire human genome. Repetitive sequences, such as *Alu*, were masked with CENSOR (Jurka et al. 1996). As a result of the alignment, the mRNA start sites of 1031 genes were mapped onto the genomic sequences. For each gene, the genomic sequence between 500 bp upstream and 100 bp downstream of the mRNAs start site was retrieved as PPR.

To check the validity of the retrieved PPRs, we searched EPD with the corresponding gene names of these PPRs. Among 1031 genes, 44 genes had their promoters registered in EPD. Forty promoters, such as the promoter of β *actin, serum albumin,* and *ferritin heavy chain,* completely coincided with the registered promoters. Only four PPRs differed from registered promoters. These four PPRs seems to be derived from

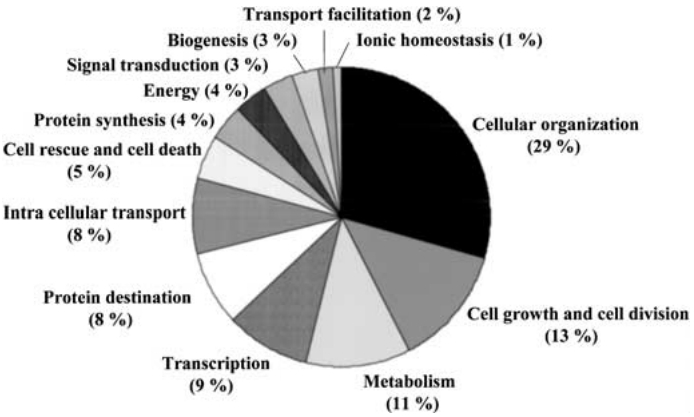

Figure 2 Functional classification of the potential promoter regions (PPRs). The promoters were classified according to which functional category the corresponding genes had been assigned in Human Info Base (HIB; http://www.mips.biochem.mpg.de/proj/human/selec_view.html). The functional categories and population of the genes belonging to each category are shown.

sequences adjacent to pseudogenes. Pseudogenes sometimes gives the highest score in BLAST search because they tend to lack introns. Considering these results, it is likely that a high proportion of the PPRs identified by our approach actually coincide with the real promoters.

Functional Classification of the Corresponding Genes

To show that PPRs of genes involved in various biological functions were included in the retrieved PPRs, we categorized the corresponding genes of the PPR, according to their functional annotations. We used Human Info Base (HIB) at MIPS (http://www.mips.biochem.mpg.de/proj/human/selec_view.html), in which human genes are functionally classified based on their homologies to the yeast genes. Among 1031 corresponding genes, 426 genes appeared in HIB. As shown in Figure 2, they are distributed along all the categories in HIB. The most frequently observed genes were those involved in cellular organization, such as α *actin* and *keratin 6*. Although the numbers were small, the PPRs of genes involved in signal transduction and cell death such as *PDGF receptor* and *TRAIL* (TNF-related apoptosis-inducing ligand) were also included in our data set. This suggests that there is no strong bias in the PPRs with regard to the functions of the corresponding genes.

Analysis of TF-Binding Sites and CpG Islands in the PPRs

To characterize the sequence elements in the PPRs, the PPR sequences were searched for possible TF-binding sites and for CpG islands. For the search of TF-binding sites, a TF-binding prediction program, TFBIND (Tsunoda and Takagi 1999) was used. CpG islands were defined as continuous regions >200 bp with GC content [%(G + C)] >50% and CpG ratio >0.6 (for more details, see Methods).

