Table 1. Boundaries of the Blocks

	Human			Mouse		
Repeat	31%			20%		
		Alu-type SINE	16%		B1-type SINE	8%
		MIR-type SINE	3%		B2-type SINE	4%
		LINE	6%		LINE	3%
		LTR	2%		LTR	3%
		MER	2%		MER	2%
		others	1%		others	0%
Gap in genomic sequence	0%			4%		
Uncharacterized	69%			76%		
Total	100%			100%		

Indicated sequences were observed at the corresponding frequencies at the boundaries of the blocks.

around the boundaries of the blocks were evaluated, the average G+C contents were 58% and 53% in the sequences inside (proximal sides to the TSSs) and outside (distal sides to the TSSs) of the blocks, respectively. The difference overall distributions of the G+C contents between them was statistically significant according to the standard *t* test ($p < 1.0e\text{-}136$), although the G+C contents vary between PPRs. The average frequencies of the dinucleotide, CpG, were 12.7 sites/200 bp and 9.0 sites/200 bp for the regions inside and outside the blocks, respectively. Again, the difference in their distributions was statistically significant ($p < 1.0e\text{-}105$). As shown in Table 3, essentially the same results were obtained from mouse PPRs. This observation also supports our claim that the sequences inside and outside of the blocks are qualitatively distinct.

Mapping of TF-Binding Sites

To study the relationship between the relative positions of the blocks and the TF-binding sites embedded in the upstream regions, we mapped previously determined TF-binding sites. For this, we used the information contained in TRANSFAC (version 7.4). This database is the most widely used database in which detailed information concerning TF-binding sites, which have been characterized by various experimental methods, is compiled (Kel et al. 2003). In the 3324 promoter pairs, there were 238 experimentally characterized TF-binding sites for human genes (further references about each of the TF-binding sites are recorded in TRANSFAC). Of these, 203 sites (85%) were located in regions proximal to the TSSs (within the −1 kb to +200 bp regions), which is consistent with previous observations that most TF-binding sites were located within this region (Praz et al. 2002; Liu et al. 2003). Among the TF-binding sites, 179 sites (88%) were located within the blocks. On the other hand, we also observed that 24 sites (12%) in human genes were located outside of the blocks, where no significant sequence similarities were found. For each of these sites, we both manually and computationally examined whether the same kind of TF-binding site could be identified in the corresponding regions of the promoter sequences of the mouse gene. All of these sites were completely missing from the corresponding regions of the mouse promoters, although there still remains a slight possibility that real TF-binding sites are located in regions distant from the TSSs, or that the TF binding sites were so diverged that they could not be identified using a computational method.

We performed similar analyses with regard to the computationally predicted TF-binding sites. Among the 1898 predicted TF-binding sites in human PPRs, 1704 (90%) were located within the blocks and 194 (10%) outside of the blocks. This corresponds well with the above results regarding the "experimentally characterized" TF sites. Essentially similar results were obtained from analyses from the mouse side, too (Table 3).

Correlation Between Sequence Conservation in the Promoters and Molecular Functions and Tissue Specificity of the Genes

We examined whether there is any correlation between sequence divergence of the promoters and molecular functions and expression patterns of the corresponding genes. We calculated the frequency of the PPRs in which blocks covered less than 50% (600 bp) of the sequences (designated as "encroached" PPRs) for each of the GO categories (Harris et al. 2004). Similarly, the frequency of those promoters was calculated for each population of the genes that showed tissue-specific patterns of gene expressions. For the expression profiles, we used the data obtained by iAFLP, which is an RT-PCR-mediated high-throughput method for de-

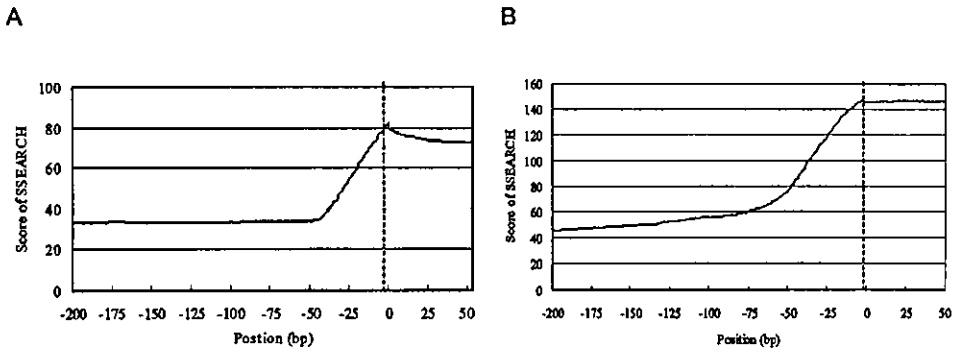


Figure 3 Sequence alignments around the boundary of the block and that of the first intron and the second exon using SSEARCH. (*A*) Sequences of human and mouse PPRs were aligned using SSEARCH with a 50-bp moving window around the boundary of the block. The broken line represents the boundary of the block calculated using LALIGN. The vertical axis represents the average score of the SSEARCH calculated for the corresponding position. The horizontal axis represents the relative position to the boundary. (*B*) Result of an analysis similar to that shown in *A*, using the proximal sequences of the 5′ end of the second exons. The broken line represents the exon–intron boundary. The horizontal axis represents the relative position to the exon–intron boundary.

Table 2. G+C Content and CpG Frequency Inside and Outside the Blocks

		Outside of −1 Kb to +200 bp	Within −1 kb to +200 bp	Within block	Outside of block
Human	Experimentally confirmed	35	203	179 (88%)	24 (12%)
	Predicted	ND	1898	1704 (90%)	194 (10%)
Mouse	Experimentally confirmed	31	108	102 (94%)	6 (6%)
	Predicted	ND	1853	1668 (90%)	185 (10%)

The sequences ±200 bp of the boundaries of the "blocks" were used for the calculation. ND = not determined.

tecting relative amounts of gene expression (Kawamoto et al. 1999; the iAFLP data used in this study are presented at http://cdna.ims.u-tokyo.ac.jp/iAFLP.xls). We tentatively defined the genes as "tissue specific" when more than 30% of the transcripts were attributed to a particular tissue.

As shown in Table 4A, the frequency of the encroached PPRs was significantly increased in the GO category of "transcription regulators", which is the group of genes of TFs ($p < 0.0002$). In the 203 TF genes, the frequency of the genes with such promoters was 39%, which was higher than the frequency calculated for any other GO category. We also observed that encroached PPRs were enriched in genes whose expression patterns were "brain specific" (Table 4B). Although statistical significance in this case was not as clear as the case of the transcription regulators, the enrichment was higher than any other tissues ($p < 0.05$).

DISCUSSION

Here we have described the first systematic and quantitative comparison of promoters regarding the manner in which and the extent to which promoter sequences are conserved between human and mouse genes. Using 3324 pairs of PPRs of human and mouse genes, we first demonstrated that the conserved parts frequently stood out against the nonconserved parts, forming blocks. The sequence similarities of around 65% in these blocks extended upstream of the TSSs and disappeared at particular points, on average, 510 bp upstream of the TSSs. This is inconsistent with the view generally held hitherto. The initial descriptions of the sequence similarity among promoters indicated that the independent alternations of the nucleotides are distributed in a gradually increasing manner in proportion to the distance from the TSSs (as shown in Fig. 1). Although the results of a previous study using 41 human–mouse promoter pairs suggested the block structure of the sequence conservation in the promoters, it was considered likely to be an artifact of the alignment program used (Jareborg et al. 1999). In the present study, we scrutinized the sequence alignments mainly using two alignment programs that are based on different algorithms and demonstrated that the block structures were observed regardless of the alignment programs in about one-third of the examined PPRs (Figs. 2, 3; for further details on the alignment programs, see Ureta-Vidal et al. 2003).

There still remains some possibility that the block structure observed in the present study was identified due to the inherent inability of the pre-existing alignment programs, most of which are designed for aligning sequences of genic (especially of protein-coding) regions. Also, we could not completely refute the possibility that alignment procedures employed here were not suitable for detecting relatively short motifs outside putative blocks separated by constitutive insertion or deletions of the nucleotides. However, we consider that such a possibility is low, because we selected relatively simple programs, LALIGN and SSEARCH, run by parameters for which no special "parameter tuning" was performed a priori. We also demonstrated that this observation was robust against the changes of the parameters (Supplementary data Figs. 2 and 3). Although it is possible further "optimization" of the programs and parameters may be useful for further precise determination of the boundaries of each of the blocks, we consider such perturbation would not greatly influence our conclusion that the segmentation occurred just around the TSSs very frequently.

It was also unlikely that our observations were obtained due to defects in our data set. Only rare data should represent spuriously identified promoter sequences resulting from erroneously cloned full-length cDNAs (truncated cDNAs), because, in most cases, the sequences could be aligned at least to some extent. If the promoters were spurious at all, they would not show any significant match against their counterparts. Mispairing of paralogs as orthologs could bring about the results observed here. As paralogs are generated by gene duplication (Frazer et al. 2003b), it is possible that there is some synteny just around the genic regions, which disappears at the boundaries of the duplication points. However, at least 80% of mouse genes have only a single identifiable homologous gene in the human genome, which should be an ortholog (Waterston et al. 2002). Also, we used the pairing information of the orthologs according to LocusLink information, in which 1:1 homologous genes are further inspected to pair orthologs (Wheeler et al. 2004). This should have excluded any remaining pseudo-orthologous pairs from our data set. Considering that the block structure was observed for more than one-third of the promoters, contamination by paralogs should not account much for our observations.