First, we analyzed important TF-binding sites as well as CpG islands. Table 1 shows that 329 PPRs (32%) out of 1031 PPRs contained TATA boxes, 872 (85%) contained Inrs, 999 (97%) contained GC boxes, and 663 (64%) contained CAAT boxes. For 493 (48%), the PPRs were located in CpG islands. As for the TATA box and the initiator, we also examined what fraction of

Table 1. Predicted TF Binding Sites and CpG Islands in the 1031 PPRs

TF definition	*Matrix ID	Hit No. (%)	Preferred region (searched region)	Cutoff value	**Consensus sequence
TATA box	V$TATA_01	329 (32%)	−40 ~ −23 (−90 ~ +27)	0.77	STATAAAWRNNNNNN
Initiator	V$CAP_01	872 (85%)	−5 ~ +6 (−55 ~ +56)	0.87	NCANNNNN
GC box	V$GC_01	999 (97%)	−74 ~ −45 (−124 ~ +5)	0.78	NRGGGGCGGGGCNK
CAAT box	V$CAAT_01	663 (64%)	−105 ~ −70 (−155 ~ −20)	0.78	NNNRRCCAATSA

	Hit No. (%)	Length (bp)	CpG ratio	GC content (%)
CpG island	493 (48%)	>200	0.6	50

The search for TF binding sites was performed using the preferred region of each TF binding motif. For example, because the preferred region of the TATA box is −40 to −23, the region of −90 to +27 was searched. Fifty-base margins were added at both ends of the preferred region because in some cases multiple mRNA start sites were observed.
*A TRANSFAC notation, which starts with an identifier that indicates vertebrates (V$), followed by an acronym for the factor (for more details, see http://transfac.gbf.de/TRANSFAC/doc/site3.html).
**The symbols used in addition to A, C, G, and T are: W = A or T; S = C or G; R = A or G; T = C or T; K = G or T; M = A or C; B = C, G, or T; D = A, G, or T; H = A, C, or T; V = A, C, or G; N = A, C, G, or T.

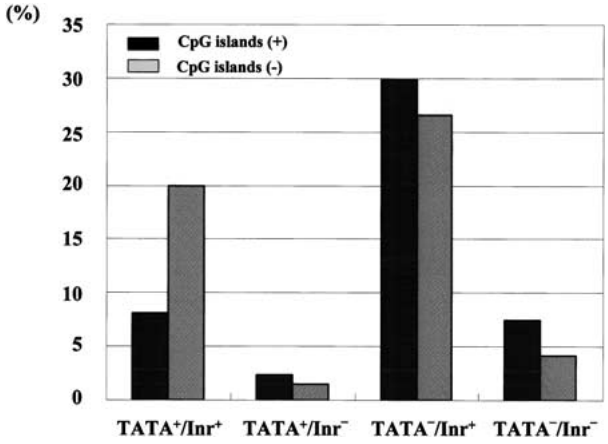

Figure 3 Population of the TATA$^+$Inr$^+$, TATA$^+$Inr$^-$, TATA$^-$Inr$^+$, and TATA$^-$Inr$^-$ potential promoter regions (PPRs) located in/outside of CpG islands. Solid bars represent population of PPRs located in CpG islands; shaded bars represent those outside of CpG islands in each category.

genes contained both, either, or neither of them. The TATA$^+$Inr$^+$, TATA$^+$Inr$^-$, TATA$^-$Inr$^+$, and TATA$^-$Inr$^-$ genes constituted 28%, 4%, 56%, and 12% of the genes, respectively.

Although these promoter elements are considered to be important, there have been almost no reports that describe what fraction of promoters contains these elements. In this respect, it is interesting to note that frequency of GC box is almost 100% whereas that of TATA box is 32%. This might suggest that GC box plays more fundamental roles in transcription than TATA box.

Next, we examined the relation between TATA boxes, Inrs, and CpG islands. We calculated the frequencies of PPRs with (+) or without (−) CpG islands separately for TATA$^+$Inr$^+$, TATA$^+$Inr$^-$, TATA$^-$Inr$^+$, and TATA$^-$Inr$^-$ PPRs (Fig. 3). For TATA$^+$Inr$^-$, TATA$^-$Inr$^+$, and TATA$^-$Inr$^-$ PPRs, the numbers of PPRs in CpG island were similar to that out of 1031 PPRs. In contrast, two-thirds of TATA$^+$Inr$^+$ PPRs were located outside of CpG islands. It is known that TATA box and Inr are the most preferred docking platforms created for the RNA polymerase II complex, which drives the most active transcription (Roeder 1996; Smale 1997). Our results may suggest that TATA$^+$Inr$^+$ promoters need not be located in CpG islands because of their strong activity.