Based on all these facts and our findings, we concluded that the block structure is, in fact, a feature of the sequence conservation in about one-third of the PPRs examined here. We consider that this discontinuous manner of the sequence conservation should be a quite frequent feature of promoters throughout the human and mouse genomes. Although we could not show whether such discontinuous conservation would be observed in more distal regions from the TSSs in the gene of the remaining population, it is significant that such dynamic changes occurred just proximal regions of the TSSs at least one-third of the PPRs. In order to understand how the transcription modulation has evolved, this information should become the fundamental data.

Within the blocks, the sequence similarity was relatively uniform (Fig. 2) with an average identity of 65%. The overall

Table 3. TF Binding Sites Inside and Outside the Blocks

	Human		Mouse	
	Inside block	Outside block	Inside block	Outside block
G+C content	58%*	53%	56%	48%
CpG frequency (sites/200 bp)	12.7**	9.0	11.0	6.0

The frequencies of the TF binding sites were calculated for each of the indicated regions. Statistical significance of the enrichment was *$p < 1.0e\text{-}136$ and **$p < 1.0e\text{-}105$.

Table 4. Correlation Between the Gene Ontology, Expression Profiles, and Sequence Conservation in the PPRs

A. GO annotation	Total number of genes	Number of genes with encroached PPRs	Frequency (%)
Transcription regulator	203	79	39*
Structural molecule	125	35	28
Enzyme	871	225	26
Enzyme regulator	91	23	25
Cell adhesion molecule	56	14	25
Defence/immunity	29	7	24
Transporter	342	81	24
Signal transducer	362	78	22
Total	3324	921	28

B. Tissue	Total number of genes with tissue-specific gene expression	Number of genes with encroached PPRs	Frequency (%)
Brain/neuron	156	53	34**
Gastrointestinal	121	35	29
Immune	98	29	30
Reproductory	137	34	25
Endocrine	17	4	24
Circulatory/blood	22	5	24
Others	148	42	29
Total	3324	921	28

The numbers and the frequencies of the genes were shown for each of the GO (*A*) and iAFLP (*B*) categories. Statistical significance of the enrichment was $*p < 0.0002$ and $**p < 0.05$, respectively (for further details on the procedure, see Methods).

sequence similarity between human and mouse at neutral sites has been estimated to be 53–54%, when assessed using relics of ancestral repeats (Waterston et al. 2002). If the regional variations of the neutral substitution rate are ignored (Hardison et al. 2003), the sequence identity is approximately 10% higher in the sites within the blocks. This difference implies that some parts of the promoters are subjected to selective pressure. Largely uniform sequence similarities within blocks were observed, maybe because the positions of the TF-binding sites are different between genes, allowing degeneracy within them to some extent. It is also possible that additional sequences as well as direct binding sites of TFs themselves should also be conserved, considering that the cognate sequences of the TFs are typically 6–10 bp long (Wray et al. 2003). Particular subregions of the promoter may not have been allowed to undergo free sequence divergence because the overall base composition or relative positions of TF-binding sites needed to be preserved. This could also explain the relatively flat patterns of sequence similarities within blocks. Extensive phylogenetic comparative analyses using forthcoming genomic sequences of other mammals (http://www.genome.gov/11007951) together with recently developed statistical methods (Elnitski et al. 2003) should lead to a more precise understanding of which sequences play a leading part, (serving as direct binding sites for TFs), and which play a supporting role.

We also observed that the sequence identity dropped just outside the blocks. It is possible that this is due to a discontinuous rate of random sequence substitution at the corresponding regions, despite the fact that the sequences themselves were continuous. However, the sequence identity outside the blocks was no more than 30%, even if the sequences were forced to be aligned (data not shown). This rate is somewhat lower than the conservation rate at neutral sites. It is unlikely that such extreme hot spots of random mutation are distributed within the regions 1 kb upstream of TSSs at such a frequency. It is more natural to suppose that totally unrelated sequences exist just outside the blocks. Consistently, the G+C content and CpG frequency were higher inside the blocks than outside (Table 3). This may also reflect that the sequences outside the blocks were foreign to the promoter sequences.

Genomic rearrangements, such as deletions, insertions, or recombination, may have taken place around the distal regions of the blocks. It is possible that the human genome has been rearranged significantly more in the course of evolution than previously thought. Although further confirmation is necessary, our result shown in Figure 2D also supports the idea that such segmentation prevails throughout the human and mouse genomes. Consistent with this possibility, recent publications have provided evidence that a large proportion of previously identified human–mouse syntenic regions contain multiple microrearrangements (Pevzner and Tesler 2003). Frazer et al. (2003a) observed genomic deletions, ranging from 0.2 to 8 kb in size, even between humans and chimpanzees. In particular, they observed integration of repetitive elements at the 3′-end boundaries of deletions in 23 out of 47 cases. In the present study, we showed that 46% of the 5′ ends of the blocks were bounded by interspersed repeats on either the human or mouse side (Table 1). Sometimes, the repetitive sequences may have acted as nucleation points for homologous recombination. In fact, it has been reported that this type of retroelement-mediated recombination has occasionally taken place in the human genome and is estimated to be responsible for at least 0.3% of human genetic disorders (Batzer and Deininger 2002).

Deletion of TF-binding sites could have accompanied some of the rearrangements. However, alterations that occurred inside the transcriptional regulatory modules in the promoters would mostly have been unfavorable for proper biological functions, and thus, would have been deleted from the population. The "block" structure we identified in the present study seemed to have formed as a consequence of such selective pressure. We observed that most of the previously characterized TF-binding sites were located within the blocks (Table 2). For these TF-binding sites, the cognate sequences as well as the relative positions of the TF-binding sites and distances to the TSSs were preserved.

Alterations that occurred outside blocks may generally have been tolerated. Some might have led to the acquisition of altered modes of transcriptional modulation. It has been reported that polymorphisms that cause an approximately twofold difference in transcription activation activities frequently occur without showing organismal phenotypes within human populations (Rockman and Wray 2002). Repetitive elements at the boundaries of the blocks could contribute to such modifications. There are a number of examples in which retroelements integrated in the vicinity of TSSs became involved in transcriptional regulation via changes in their sequences (Norris et al. 1995; Vansant and Reynolds 1995; Hamdi et al. 2000). It is likely that such variations have accumulated during evolution and have laid the genetic background to drive speciation during certain periods of time.

Intriguingly, we observed that the blocks in the PPRs were most encroached in the genes encoding transcription factors and genes whose expression patterns are brain specific (Fig. 3). This suggests that alterations within the proximal regions of the TSSs have been accumulated for these gene populations. It is possible that evolutionary diversification between humans and mice has been caused by slight changes in the regulation by TFs, which are located at the apexes of the regulatory hierarchy of transcriptional networks, rather than changes of the downstream proteins. Moreover, the evolutional changes may be the most significant in the genes expressed and functioning in the brain, which is the most distinctly different organ between humans and mice. Further characterization of the TF-binding sites that are similar to or distinctive in mice and humans as well as cross-validation of expression analyses should help to elucidate the molecular mechanisms underlying the alterations in transcriptional modulation responsible for the speciation of humans and mice. To this end, the present work has provided a first glimpse of how the modulation of transcriptional networks is likely to have differentially evolved between humans and mice.

MATERIALS AND METHODS

Promoter Data Set

The putative promoter regions were extracted by computational mapping of the 5′ ends of the human and mouse full-length cDNA sequences onto the corresponding genomic sequences obtained from UCSC Genome Browser (human: hg13; mouse: mm2). In total, 400,225 human and 580,209 mouse cDNAs were used to retrieve 8793 human and 6875 mouse promoters by the sequential use of BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start; BLAT) and SIM4 (http://pbil.univ-lyon1.fr/sim4.php; SIM4). The identified promoters were located about 4 kb upstream of the 5′ ends of the previously registered public cDNA sequences on average. Among the retrieved promoters, 3324 were correlated with each other as putative mutually orthologous genes using the table obtained from ftp://ftp.ncbi.nih.gov/pub/HomoloGene/. The statistics of the generated promoter data set are provided as Supplemental data Table 1. Details of the procedures for cDNA mapping and promoter pairing are described in Suzuki et al. (2004). Further information on the gene definitions used for the present study is also available in Supplemental data Table 1. As described there, at least two-thirds of the promoters were supported by three independently isolated full-length cDNAs. Considering that the average frequency of the full-length cDNAs (full-length-ness) in each of the libraries is >70%, there should be little chance that all of them are truncated. Also, we discarded all of the CDS-minus cDNAs, which increased the full-length-ness even more (for further discussion of this issue, please refer to Suzuki et al. 2001).

Sequence Alignment of the Promoters

LALIGN was obtained from http://www.ch.embnet.org/software/LALIGN_form.html and used for aligning sequences of the promoters with the default settings in the main text. The results of similar analyses using different parameter sets are shown in Supplemental data Figure 2. When LALIGN results split the sequence alignments allowing a large gap(s), most distal positions were recognized as the boundaries of the blocks. A graphical view of the sequence alignment and calculated sequence identities are shown in Supplemental data Figure 1.