Association Between the PPRs with the Expression Profiles

To study the relationship between the promoter element in the PPRs and the transcriptional level of genes, we associated the PPRs with the expression profiles obtained by iAFLP. iAFLP is an RT-PCR-mediated gene quantification method (Kato 1997; Kawamoto et al. 1999) and now expression levels among 30 kinds of human tissues are available for many human genes (http://bodymap.ims.u-tokyo.ac.jp). We searched the

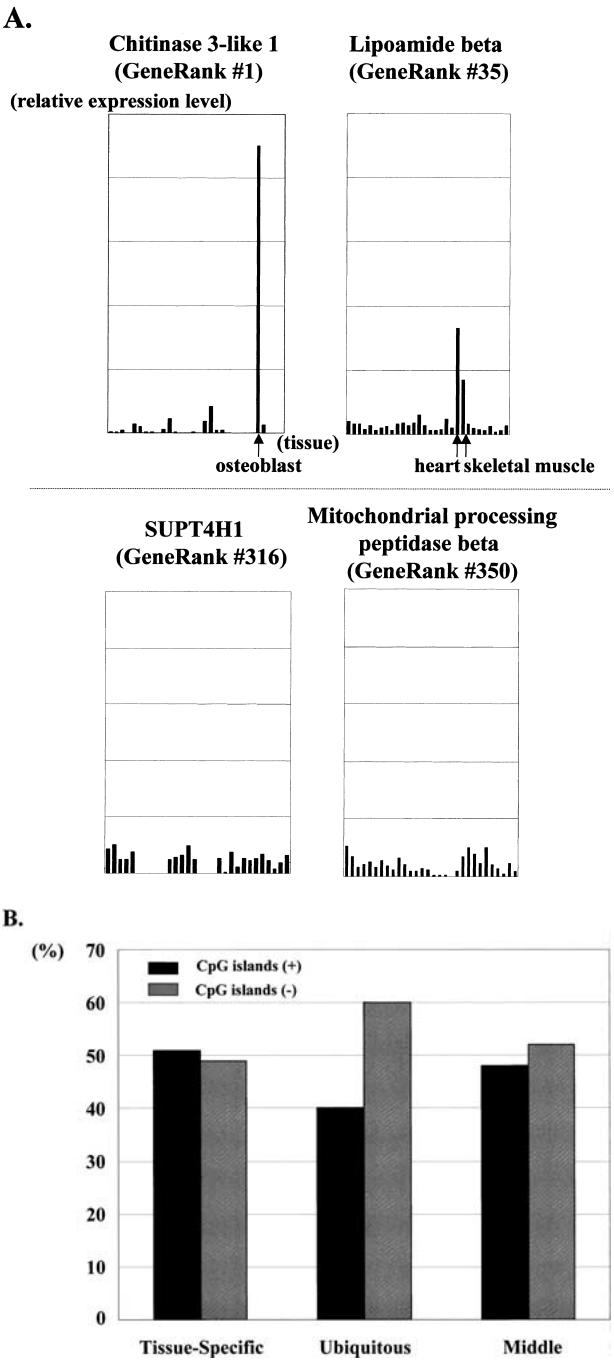

Figure 4 (*A*) Expression profiles of *chitinase 3-like 1* (GeneRank no. 1), *lipoamide beta* (GeneRank no. 35), *SUPT4H1* (GeneRank no. 316), and *mitochondrial-processing peptidase beta* (GeneRank no. 350) observed by iAFLP. Vertical axes represent the relative expression level; horizontal axes represent the tissue distributions. The expression level was designated so that the total values should be 30. (*B*) Populations of tissue-specific, ubiquitous, and middle genes located in/outside of CpG islands. Solid bars represent population of potential promoter regions (PPRs) located in CpG islands; shaded bars represent PPRs outside of CpG islands.

iAFLP database and retrieved expression profiles for 350 PPRs.