For aligning nongenic regions, the putative syntenic regions were obtained according to the information from the UCSC genome alignment map (http://genome.cse.ucsc.edu/goldenPath/14nov2002/vsMm2/axtTight/). The alignments located within 100 kb for the Ensembl regions (http://genome.ucsc.edu/goldenPath/14nov2002/database/) were excluded and the 183,733 boundary sequences ranging from −1 kb to +200 bp were retrieved. Using these sequences, the alignments were generated using LALIGN.

SSEARCH was obtained from ftp://ftp.virginia.edu/pub/fasta/ as FASTA package programs. SSEARCH was used with default parameters for the detailed alignment of the sequences at the distal regions of the blocks and the proximal regions of the 5′ ends of the second exons. The results of a similar analysis using different parameter sets are shown in Supplemental data Figure 3.

Search for the Repetitive Sequences in the Promoters

The positions of the boundaries of the blocks were compared with those of annotated repetitive sequences. For positional information about the repetitive sequences, http://genome.ucsc.edu/goldenPath/14nov2002/bigZips/chromOut.zip and http://genome.ucsc.edu/goldenPath/mmFeb2002/bigZips/chromOut.zip were used for the human and mouse genomes, respectively. Classification of the repetitive sequences was also as described there.

Computational Prediction of the Putative TF-Binding Sites in the Promoters

For information about previously experimentally characterized TF-binding sites, TRANSFAC Professional 74 was used. For the computational prediction of the putative TF-binding sites, the promoter sequences were surveyed using MATCH. For the predictions, the cutoff value set of minFP.prf, which has been demonstrated to minimize "false positives", were used.

Relating GO Criteria and Expression Profiles With the Sequence Divergence in the Promoters

The correlation tables between GO terms and RefSeq IDs were obtained from http://www.geneontology.org/. For each GO term, the frequencies of the promoters whose block lengths were greater or less than 600 bp were determined. As for the expression profiles, for those genes whose relative expression level was limited to a particular organ by more than 0.3, a similar calculation was performed. Classification of the organs is shown together with the iAFLP data file (http://cdna.ims.u-tokyo.ac.jp/iAFLP.xls). A detailed characterization of the iAFLP data will be published elsewhere.

Statistical significance of the difference in the frequencies of the encroached PPRs was evaluated by calculating hypergeometric distribution using the following equation:

$$\sum_{x=k}^{M} \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$$

where $N = 3324$, $n = 921$, $M = 203$, $k = 79$ ("transcriptional regulators") in the case of GO terms and $N = 3324$, $n = 921$, $M = 156$, $k = 53$ ("brain specific") in the case of expression profiles.

ACKNOWLEDGMENTS

We thank T. Hasui for his support with computational analyses of the promoters. We are grateful to K. Abe, M. Morinaga, M. Ishizawa, M. Kawamura, T. Mizuno, A. Kanai, and H. Hata for their excellent technical support. We are thankful to K. Koyanagi, M. Nakao, and E. Nakajima for helpful discussions and critical reading of the manuscript. This study was supported by a Grant-In-Aid for Scientific Research on Priority Areas and by Special Coordination Funds for Promoting Science and Technology (SCF), both from the Ministry of Education, Culture, Sports, Science, and Technology in Japan. This study was also supported by a Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science, and Technology of the Japanese Government to Y.H.

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

Batzer, M.A. and Deininger, P.L. 2002. *Alu* repeats and human genomic diversity. *Nat. Rev. Genet.* **3:** 370–379.

Boguski, M.S. 2002. Comparative genomics: The mouse that roared. *Nature* **420:** 515–516.

Carninci, P. and Hayashizaki, Y. 1999. High-efficiency full-length cDNA cloning. *Methods Enzymol.* **303:** 19–44.

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

Deininger, P.L. and Batzer, M.A. 2002. Mammalian retroelements. *Genome Res.* **12:** 1455–1465.

Elnitski, L., Hardison, R.C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M.J., Schwartz, S., Miller, W., and Chiaromonte, F. 2003. Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13:** 64–72.

Frazer, K.A., Chen, X., Hinds, D.A., Pant, P.V., Patil, N., and Cox, D.R. 2003a. Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* **13:** 341–346.

Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I., and Hardison, R.C. 2003b. Cross-species sequence comparisons: A review of methods and available resources. *Genome Res.* **13:** 1–12.

Hamdi, H.K., Nishio, H., Tavis, J., Zielinski, R., and Dugaiczyk, A. 2000. *Alu*-mediated phylogenetic novelties in gene regulation and development. *J. Mol. Biol.* **299:** 931–939.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13:** 13–26.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32:** D258–261.

Huang, X., Miller, W., Schwartz, S., and Hardison, R.C. 1992. Parallelization of a local similarity algorithm. *Comput. Appl. Biosci.* **8:** 155–165.

Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9:** 815–824.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., et al. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685–690.

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

Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E. 2003. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **31:** 3576–3579.

King, M.C. and Wilson, A.C. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107–116.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Liu, R., McEachin, R.C., and States, D.J. 2003. Computationally identifying novel NF-κB-regulated immune genes in the human genome. *Genome Res.* **13:** 654–661.

Mitchell, P.J. and Tjian, R. 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245:** 371–378.

Nadeau, J.H. and Taylor, B.A. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci.* **81:** 814–818.

Norris, J., Fan, D., Aleman, C., Marks, J.R., Futreal, P.A., Wiseman, R.W., Iglehart, J.D., Deininger, P.L., and McDonnell, D.P. 1995. Identification of a new subclass of *Alu* DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *J. Biol. Chem.* **270:** 22777–22782.

Novina, C.D. and Roy, A.L. 1996. Core promoters and transcriptional control. *Trends Genet.* **12:** 351–355.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.

Pearson, W.R. 1996. Effective protein sequence comparison. *Methods Enzymol.* **266:** 227–258.

Pevzner, P. and Tesler, G. 2003. Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes. *Genome Res.* **13:** 37–45.

Praz, V., Perier, R., Bonnard, C., and Bucher, P. 2002. The Eukaryotic Promoter Database, EPD: New entry types and links to gene expression data. *Nucleic Acids Res.* **30:** 322–324.

Rockman, M.V. and Wray, G.A. 2002. Abundant raw material for *cis*-regulatory evolution in humans. *Mol. Biol. Evol.* **19:** 1991–2004.

Roeder, R.G. 1996. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21:** 327–335.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

Suzuki, Y. and Sugano, S. 2003. Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. *Methods Mol. Biol.* **221:** 73–91.

Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Morishita, S., et al. 2001. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.* **2,5:** 388–393.

Suzuki, Y., Yamashita, R., Shirota, M., Sakakibara, Y., Chiba, J., Mizushima-Sugano, J., Kel, A.E., Arakawa, T., Carninci, P., Kawai, J., et al. 2004. Large-scale collection and characterization of promoters of human and mouse genes. *In Silico Biol.* **4:** 0036.

Tautz, D. 2000. Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* **10:** 575–579.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Ureta-Vidal, A., Ettwiller, L., and Birney, E. 2003. Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4:** 251–262.

Vansant, G. and Reynolds, W.F. 1995. The consensus sequence of a major *Alu* subfamily contains a functional retinoic acid response element. *Proc. Natl. Acad. Sci.* **92:** 8229–8233.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Wheeler, D.L., Church, D.M., Edgar, R., Federhen, S., Helmberg, W., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., et al. 2004. Database resources of the National Center for Biotechnology Information: Update. *Nucleic Acids Res.* **32:** D35–40.

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., and Romano, L.A. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20:** 1377–1419.

WEB SITE REFERENCES

http://dbtss.hgc.jp/; DBTSS.
http://fantom.gsc.riken.go.jp/; FANTOM.
ftp://ftp.virginia.edu/pub/fasta/; SSEARCH.
http://genome.ucsc.edu/cgi-bin/hgBlat?command=start; BLAT.
http://genome.ucsc.edu/downloads.html; UCSC Genome Browser.
http://pbil.univ-lyon1.fr/sim4.php; SIM4.
http://www.ch.embnet.org/software/LALIGN_form.html; LALIGN.
http://www.ensembl.org/; Ensembl.
http://www.epd.isb-sib.ch; Eukaryotic Promoter Database.
http://www.gene-regulation.com/; TRANSFAC.
http://www.geneontology.org/; GO.
http://www.ncbi.nlm.nih.gov/RefSeq/; RefSeq.
http://cdna.lms.u-tokyo.ac.jp/tAFLP.xls; tAFLP Expression Data.
http://www.genome.gov/11007951; NHGRI Genome Projects.
ftp://ftp.ncbi.nih.gov/pub/HomoloGene/; HomoloGene.
http://genome.ucsc.edu/goldenPath/14nov2002/database/; Ensembl at UCSC.
http://genome.ucsc.edu/goldenPath/14nov2002/bigZips/chromOut.zip; Human Genome.
http://genome.ucsc.edu/goldenPath/mmFeb2002/bigZips/chromOut.zip; Mouse Genome.
http://cdna.lms.u-tokyo.ac.jp/tAFLP.xls; tAFLP expression data.
http://genome.cse.ucsc.edu/goldenPath/14nov2002/vsMm2/axtTight/; Human–Mouse Alignment.