Some of the genes showed highly tissue-specific expression patterns. We classified the 350 PPRs according to the tissue specificity of the expression using GeneRank (http://bodymap.ims.u-tokyo.ac.jp). GeneRank is a program that calculates the scores of tissue specificity based on rigorous indexes as to whether the expression pattern is statistically biased (in prep.). We selected the top 10% of the genes (GeneRank nos. 1–35) according to the GeneRank scores and defined these genes as tissue-specific genes. The most significant tissue-specific expression pattern was observed for *chitinase 3-like 1* in osteoblast (Fig. 4A). The genes ranked at bottom 10% (GeneRank nos. 316–350) were defined as ubiquitously expressed genes and the genes ranked in between (GeneRank nos. 36–315) were defined as middle genes. The expression patterns of *chitinase 3-like 1* (GeneRank no. 1), *lipoamide beta* (GeneRank no. 35), SUPT4H1 (GeneRank no. 316), and *mitochondrial-processing peptidase beta* (GeneRank no. 350) are shown in Figure 4A.

We examined the association between the presence of the CpG islands in the PPRs and the tissue-specific expression of the corresponding genes because CpG islands are thought to be associated with housekeeping genes (Larsen et al. 1992; Cross and Bird 1995). The frequencies of CpG (+) PPRs were almost the same with that of CpG (−) PPRs for tissue-specific genes or for middle genes (Fig. 4B; see also Kusuda et al. 1993). This coincides with the result of all the 1031 PPRs (48%, see Table 1). However, the frequency of CpG (+) PPRs was significantly reduced for ubiquitous genes. This discrepancy with the previous view was unexpected. At present, we do not have a good explanation for this observation.

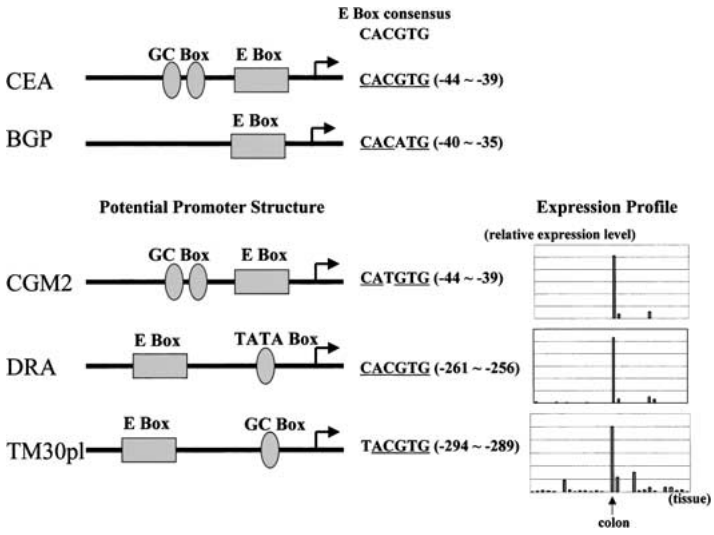

Figure 5 Potential promoter structures and expression profiles of the *CGM2* gene, the *DRA* gene, and the *TM30pl* gene are shown. The promoter structures were predicted from corresponding potential promoter region (PPR) sequences using TFBIND. Previously reported promoter structures of the *CEA* gene and the *BGP* gene are shown at *top*. Consensus sequence of E box and the sequences of predicted E boxes are shown to the *right* of the promoter structures. The nucleotides that match the consensus sequence are underlined. Each position of the predicted E box is also shown. The expression profile observed by iAFLP is shown at *right* for each gene (for more details, see http://bodymap.ims.u-tokyo.ac.jp).