Received February 10, 2004; accepted in revised form June 23, 2004.

nature biotechnology

Promoter prediction analysis on the whole human genome

Vladimir B Bajic[1], Sin Lam Tan[1], Yutaka Suzuki[2] & Sumio Sugano[2]

Promoter prediction programs (PPPs) are important for *in silico* gene discovery without support from expressed sequence tag (EST)/cDNA/mRNA sequences, in the analysis of gene regulation and in genome annotation. Contrary to previous expectations, a comprehensive analysis of PPPs reveals that no program simultaneously achieves sensitivity and a positive predictive value >65%. PPP performances deduced from a limited number of chromosomes or smaller data sets do not hold when evaluated at the level of the whole genome, with serious inaccuracy of predictions for non-CpG-island-related promoters. Some PPPs even perform worse than, or close to, pure random guessing.

Recent availability of the human genome draft[1,2] has enabled analyses on the whole genome. However, no such large-scale analysis has been made regarding the performance of PPPs. Promoters are crucial control regions for transcriptional activation of every gene[3,4]. Development of PPPs has received a lot of attention[5–18] (see also reviews[19–21]). However, until the appearance of PromoterInspector[17], PPPs suffered from low accuracy. Following PromoterInspector, several efficient PPPs have been developed[5–9,12–16,18].

PPPs are built on different concepts. The underlying principle is that properties of promoter regions are different from properties of other genomic DNA. Many concepts are used, such as the presence of the CpG islands[7,8,14–16] close to transcription start site (TSS) locations, the presence of specific transcription factor binding sites (TFBSs)[9–13,18], possible higher density of potential TFBSs[19–20], statistical properties of proximal and core promoters as opposed to other genomic sequences[5–11], homology with orthologous promoters[18] and restricting the promoter prediction domain using information from mRNA transcripts[22]. Recognition technologies employed in PPPs are based on neural networks[5–7,10–12], linear and quadratic discriminant analyses[8,16,18], Relevance Vector Machine[9], statistical properties of promoter regions[5,7–9,12,14–17], interpolated Markov model[12,13], or a combination of these[5–17]. In general, PPPs perform better for a particular category of genomic sequences, such as G+C rich[9], whereas others[7,8,14–16] are more appropriate for the CpG-island-related promoters. One report[8] claims efficient recognition of non-CpG-island-related promoters.

PPPs are important *in silico* tools for guiding experimental biologists. Once the approximate putative regions for promoters have been detected using PPPs, reporter gene assays based on a series of deletion mutants can be used to further narrow down the DNA regions that play the most important role in the promoter activities. Conventional 5′ random amplification of cDNA ends (RACE) or other contemporary cap-selection methods, such as oligo-capping, on individual genes can also be used in experimental validation of the exact TSS positions. Wet-lab biologists have two principal tasks in which they need the assistance of PPPs: first, in a search for TSSs and alternative (fluctuating) TSSs in short segments of DNA[23]; second, in a search for unknown genes in targeted chromosomal segments or whole chromosomes/genomes. A large number of predictions very distant from real promoter sites can make laboratory tests infeasible; thus, a rigorous assessment of PPP performance is needed.

Although several PPPs have been developed, wet-lab biologists do not have clear information about the benefits and shortcomings of using a particular PPP. Does masking repeats enhance the performance of PPPs? The spectrum of promoters that can be detected and the costs (in terms of false-positive predictions) of making one true-positive prediction, vary for different PPPs. What is the real performance and the best way to use individual PPPs? Can the performance observed on a few chromosomes or specific data sets, generally used in illustrating the performance of PPPs, be extrapolated to the whole human genome? For many PPPs[5,7–9,14,15,24], authors report very good performance on chromosome 22, but it is the second most G+C-rich human chromosome and atypical. Taking the whole human genome as the reference is thus essential to reduce the bias that different genomic test sets introduce.

There is an urgent need to provide clearer answers to these questions, to set up standards for assessment of PPPs and to demonstrate how PPP performance can be enhanced. To address these problems, we performed a comparative promoter prediction analysis on the whole human genome.

We selected eight representative PPPs that can analyze large genomic sequences and report strand-specific TSS predictions (**Supplementary Methods** online and **Table 1**). Five of the programs compared (DragonPF, DragonGSF, McPromoter, NNPP2.2, Promoter2.0) use artificial neural networks (ANNs) as part of their design; four pro-

[1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, 119613 Singapore. [2]Human Genome Center, University of Tokyo, 4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan. Correspondence should be addressed to V.B.B. (bajicv@i2r.a-star.edu.sg).

Published online 4 November 2004; doi:10.1038/nbt1032

grams (DragonPF, DragonGSF, Eponine, FirstEF) use G+C content in their algorithms. Three programs (Eponine, NNPP2.2, Promoter2.0) explicitly use the TATA-box motif. Three programs (CpGProD, DragonGSF, FirstEF) use different versions of the concept of CpG islands. One program (CpGProD) uses rules based on statistics. The criteria for selecting PPPs for this analysis are given in Supplementary Methods online.

For reference, we used a large set of TSS locations based on full-length, oligo-capped cDNA sequences from the database of transcription start sites (DBTSS)[23,25]. This is the largest and most diverse human TSS data set based on experimental evidence used to date in the assessment of PPP performances. We also provide an analysis of the effects of masking repeats on promoter predictions in the human genome, establish standards for evaluation of PPPs, demonstrate how to improve the prediction performance of most of PPPs used with masking repeats and, similarly, how to combine some preexisting PPPs.

RESULTS

A total of 2,861,142,542 base pairs in the human genome sequence were examined using the eight selected PPPs. The accuracy of their predictions was assessed relative to the reference TSS set of 7,597 genes (Supplementary Methods online). The reference TSS locations were taken from DBTSS, making a total of 7,597 different TSSs. The selection of parameter settings for all PPPs is explained in Supplementary Methods online. Predictions of individual programs, which were no more than 1,000 nucleotides apart from the closest neighboring prediction, have been merged into a cluster. Each such cluster is represented by a new TSS prediction obtained as the average of all predictions within the cluster. When masking repeats is applied, we eliminate the clustered predictions of TSSs from the masked regions.

Performance on human genome data set. We applied several criteria to assess prediction results of eight PPPs using the whole human genome data set. These criteria included, sensitivity, positive predictive value (p.p.v.), number of true positives and false positives, average

Table 1 Details about the operation principles, basic technical data and best use of eight analyzed PPPs

Prediction program	Operating principle	Technical data	Best use	URL	Reference
CpGProD (CpG-island-promoter detection)	Statistical rule-based system. Detects only CpG-island-related promoters.	CpG-island boundaries and parameters. Suggests strand. We used midpoint of CpG island as predicted TSS.	No clustering of predictions and requires RepeatMasker[a]	http://pbil.univ-lyon1.fr/software/cpgprod_query.html	14
DragonPF (Dragon promoter finder version 1.5)	ANN, overlapping pentamer matrix models of promoters, exons and introns. Separate modules for promoters in G+C-rich and G+C-poor regions.	Content analysis of region around predicted TSS. Binding sites determined based on TRANSFAC[29] public database v.6. Analyzes putative start codons downstream of TSSs. Interactive[b,c].	Clustering of predictions and requires RepeatMasker[a]	http://research.i2r.a-star.edu.sg/promoter/promoter1_5/DPF.htm	5
DragonGSF (Dragon gene start finder version 1.0)	ANNs, concept of CpG island combined with predictions of DragonPF.	Content analysis of region around predicted TSS. Binding sites determined based on TRANSFAC[29] public database v.6. Interactive[b].	No clustering of predictions	http://research.i2r.a-star.edu.sg/promoter/dragonGSF1_0/genestart.htm	6,7
Eponine	Relevance Vector Machine based on a TATA-box motif in a G+C-rich domain.	Basic. Has a maximum sequence length of 1,024,000 nt[d]. No space is allowed in the header of the FASTA sequences. Download version can be obtained from: http://www.sanger.ac.uk/Users/td2/eponine/.	Clustering of predictions	http://servlet.sanger.ac.uk:8080/eponine/	9
FirstEF (first exon finder)	Quadratic discriminant analysis of promoters, first exons and first donor site. Uses the concept of CpG island.	Classifies predictions as CpG related or non-CpG related.	Clustering of prediction	http://rulai.cshl.org/tools/FirstEF/	8
McPromoter (McPromoter MM:II)	ANN, interpolated Markov model, different physical properties of promoter regions and statistical properties of promoters versus nonpromoters.	Gives graphical presentation of the evolution of scores. To run McPromoter, each sequence has to be in a separate file. The web version is restricted to 20 kbp sequences, whereas the download version is not. Download version very slow. Contact author Uwe Ohler (e-mail:ohler@mit.edu) to get binary executables.	Clustering of prediction and requires RepeatMasker[a]	http://genes.mit.edu/McPromoter.html	12,13
NNPP2.2 (neural network promoter prediction version 2.2.)	Three time-delay ANNs trained to recognize TATA box and Inr, as well as their mutual distance.	Basic. To get a local copy, contact author Martin G. Reese (e-mail: martinr@bdgp.lbl.gov)	Clustering of prediction and requires RepeatMasker[a]	http://www.fruitfly.org/seq_tools/promoter.html	10
Promoter2.0	ANN trained to recognize a combination of four TFBSs (TATA box, CCAAT box, GC box, Inr) and their mutual distances.	Classifies predictions as highly likely, medium likely, marginal[d]. Contact author Steen Knudsen (e-mail: steen@cbs.dtu.dk) to obtain a copy for noncommercial purposes.	Clustering of prediction and requires RepeatMasker[a]	http://www.cbs.dtu.dk/services/Promoter/	11