Association Between the TF-Binding Sites in the PPRs and the Expression Profile

We examined whether there is correlation between the presence of a certain TF-binding sites in PPRs and the expression profiles revealed by iAFLP. We selected the *CGM2* gene (carcinoembryonic antigen family member 2), *DRA* gene (colon mucosa-associated gene), and *TM30pl* gene [fibroblast tropomyosin TM30 (pl)] genes, which showed highly colon-specific expression patterns (GeneRank nos. 11, 36, and 50, respectively). In Figure 5, the potential promoter structure and the expression profile of each gene are shown. In each of the PPRs, the consensus sequence of the E box was observed. The *CGM2* gene has other members of a closely related gene family, the *CEA* gene (carcinoembryonic antigen) and the *BGP* gene (biliary glycoprotein), both of which are also expressed in colon (Thompson et al. 1991). As for the *CEA* gene and the *BGP* gene, the promoter structures have been well characterized and the E boxes in their promoters have been reported to play essential roles in their transcriptional regulations (Hauck et al. 1994; Hauck and Stanners 1995). Because sequence homology of the PPRs of the *CGM2* gene and the promoter of the *CEA* gene was >80% (data not shown), it is suggested that the E boxes in the PPR of the *CGM2* gene are also involved in the transcriptional regulation. Also, the E boxes were observed in the PPRs of the *DRA* gene and the *TM30p1* gene, which showed similar expression patterns with the *CGM2* gene. It is likely that these E boxes have functional roles for these genes. Although more experimental analyses should be done before it can concluded that these E boxes actually have functional roles, this approach should give a useful clue to infer the involvement of certain TFs in the transcriptional regulation of a gene.

Mapping of the SNPs onto the PPR Sequences

Promoter regions are also important targets for exploring single nucleotide polymorphisms (SNPs; Brookes 1999). Recently, several attempts have been reported to associate SNPs in the promoters with disease predispositions. One of the most successful results may be analysis of a SNP in the promoter of the *TNF* gene. This SNP has been shown to affect the affinity of a transcrip-

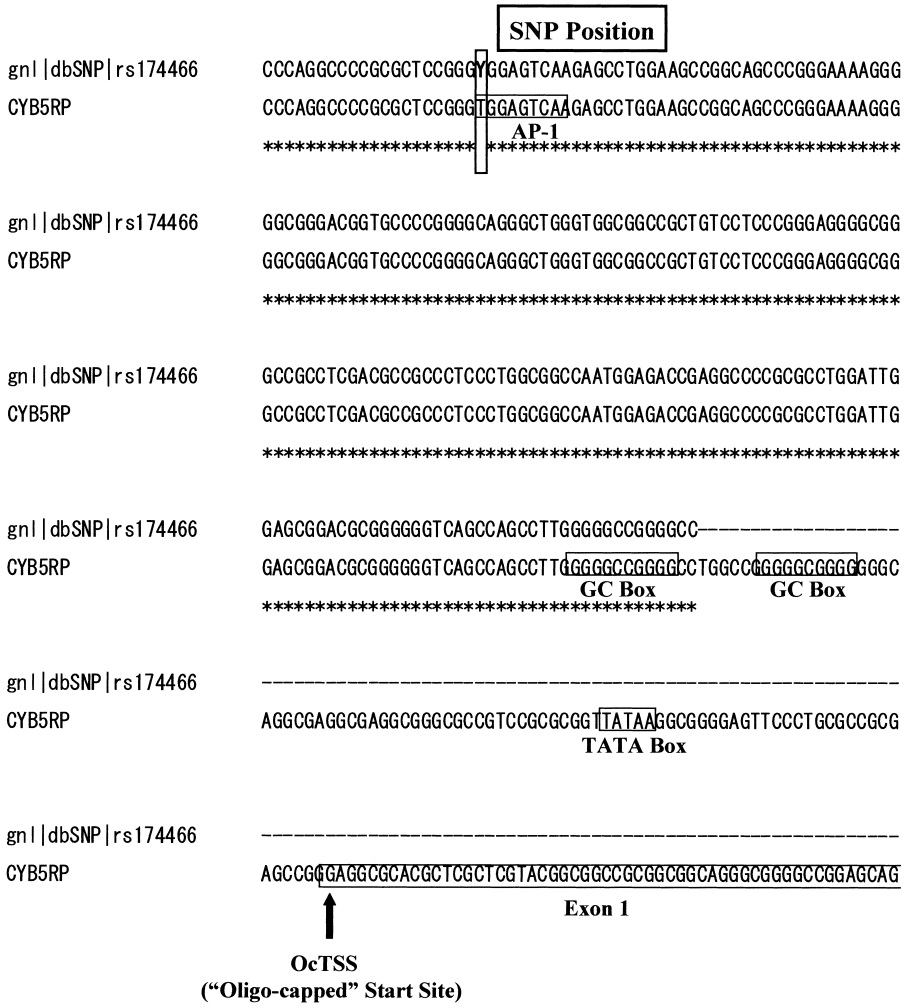

Figure 6 Identification of a SNP in the potential promoter region (PPR) of *CYB5RP.* The SNP position is shown by a bold letter Y (C or T) and a box. The DDBJ/EMBL/GenBank accession number of the corresponding SNP in dbSNP is shown at *left*. Consensus sequences of TF-binding sites predicted by TFBIND are also shown by boxes.

tion factor, *OCT-1,* and increase the susceptibility to severe malaria infection (Knight et al. 1999).

We searched the PPRs for SNPs reported in dbSNP (http://www.ncbi.nlm.nih.gov/SNP/index.html). We downloaded 595,893 SNP data (as of September 25, 2000) and mapped 119 SNPs successfully onto the PPRs. Some of these SNPs turned out to be located in the TF-binding consensus sequences. Figure 6 shows the case of the *CYB5RP* (delta-6 fatty acid desaturase) gene. When the allele contains C instead of T, the consensus sequence of AP-1-binding site would be destroyed because of this SNP. Although it has not been reported that the *AP-1* is involved in the transcriptional regulation of the *CYB5RP* gene, this information should be a useful clue to identify the promoter SNPs that have functional consequences.

In this report, we identified and characterized the PPRs of 1031 kinds of human genes. In addition to the promoter elements described here, other factors should also have important roles in transcription. For example, enhancers located distantly from the mRNA start site, and whether the chromatin in the promoter is formed or remodeled, should also be considered. Even though our data are based on the sequence adjacent to the transcription start sites only, because almost no reports have described genome-wide features of promoters, the present study should lay groundwork for better understanding of the mechanism of transcription.

METHODS

Construction of Oligo-Capped cDNA Libraries and Sequencing Analysis

Oligo-capped cDNA libraries were constructed as reported previously (Suzuki et al. 1997, 2000; Sambrook et al. 1989). Clones (100,000) were randomly isolated from 34 kinds of these cDNA libraries and one-pass sequenced from their 5′ ends using an ABI 377 XL Auto-Sequencer. The list of cDNA libraries was described elsewhere (Suzuki et al. 2000).

Identification of the PPRs

Sequence similarity was searched against Genbank (Release 102.0) using BLASTN (Altschul et al. 1990). The cDNAs matched with known genes were collected after removing the oligo-capped 5′-oligonucleotide sequence from each 5′ end. The cDNAs lacking the reported translation initiator ATG were removed from the database as erroneous products of the oligo-capping method. The selected cDNAs were clustered to create a nonredundant set of 2251 cDNAs by round-robin search of BLASTN. When the multiple start sites were observed (Y. Suzuki, H. Taira, S. Tsunoda, J. Sese, J.S. Mizushima, H. Hata, T. Ota, T. Isogai, T. Tanaka, Y. Sakaki, A. Suyama, S. Morishita, Y. Nakamura, K. Okubo, and S. Sugano, in prep.), cDNAs containing the longest 5′ ends were selected as representatives. For alignment of the 5′ end of the cDNA with the genome sequence, genome sequences were downloaded from FTP sites of GenBank (ftp://ncbi.nlm.nih.gov/genbank/genomes/H_sapiens/) on February 8, 2000, when draft and finished sequences altogether had covered about 60% of the entire human genome. The genomic sequences were first roughly searched with BLAST, using 100-bp sequences from the 5′ ends of oligo-capped cDNAs. The exact alignment between cDNAs and genome sequences was confirmed with CLUSTALW (Thompson et al. 1994). When 23 bp out of 25 bp from the 5′