[a]RepeatMasker is available at http://pbil.univ-lyon1.fr/software/cpgprod.html. [b]Match program from Biobase,Germany and TRANSFAC database, a version of which is provided as a part of the TRANSPLORER Professional version 1.2 package of Biobase and can be downloaded from http://www.biobase.de/download/com/transplorer/demo/index.html. [c]The web version also allows reporting location of gaps in the input sequences, matching the selected segment around predicted TSS location with promoters from the Eukaryotic Promoter Database (EPD; http://www.epd.isb-sib.ch/) using BLAST, finding locations of candidate translation initiation sites (TISs) in genomic sequence downstream of the predicted TSSs and matching with BLAST the segment downstream of the predicted TISs with the 'nr' (all nonredundant GenBank CDS translations+RefSeq Proteins+PDB+SwissProt+PIR+PRF) database. [d]Analyzes only one strand. To analyze both strands you need to reverse-complement the original sequence before submitting it for analysis. In that case, program has to be run twice: first for the forward strand; second for the reverse strand.

Box 1 Criteria used to assess PPP performance quality

When one or more predictions fall in the region [–2000,+2000] relative to the reference TSS location, then the respective gene is counted as a true positive. When the known gene is missed by this count, it represents a false negative. Every prediction that falls on the annotated part of the gene in the segment [+2001, EndOfTheGene] is counted as a false positive, although we are aware that some of the predictions in this region could represent real promoters (see **Supplementary Methods** online). We did not consider other predictions for counting true-positive and false-positive scores. Counting true-positive predictions is either relaxed or the same as in the original reports on PPPs used. Counting false-positive predictions is in some cases different (see **Supplementary Methods** online) but, in our study, it is possible to compare performances as the same criteria are used for all PPPs. Results for different distance criteria and distributions of predictions around experimental TSSs are given in **Supplementary Table 3** online and **Supplementary Figure 1** online.

The cost of making one true-positive prediction, that is, how expensive is it to rely on a particular PPP, has a direct impact on the cost of the follow-up laboratory experiments for verifying predictions. To quantify these answers and express the prediction quality of individual programs we used sensitivity, p.p.v., correlation coefficient, average score measure (ASM) and true-positive cost.

Sensitivity is the proportion of correct predictions of TSSs relative to all experimental TSSs:

Sensitivity = true positive/(true positive + false negative).

p.p.v. (positive predictive value) is the proportion of correct predictions of TSSs out of all counted positive predictions:

p.p.v. = true positive/(true positives + false positives).

The Pearson correlation coefficient (CC) is defined as:

$$CC = (\text{true positive} \times \text{true positive} - \text{false positive} \times \text{false negative}) / ((\text{true positive} + \text{false positive})(\text{true positive} + \text{false negative})(\text{true negative} + \text{false positive})(\text{true negative} + \text{false negative}))^{1/2}$$

ASM[21] is the averaged rank position of the compared PPPs. It enables meaningful comparison of PPPs that achieve different sensitivity and p.p.v. scores. It uses 11 different performance indicators and calculates the average rank position of each PPP based on these indicators. The performance is better if the ASM score is smaller.

True-positive cost is the average amount of false-positive predictions required to achieve one true-positive prediction. The smaller its value, the less costly is the use of a particular PPP. This may prove useful in planning wet-lab experiments.

True-positive cost = false positive/true positive.

The most balanced behavior of a *P* is obtained if sensitivity and p.p.v. are approximately equal (the case of DragonGSF). If these two indicators have very different values going in favor of p.p.v. (the case of Eponine), then the system will make fewer false-positive predictions at the expense of true-positive predictions. Alternatively, if sensitivity is much higher than p.p.v., the program will make more true-positive predictions but will also produce many more false-positive predictions (the case of FirstEF). The gain or reduction in true-positive predictions has to be considered together with the increased cost of the follow-up experiments (that is, in terms of the increased number of false-positive) predictions, as well as the changed coverage of predicted promoters.

score measure, Pearson correlation coefficient and true-positive cost (for more information, see Box 1).

Table 2 shows that no PPP achieved simultaneously balanced sensitivity and p.p.v. >65%. The highest previously estimated[8] simultaneously achieved sensitivity and p.p.v. were ≥83%. However, this did not hold at the whole-genome level. Only one program[7] achieved simultaneously sensitivity and p.p.v. >62%, which means that it can recognize about two-thirds of all promoters while making something more than one false-positive prediction for every two true-positive predictions.

The prediction inaccuracy was most serious for non-CpG-island-related promoters. Essentially, no program could predict non-CpG-island-related promoters satisfactorily. A previous report[8] claimed that FirstEF predicts non-CpG-island-related promoters with p.p.v. = 60%, but we found that p.p.v. = 5.57% when evaluated using human genome and DBTSS data. As FirstEF predicts CpG-island-related promoters quite accurately (sensitivity = 77%; p.p.v. = 51%), obviously the non-CpG-island-related promoters have some distinct features that pose problems for accurate predictions.

For almost all PPPs, we observed discrepancies between reported sensitivity/p.p.v. values and those assessed using the whole human genome. Eponine achieves a sensitivity = 40.07% and p.p.v. = 66.97% on the human genome, whereas reported values[9] were sensitivity = 53.5% and p.p.v. = 72.73%. Likewise, the p.p.v. reported for FirstEF[8] on a data set derived from chromosomes 21 and 22 could not be confirmed when we used it to analyze these whole chromosomes (Supplementary Table 1 online) nor the human genome. According to our analysis, the performance of McPromoter has improved compared with its reported one[12] on chromosome 22, from sensitivity = 52.8% and p.p.v. = 62.6% to sensitivity = 57.92% and p.p.v. = 74.13%, but our criteria were different. Neural Network Promoter Prediction version 2.2 (NNPP2.2)[10] and Promoter2.0[11] produce predictions close to, or worse than, random guessing. On three whole chromosomes, DragonGSF[7] achieves a p.p.v. = 78%, but on the human genome, it only achieves a p.p.v. = 62.98%. For other interpretations of PPP performance, see Box 2.

Effect of repeats. Masking repeats in the human genome using RepeatMasker (Smit, A.F.A. & Green, P. RepeatMasker at http://repeatmasker.org/) substantially benefits the performance of several PPPs, the most evident improvement being demonstrated for DragonPF, McPromoter, NNPP2.2 and Promoter2.0.

A mild positive effect is observed for FirstEF, whereas essentially no benefits are achieved for DragonGSF and Eponine using RepeatMasker. Results are summarized in Table 2.

Combination of predictions. To find if proper combinations of PPPs can show beneficial effects on the overall prediction performance, we carried out an additional experiment and found that combining clustered predictions of PPPs can result in improved prediction quality. We analyzed combinations of two simple rules as applied to PPPs. PPPs were tagged as: 1, 2, 3, 4, 5, 6, 7, 8 corresponding to DragonGSF, DragonPF (expected sensitivity 0.65), Eponine, FirstEF, McPromoter (threshold –0.005), NNPP2.2 (threshold 0.99), Promoter2.0 and CpGProD (threshold 0.0), respectively (see Box 3 and Table 3).

Box 2 Interpretation of performance of PPPs

Results shown in **Table 2** were generated using the maximum allowed distance from the real TSS of 2,000 nucleotides. The true-positive cost column gives very useful information about the cost of making one true-positive prediction. For example, if you use FirstEF then for each true-positive prediction you make, you will have to make almost two false positives. If you restrict consideration only to CpG-island-related promoters, then the use of FirstEF is much more efficient and for every true-positive prediction you will make approximately one false-positive prediction.

However, if you intend to use FirstEF to search for non-CpG-island-related promoters, such a search will be very expensive and for each true-positive prediction you make, you will also make almost 17 false-positive predictions (if you do not use RepeatMasker). This information may prove crucial for wet-lab biologists planning laboratory experiments. The cheapest in this sense are CpGProD and Eponine, which will allow you to make one true-positive prediction by making less than 0.5 false-positive predictions. However, the price is that CpGProD will be able to predict only promoters contained within the CpG islands (37% of promoter population), whereas Eponine will predict about 40% of the promoters and only those characterized by the presence of the TATA-box motif in the G+C-rich promoters. The next best, with a slight increase in the cost, is DragonGSF where you will have to make approximately 0.6 false-positive predictions for every true-positive prediction. However, it will cover about 65% of all promoters, with the preference to the CpG-island-related ones. We also provide a rough estimate of the number of false-positive predictions in the human genome in the last column of **Table 2**. This estimate is calculated using the formula:

Estimated number of false positives on human genome = total clustered (column 6) × false positives/total of number of clustered (column 5) predictions used in counting true and false positives

Note that due to approximate character of this formula, the number of estimated false positives for FirstEF do not add up when summing estimates for CpG-island-related and non-CpG-island-related promoters.