ends of the cDNAs were matched, the corresponding genomic sequences were retrieved. The promoters were defined as the sequences extending from 500 bp upstream to 100 bp downstream of the mapped 5′ ends of the oligo-capped cDNAs. In the numbering scheme used here, the mRNA start site identified by the longest oligo-capped cDNA is designated as +0. Negative and positive integers indicate 3′ and 5′ relative to +0, respectively. Repetitive sequence elements, such as *Alu*, were masked using CENSOR (Jurka et al. 1996). Four PPRs that had been derived from the nonspecific genomic sequences with high homology to the 5′ ends of the cDNAs used for the search, such as sequences adjacent to pseudogenes were excluded from the data set as erroneous products.

Identification of the TF-Binding Sites and CpG Islands in the PPRs

The search of possible TF-binding sites using TFBIND was performed as described previously (Tsunoda and Takagi 1999). For calculating matching scores, Bucher's calculating method (Bucher 1990) and TF frequency matrices in TRANSFAC (Rel. 4.0; http://transfac.gbf.de/index.html) were used. The optimized cutoff values, preferred region, and searched region of each TF-binding motif are shown in Table 1. On each TF-binding matrix, the number of promoters whose matching scores were above the optimized cut-off values were counted. The methods for optimizing the cutoff values and determining the preferred regions were described elsewhere (Tsunoda and Takagi 1999).

For the analysis of CpG islands, the moving average for %(G + C) and the CpG ratio were calculated for each sequence, using a 100-bp window moving along the sequence at 1-bp intervals. The CpG ratio was calculated according to the standard method (Gardiner-Garden and Frommer 1987): (number of CG $\times$ N)/(number of C $\times$ number of G), where N is the total number of nucleotides in the sequence being analyzed. CpG islands were defined as regions >200 bp with %(G + C) > 50% and CpG ratio > 0.6. Repetitive sequence elements were not included in the calculation.

Resources for Databases and Computer Programs

Human genomic sequences were from ftp://ncbi.nlm.nih.gov/genbank/genomes/H_sapiens/ as of February 8, 2000. HIB were from http://www.mips.biochem.mpg.de/proj/human/selec_view.html. All of the iAFLP data was from http://bodymap.ims.u-tokyo.ac.jp. TRANSFAC was from http://transfac.gbf.de/index.html (Heinemeyer et al. 1999). TFBIND was from T. Tsunoda (Tsunoda and Takagi 1999). GenRank was from J. Sese (University of Tokyo). BLASTN, CLUSTALW, and CENSOR were from the Human Genome Center in the Institute of Medical Sciences, University of Tokyo.

ACKNOWLEDGMENTS

We thank Y. Shirai, Y. Takahashi, T. Tsunoda, T. Mizuno, M. Morinaga, M. Kawamura, and K. Mizuno for their excellent sequencing work. We are also grateful to M. Hida, M. Sasaki, and T. Ishihara for their technical support, D. Ramana, M. Zhang, S. Watanabe, K. Kataoka, J. Imai, T. Komatsu, M. Watanabe, T. Togashi, and N. Osada for helpful discussions, and E. Nakajima for critical reading of the manuscript. This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports, and Culture of Japan and by special coordination funds for promoting science and technology (SCF) from the Science and Technology Agency (STA) of Japan.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Berk, A.J. and Sharp, P.A. 1977. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell* **12:** 721–732.

Brookes, A.J. 1999. The essence of SNPs. *Gene* **234:** 177–186.

Bucher, P. 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212:** 563–578.

Costello, J.F., Fruhwald, M.C., Smiraglia, D.J., Rush, L.J., Robertson, G.P., Gao, X., Wright, F.A., Feramisco, J.D., Peltomaki, P., Lang, J.C., et al. 2000. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat. Genet.* **24:** 132–138

Cross, S.H. and Bird, A.P. 1995. CpG islands and genes. *Curr. Opin. Genet. Dev.* **5:** 309–314.