Table 2 Prediction results on the whole human genome

Program	Sensitivity (%)	p.p.v. (%)	No. of true positives	No. of false positives	No. of predictions in genome after clustering	Avg. score measure	Rank by avg. score measure	Correlation coefficient	Rank by correlation coefficient	True-positive cost	Masking repeats is beneficial	Avg. distance in nt between clustered predictions	Estimated no. of false-positive predictions
CpGProD	47.26	51.84	3,590	3,335	34,067	5.2727	5–6	0.4950	6	0.9290	Built-in	90,704	16,295
(0.0)[a]	47.26	51.84	3,590	3,335	34,067	5.5455	6	0.4950	6	0.9290		90,704	16,295
CpGProD	37.09	69.79	2,818	1,220	16,215	4.0000	4	0.5088	5	0.4329	Built-in	190,566	4,889
(0.3)[a]	37.09	69.79	2,818	1,220	16,215	4.0909	4	0.5088	5	0.4329		190,566	4,889
DragonGSF	65.21	62.99	4,954	2,911	36,043	1.9091	1	0.6409	1	0.5876	No	158,763	13,313
	61.79	64.80	4,694	2,550	32,224	2.2727	1	0.6328	2	0.5432		95,892	11,326
DragonPF	56.05	21.30	4,258	15,729	151,031	8.5455	9	0.3455	9	3.6940	Yes	37,888	116,580
(50%)	53.85	32.23	4,091	8,604	84,975	7.9091	7	0.4166	9	2.1032		36,364	56,353
DragonPF	67.65	19.68	5,139	20,971	197,939	8.2727	7–8	0.3649	7	4.0808	Yes	28,909	155,747
(55%)	64.68	30.43	4,914	11,235	107,439	8.0909	9	0.4436	7	2.2863		28,760	73,094
DragonPF	80.93	15.05	6,148	34,708	318,421	8.2727	7–8	0.3490	8	5.6454	Yes	17,971	264,871
(65%)	77.28	24.62	5,871	17,972	165,098	8.0000	8	0.4362	8	3.0611		18,716	122,019
Eponine	40.08	66.98	3,045	1,501	22,569	3.9091	3	0.5181	4	0.4929	No	253,546	7,201
	39.91	67.33	3,032	1,471	21,963	4.0000	3	0.5184	4	0.4852		140,692	6,940
FirstEF	80.98	35.18	6,152	11,336	103,134	5.2727	5–6	0.5337	3	1.8427	Yes	55,484	66,484
	79.41	39.37	6,033	9,291	83,669	5.1818	5	0.5591	3	1.5400		36,931	50,455
FirstEF	4.38	5.61	333	5,608	40,398	10.4545	10	0.0496	14	16.8408	Yes	141,647	38,056
(CpG–)	4.12	6.25	313	4,697	31,620	10.8182	10	0.0507	14	15.0064		97,724	29,585
FirstEF	76.99	50.52	5,849	5,728	62,737	3.7273	2	0.6237	2	0.9793	Yes	91,211	30,893
(CpG+)	75.64	55.57	5,746	4,594	52,050	2.8182	2	0.6483	1	0.7995		59,366	23,020
NNPP2.2	92.77	2.78	7,048	246,792	2,043,202	11.2727	13	0.1605	10–11	35.0159	Yes	2,801	1,957,415
(0.90)[b]	77.12	4.08	5,859	137,800	1,073,716	11.4545	12	0.1773	10	23.5194		2,877	1,015,937
NNPP2.2	85.43	3.02	6,490	208,681	1,758,771	11.0000	11	0.1605	10–11	32.1542	Yes	3,257	1,682,267
(0.95)[b]	69.00	4.41	5,242	113,535	901,038	11.4545	13	0.1745	11	21.6587		3,429	850,460
NNPP2.2	56.50	4.27	4,292	96,335	834,712	11.0909	12	0.1552	12	22.4452	Yes	6,855	791,815
(0.99)[c]	43.32	6.11	3,291	50,594	413,907	11.2727	11	0.1627	12	15.3734		7,465	385,246
Promoter	57.23	3.27	4,348	128,789	1,255,812	12.0000	14	0.1367	13	29.6203	Yes	4,557	1,205,374
2.0[c]	44.07	4.90	3,348	65,048	603,914	12.0909	14	0.1469	13	19.4289		5,116	570,258
McPromoter	27.13	78.39	156	43	1,209						Yes		
(+0.005)[d]	26.96	87.08	155	23	956								
McPromoter	55.65	70.95	320	131	3,132						Yes		
(–0.005)[d]	54.96	79.20	316	83	2,294								

The upper number in a cell represents the value obtained without use of RepeatMasker. The lower number in a cell corresponds to the case when RepeatMasker is used. [a]CpGProD by design requires use of RepeatMasker. [b]Predictions do not seem useful in a genome-wide search, because pure random predictions at every 4,000 nt will produce sensitivity (Se) = 100%, whereas in the case of these systems many more predictions are made and Se < 100% is achieved. [c]Predictions are close to the random guessing which will produce Se = 100%, whereas these systems produce Se < 58%, making their use as single predictors highly questionable. [d]Results for McPromoter relate only to chromosomes 4, 21 and 22.

By examining all possible combinations of predictions of these programs on chromosomes 4, 21 and 22, we demonstrate that such combinations can improve the overall performance[26]. Previous studies[22,27–28] in promoter and gene predictions have reported the benefit of such approaches. Table 3 presents results for all combinations of PPPs that achieve a correlation coefficient greater than the highest correlation coefficient (0.7089) contained in Supplementary Table 2 online. Because McPromoter's results were not available for the whole human genome, the analysis of the whole genome would be deficient and thus we used only chromosomes 4, 21 and 22 for this analysis.

We have demonstrated that the use of two simple rules (see Box 3) improves accuracy. After examining all possible combinations, the results that outperform those from Supplementary Table 2 are presented in Table 3. The combination that resulted in the highest correlation coefficient (0.7250) was obtained by combining predictions of five programs.

Box 3 The combination of PPPs

To assess the effect on performance of combining PPPs, an experiment was performed in the following manner. We first scanned DNA in windows 2,000 nucleotides in length that were not overlapping. We considered whether the windows contained predictions of different PPPs and from particular PPP subgroups. When the conditions of the rules were satisfied for a window, we judged that window to contain predictions of the combination considered. Then, we represented the window by a new prediction selected as the midpoint of the window. This new prediction of TSSs was subjected to the same criteria used previously in counting true positives and false positives with the distance criterion [–2000,+2000]. The rules used were:

Rule 1: Window k contains predictions of at least s programs.
Rule 2: Window k contains predictions of at least p programs from the selected subgroup of programs.

Rule 1 considers all eight PPPs. Rule 2 considers only PPPs from selected subgroups. We analyzed all possible PPP subgroups, while varying s and p from 1 to 8, thus covering all possible choices. In the case when $s < p$, then Rule 1 can be ignored and the content of the window is controlled only by Rule 2. Results are shown in Table 3. We illustrate the interpretation of Table 3 content with data in row 7 that corresponds to a combination which has resulted in the highest Pearson correlation coefficient of 0.7250: each window must contain at least predictions from any three of the eight programs used; two of these programs have to be from the group (DragonGSF (1), Eponine (3), FirstEF (4), McPromoter (5)). The actual number of true-positive and false-positive predictions, as well as the resulting sensitivity and p.p.v. are provided in columns 3, 4, 10 and 11, respectively. It is interesting to note that CpGProD does not appear in these combinations, probably due to strong overlap of its predictions with those of FirstEF and DragonGSF.

Table 3 Results of combined promoter prediction on chromosomes 4, 21 and 22

Rule 1	Rule 2	No. of true positives	No. of false positives	Tags of programs used by Rule 2	Sensitivity	p.p.v.	Correlation coefficient	True-positive cost
3	2	358	81	1 3 4	0.6226	0.8155	0.7126	0.2263
1	2	364	85	1 3 4	0.6330	0.8107	0.7164	0.2335
3	1	374	100	1 3	0.6504	0.7890	0.7164	0.2674
1	3	379	106	1 2 3 4 5	0.6591	0.7814	0.7177	0.2797
3	2	382	112	1 4 5	0.6643	0.7733	0.7167	0.2932
3	2	393	118	1 3 4 5	0.6835	0.7691	0.7250	0.3003
1	2	397	129	1 3 4 5	0.6904	0.7548	0.7219	0.3249

DISCUSSION

Previous comparison analyses of PPPs were either limited to specifically selected data sets[16–19,21] or included only a few programs, focusing mainly on chromosome 22[5,7–9,14,15,24]. One analysis covered human chromosomes 4, 21 and 22, but included only three programs[7]. The present study includes eight PPPs, the whole human genome, the largest and most diverse experimental reference TSS data set used to date for PPP comparison, as well as an assessment of effects of masking repeats in promoter prediction on a large scale. It also provides a performance assessment of the current PPP technologies on the human genome level. We did not analyze recognition of alternative promoters because, currently, there is no large, sufficiently accurate and statistically diverse data set to be used as a reference. However, the present study has made progress in assessing PPPs using reliable information of a large number of human gene promoters.

The price of making one true-positive prediction directly influences the cost of the follow-up laboratory experiments when verification of predictions is required. The present study provides useful clues in this direction. If a single program were the preference, the most natural choice would be DragonGSF or Eponine because these two programs provide good coverage of promoters and are reasonably cheap; approximately for every two true-positive predictions, they give something more than one false-positive prediction.

If finding TSS predictions of the greatest accuracy is the preference, then CpGProD is the choice, with low sensitivity (37%) and restriction exclusively to CpG-island-related promoters. The second best in this regard is Eponine (sensitivity = 40%), which is restricted to G+C-rich promoters containing a TATA-box. CpGProD requires the use of RepeatMasker.

The best general purpose PPPs appear to be DragonGSF (which has a preference for CpG-island-related promoters) and FirstEF. DragonPF and FirstEF predict the most diverse sets of promoters. Very good performance is obtained by McPromoter, but it is very slow, which prevents its application in large-scale analyses. Neither Promoter2.0 nor NNPP2.2 seem to be a good choice even for the analysis of short DNA segments, particularly considering the cost of obtaining one true-positive prediction.

Our study demonstrates that, based on the current technology, it is not possible to extrapolate the performance of PPPs to the whole human genome, even from results from the three complete chromosomes[7] (Table 2 and Supplementary Table 2 online). A probable reason is that the structure of promoters on different chromosomes varies and suggests that this variation is not well covered by the algorithms used. So that users don't have unrealistic expectations, we therefore propose that any future performance assessment of PPPs should be supported by the whole genome and a repository of experimental data with a large number of promoters of different nature, such as DBTSS.

For most of the PPPs, the clustering of predictions is highly beneficial. We strongly advise users to apply this in the same manner as used in this study following suggestions given in Table 1. In addition, for several PPPs, the use of masking repeats is demonstrated as extremely

beneficial. We strongly recommend PPP users to adopt this practice on the basis of data presented in Tables 1 and 2.

Our data also emphasize that the combination of specific PPPs (Table 3) for the large-scale localization of the 5′ end of genes on the whole genomes is more beneficial than the use of single PPPs, although we were able to test this only on chromosomes 4, 21 and 22. The best use of individual PPPs and additional features provided to the users, are summarized in Table 1.

On the basis of our study, we have identified several aspects of promoter prediction software that require a great deal of improvement. First, selection of the right biological signals to be implemented in PPPs remains an open issue. The most efficient solutions currently available embark on the use of CpG islands, first exon properties, TATA-box and several other properties from the core promoter region, but do not possess sufficient sensitivity or the positional accuracy of TSS predictions. This suggests that, although these features are important, they are not sufficient for accurate TSS location by computational means.

Second, the bottleneck of the current technology is detection of non-CpG-island-related promoters. This is a considerable problem as this type of promoter represents a high proportion of all promoters found in the human genome[23].

Third, current PPPs are not able to precisely determine the TSS and thus cannot effectively detect alternative TSSs. Many PPPs recognize certain features of the 5′end of genes in a relatively efficient manner, but generally lack the ability to pinpoint the exact TSS location. To reach another level in PPP performance, we suggest the use of the distance criterion of [−20,+20] relative to experimental TSS accompanied with the average frequency of one prediction in 100,000 nucleotides.

Fourth, current PPPs with high sensitivity still show a high frequency of promoter prediction on the whole human genome. Clearly, high sensitivity and low frequency of promoter prediction is the goal.

Finally, Liu and States[22] have proposed that the combination of predictions of PPPs and transcript data should improve TSS prediction accuracy on a genome-wide level. Further work should analyze this in detail.

The present study provides useful hints for using individual PPPs under different circumstances. It shows how to greatly improve performance of several PPPs through the use of RepeatMasker, and to a lesser extent by combining predictions of PPPs. These analyses should expand the potential utility of the preexisting PPPs and provide a firm foundation for developing better PPPs. Additional technical details about individual PPPs are provided in Supplementary Notes online and in Table 4.

Table 4 Specifications for standalone PPPs analyzed in this study

Program	Operating system	License included	Price	Source code	Executable format	Max. sequence length	Sequence input format	Max. no. sequences in input file	Handling of gaps
CpgProD (CpG-island-promoter detection)	Pentium (Windows, Linux) SPARC (Solaris) SGI, Macintosh	Open source	Free download	Yes	Binary	No limit	FASTA, the sequence should be masked with RepeatMasker	No limit	Does not report gaps
DragonGSF (Dragongene start finder version 1.0)	Pentium (Windows, Linux) SPARC (Sun, Solaris)	Commercial, bundled in TRANSPLORER Professional v.1.2 (Biobase, Germany)	Academic $115	No	Binary	No limit	GenBank, EMBL, FASTA	No limit	Ignores gaps
DragonPF (Dragonpromoter finder version 1.5)	Pentium (Windows, Linux) SPARC (Sun, Solaris)	Commercial, bundled in TRANSPLORER Professional v.1.2 (Biobase, Germany)	Academic $115	No	Binary	No limit	GenBank, EMBL, FASTA	No limit	Ignores gaps
Eponine	All platforms that support JAVA	Open source (LGPL)	Free	No	JAR	1 Mb per sequence	FASTA, plain	No limit	Does not report gaps
McPromoter (McPromoter MM:II)	Pentium (Linux)	Contact author Uwe Ohler (e-mail: ohler@mit.edu)	Contact author	No	Binary	No limit	FASTA	One sequence	Long stretches of ambiguous symbols (>50 bp) skipped; short stretches replaced randomly
FirstEF (first exon finder)	Pentium (Linux), SPARC (Sun,Solaris) MIPS (Silicon Graphics, IRIS), Alpha (OSF1)	Required	Free for nonprofit users	No	Perl and Binary	No limit	FASTA	No limit	Does not report gaps
NNPP (neural network promoter prediction version 2.2)	Pentium (Linux), SPARC (Sun,Solaris) Alpha (OSF1)	Contact author Martin G Reese (e-mail: martinr @bdgp.lbl.gov)	Free for nonprofit users	No	Perl and Binary	No limit	FASTA	No limit. Produces two files if analyzes both strands.	Does not report gaps
Promoter2.0	Pentium (Linux), MIPS (Silicon Graphics, IRIS) SPARC (Sun, Solaris), Alpha (OSF1), Power (AIX)	Contact author Steen Knudson (e-mail: steen@cbs.dtu.dk)	Free for nonprofit users	No	Binary, needs Unix tools: gawk, echo, nawk, uname	No limit	FASTA	No limit. To analyze reverse strands, users must reverse complement sequence and redo analysis	All symbols that are not A,C,G,T converted to X before processing. Does not report gaps

EMBL, European Molecular Biology Laboratory.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We are grateful to Riu Yamashita and Kenta Nakai for assisting in constructing and maintaining DBTSS.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details).

Published online at http://www.nature.com/naturebiotechnology/

1. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Weinzierl, R.O.J. Mechanisms of Gene Expression: Structure, Function, and Evolution of the Basal Transcriptional Machinery (Imperial College Press, London, 1999).
4. Pedersen, A.G., Baldi, P., Chauvin, Y. & Brunak, S. The biology of eukaryotic promoter prediction—a review. *Comput. Chem.* **23**, 191–207 (1999).
5. Bajic, V.B. *et al.* Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates. *J. Mol. Graph. Model.* **21**, 323–332 (2003).
6. Bajic, V.B. & Seah, S.H. Dragon Gene Start Finder identifies approximate locations of the 5′ ends of genes. *Nucleic Acids Res.* **31**, 3560–3563 (2003).
7. Bajic, V.B. & Seah, S.H. Dragon Gene Start Finder: an advanced system for finding approximate locations of the start of gene transcriptional units. *Genome Res.* **13**, 1923–1929 (2003).
8. Davuluri, R.V., Grosse, I. & Zhang, M.Q. Computational identification of promoters and first exons in the human genome. *Nat. Genet.* **29**, 412–417 (2001).
9. Down, T.A. & Hubbard, T.J. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12**, 458–461 (2002).
10. Reese, M.G. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput. Chem.* **26**, 51–56 (2001).
11. Knudsen, S. Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics* **15**, 356–361 (1999).
12. Ohler, U., Liao, G.C., Niemann, H. & Rubin, G.M. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3**(12), RESEARCH0087. Epub 2002 Dec 20 (2002).
13. Ohler, U., Stemmer, G., Harbeck, S. & Niemann, H. Stochastic segment models of eukaryotic promoter regions. *Proc. Pac. Symp. Biocomput.* **5**, 380–391 (2000).
14. Ponger, L. & Mouchiroud, D. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* **18**, 631–633 (2002).
15. Hannenhalli, S. & Levy, S. Promoter prediction in the human genome. *Bioinformatics* **17**, S90–S96 (2001).
16. Ioshikhes, I.P. & Zhang, M.Q. Large-scale human promoter mapping using CpG islands. *Nat. Genet.* **26**, 61–63 (2000).
17. Scherf, M., Klingenhoff, A. & Werner, T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* **297**, 599–606 (2000).
18. Solovyev, V.V. & Shahmuradov, I.A. PromH: Promoters identification using orthologous genomic sequences. *Nucleic Acids Res.* **31**, 3540–3545 (2003).
19. Fickett, J.W. & Hatzigeorgiou, A.G. Eukaryotic promoter recognition. *Genome Res.* **7**, 861–878 (1997).
20. Prestridge, D.S. Computer software for eukaryotic promoter analysis. *Methods Mol. Biol.* **130**, 265–295 (2000).
21. Bajic, V.B. Comparing the success of different prediction software in sequence analysis: a review. *Brief. Bioinform.* **1**, 214–228 (2000).
22. Liu, R. & States, D. J. Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome Res.* **12**, 462–469 (2002).
23. Suzuki, Y. *et al.* Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.* **2**, 388–393 (2001).
24. Scherf, M. *et al.* First pass annotation of promoters on human chromosome 22. *Genome Res.* **11**, 333–340 (2001).
25. Suzuki, Y. *et al.* DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.* **30**, 328–331 (2002).
26. Ripley, B.D. *Pattern Recognition and Neural Networks* (Cambridge University Press, Cambridge, UK, 1996).
27. Murakami, K. & Takagi, T. Gene recognition by combination of several gene-finding programs. *Bioinformatics* **14**, 665–675 (1998).
28. Rogic, S., Ouellette, B.F. & Mackworth, A.K. Improving gene recognition accuracy by combining predictions from two gene-finding programs. *Bioinformatics* **18**, 1034–1045 (2002).
29. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**(1), 374–8 (2003).