Gardiner-Garden, M. and Frommer, M. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196:** 261–282.

Hauck, W. and Stanners, C.P. 1995. Transcriptional regulation of the carcinoembryonic antigen gene. Identification of regulatory elements and multiple nuclear factors. *J. Biol. Chem.* **270:** 3602–3610.

Hauck, W., Nedellec, P., Turbide, C., Stanners, C.P., Barnett, T.R., and Beauchemin, N. 1994. Transcriptional control of the human biliary glycoprotein gene, a CEA gene family member down-regulated in colorectal carcinomas. *Eur. J. Biochem.* **223:** 529–541.

Heinemeyer, T., Chen, X., Karas, H., Kel, A.E., Kel, O.V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F., and Wingender, E. 1999. Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res.* **27:** 318–322.

Jurka, J., Klonowski, P., Dagman, V., and Pelton, P. 1996. CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Computers Chem.* **20:** 119–122.

Kato, K. 1997. Adaptor-tagged competitive PCR: A novel method for measuring relative gene expression. *Nucleic Acids Res.* **25:** 4694–4696.

Kawamoto, S., Ohnishi, T., Kita, H., Chisaka, O., and Okubo, K. 1999. Expression profiling by iAFLP: A PCR-based method for genome-wide gene expression profiling. *Genome Res.* **9:** 1305–1312.

Knight, J.C., Udalova, I., Hill, A.V., Greenwood, B.M., Peshu, N., Marsh, K., and Kwiatkowski, D. 1999. A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria. *Nat. Genet.* **22:** 145–150.

Kusuda, J., Hirata, M., Toyoda, A., Takahashi, I., and Hashimoto, K. 1993. A search for CpG islands associated with genes in human genomic sequences compiled in the DNA database. *Genome. Inform. Works.* **4:** 239–244.

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

Maruyama, K. and Sugano, S. 1994. Oligo-capping: A simple method to replace the cap structure of eucaryotic mRNAs with oligoribonucleotides. *Gene* **138:** 171–174.

McKnight, S.L. and Kingsbury, R. 1982. Transcriptional control signals of a eukaryotic protein-coding gene. *Science* **217:** 316–324.

Mitchell, P.J. and Tjian, R. 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245:** 371–378.
Novina, C.D. and Roy, A.L. 1996. Core promoters and transcriptional control. *Trends Genet.* **12:** 351–355.
Perier, R.C., Praz, V., Junier, T., Bonnard, C., and Bucher, P. 2000. The eukaryotic promoter database (EPD). *Nucleic Acids Res.* **28:** 302–303.
Roeder, R.G. 1996. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21:** 327–335.
Sambrook, J., Fritsh, E.F., and Maniatis, T. 1989. *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Schaefer, B. 1995. Revolutions in rapid amplification of cDNA ends: New strategies for polymerase chain reaction cloning of full-length cDNA ends. *Anal. Biochem.* **227:** 255–273.
Smale, S.T. 1997. Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim. Biophys. Acta* **1351:** 73–88.
Suzuki, Y., Yoshitomo, K., Maruyama, K., Suyama, A., and Sugano, S. 1997. Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. *Gene* **200:** 149–156.
Suzuki, Y., Ishihara, D., Sasaki, M., Nakagawa, H., Hata, H., Tsunoda, T., Watanabe, M., Komatsu, T., Ota, T., Isogai, T., Suyama, A., and Sugano, S. 2000. Statistical analysis of the 5′ untranslated region of human mRNA using oligo-capped cDNA libraries. *Genomics* **64:** 286–297.
Thompson, J.A., Grunert, F., and Zimmermann, W. 1991. Carcinoembryonic antigen gene family: Molecular biology and clinical perspectives. *J. Clin. Lab. Anal.* **5:** 344–366.
Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680
Tsunoda, T. and Takagi, T. 1999. Estimating transcription factor bindability on DNA. *Bioinformatics* **15:** 622–630.

Received September 5, 2000; accepted in revised form February 5, 2001.