In Silico Biology 4, 0036 (2004); ©2004, Bioinformation Systems e.V.

Large-scale collection and characterization of promoters of human and mouse genes

Yutaka Suzuki[1]*, Riu Yamashita[1], Matsuyuki Shirota[1], Yuta Sakakibara[1,2], Joe Chiba[2], Junko Mizushima-Sugano[1], Alexander E. Kel[3], Takahiro Arakawa[4], Piero Carninci[4,5], Jun Kawai[4,5], Yoshihide Hayashizaki[4,5], Toshihisa Takagi[1], Kenta Nakai[1] and Sumio Sugano[1]

[1] Human Genome Center, The Institute of Medical Science, The University of Tokyo: 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan;

[2] Department of Biological Science and Technology, Science University of Tokyo, 2641 Yamazaki, Noda-shi, Chiba, 278-8510, Japan;

[3] BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany;

[4] Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Main Campus, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan;

[5] Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

* Corresponding author; email: ysuzuki@ims.u-tokyo.ac.jp
Phone: +81-3-5449 5343; Fax: +81-3-5449 5416

Edited by E. Wingender; received June 10, 2004; revised and accepted July 20, 2004; published July 23, 2004

Abstract

We report the generation and initial characterization of a large-scale collection of sequences of putative promoter regions (PPRs) of human and mouse genes. Based on our unique collection of 400,225 and 580,209 human and mouse full-length cDNAs, we determined exact transcriptional start sites (TSSs). Using positional information of the TSSs, we could retrieve adjacent sequences as PPRs for 8,793 and 6,875 human and mouse genes, respectively. The positions of the PPRs were 4 kb upstream to previously reported 5'-ends of cDNAs on average, demonstrating that full-length cDNA information is indispensable for this purpose. Among those PPRs supported by experimentally validated TSSs, 3,324 could be paired as mutually homologous genes between human and mouse and were used for the comprehensive comparative studies. The sequence identities in the proximal regions of the TSSs were 45% on average, and 22,794 putative transcription factor binding sites that are conserved between human and mouse were identified. The data resource created in the present work and the results of the sequences' initial characterization should lay the firm foundation for deciphering the transcriptional modulations of human genes. All the data were deposited and made available through a database for comparative studies, DBTSS.

Key words: full-length cDNA, promoter, comparative genomics, transcriptional start sites

Introduction

In order to understand the transcriptional network of human genes, it is essential to characterize their

regulatory regions, which include regions called promoters. To this end, one of the challenges confronted by both experimental and bioinformatics researchers has been to decode what kind of functional sequence elements reside in which parts of the promoters and how they serve as modulators of transcription. A large number of regulatory proteins, which are collectively called transcription factors (TFs), have been identified and their sequence-specific binding to promoter elements has been shown to play the central role in regulation [Mitchell and Tjian, 1989; Novina and Roy, 1996]. As many of the TF binding sites are short (6-12 bp) and their consensus sequences are often degenerated, it was an intricate problem to discriminate the genuine TF binding sites, which have biological significance *in vivo*, from insignificant sequences, which occur randomly and very frequently in the large volume of human genomic sequences [Fickett and Wasserman, 2000].

Comparative study of human and other organisms' sequences, namely comparative genomics, is a powerful method to extract biologically meaningful information as to which parts of the genomic sequences are likely to have functional relevance. It is expected that the functionally important regions, such as exons and promoter elements, are evolutionally conserved and could be discriminated from non-conserved ones, which are supposed to be subject to fewer functional constraints [Hardison, 2000; Boguski, 2002]. The almost-complete sequencing of both human and mouse genomes [Lander *et al.*, 2001; Venter *et al.*, 2001; Waterston *et al.*, 2002] provided us with the basic material with which to initiate large-scale comparative studies of the promoters. If the positional information of the transcriptional start sites of mRNAs (TSSs) were available, the promoter sequences could be identified by computational mapping of the TSS onto the genomic sequences, since in most cases, the promoters are located just proximal to or overlapping with the TSS. Once promoter sequences were retrieved, they could be subjected to further analyses for the presence of particular TF binding sites.

Transcriptional start sites correspond to the 5'-ends of the full-length cDNAs. Therefore, obtaining the TSS information is equivalent to obtaining the 5'-end information of the full-length cDNAs. However, it was often difficult to obtain the 5'-end sequences of the full-length cDNAs from public databases. For most of the cDNAs registered there, the exact TSSs had not been determined either by S1 mapping, primer extension or 5'RACE and their authentic 5'-ends remain uncharacterized. Since these cDNAs cannot be regarded as full-length cDNAs in a strict sense, it would be inappropriate to use their 5'-end information for promoter retrieval. Indeed, even for the cDNAs registered in one of the most reliable cDNA databases, RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/)[Suzuki *et al.*, 2002; Pruitt *et al.*, 2003], about half of the 5'-ends of RefSeq sequences should be extended towards the 5'-end according to our previous observation [Suzuki *et al.*, 2002].

We have developed a method of constructing a full-length enriched cDNA library using a cap selection method, "oligo-capping", and have been collecting the full-length cDNAs [Suzuki and Sugano, 2003]. Based on the human genomic DNA and full-length cDNA data, we recently reported identification and computational characterization of human gene promoters on a large-scale [Suzuki *et al.*, 2001]. The 5'-end one-pass sequences of 217,402 of the full-length cDNAs were mapped onto the human genomic sequences and adjacent promoter sequences were identified [Suzuki *et al.*, 2002].

In the present study, we expanded the human full-length cDNA as well as applying a similar strategy to mouse data. For mouse cDNA data, we used full-length cDNA sequences, which were derived from cDNA libraries constructed by another cap selection method, the "cap trapper" method [Carninci and Hayashizaki, 1999; Kawai *et al.*, 2001; Okazaki *et al.*, 2002]. It is estimated that more than 80% of the cDNA clones isolated from the cDNA libraries constructed either by the oligo-capping method or by the cap-trapper method should represent full-length cDNAs [Carninci and Hayashizaki, 1999; Suzuki and Sugano, 2003]. Here we report generation and initial characterization of a large-scale dataset of promoter sequences and construction of a database, DBTSS, for comparative studies of promoters of human and mouse genes.

Materials and methods

Processing of the full-length cDNA sequences and Mapping of the TSSs on the Genomic Sequences

For human TSSs, each sequence produced by the oligo-capping method was first processed to trim its vector site and its low quality parts. We also used FANTOM 5'-end sequences from Genbank (acc. No. BB561685-BB667065, BB838020-BB873800) and our full-length cDNA data to determine mouse TSSs. They were compared with human or mouse RefSeq using BLASTN. If a sequence alignment displayed an identity greated than 95% and a e-value less than 1.0e-100, it was regarded as identical to the RefSeq sequence. Sequences that had multiple hits in RefSeq were discarded. Then, the exact positions of the TSSs on the human (build 31) or mouse (mm2) genomic sequences were determined (http://genome.ucsc.edu/downloads.html), using the sim4 program (http://pbil.univ-lyon1.fr/sim4.html) [Florea *et al.*, 1998]. In order to identify precise TSS information, we removed all the entries that were not mapped on the human genome sequence from their first base. Where fluctuating TSSs were observed, the most frequently used TSSs were defined as representatives. If the "most frequent TSSs" were multiple, we defined the median of them as a representative.

Generation of the correlation table between human and mouse counterparts

In order to generate the relational table between human and mouse counterparts, human and mouse representative transcripts were compared with each other using BLAST with a cut-off e-value of 1.0e-100. For the datasets of representative human and mouse transcripts, RefSeq and RTPS (representative transcripts and protein sequences from the FANTOM project) were used, respectively. The generated pairs were further sorted by having at least one Ref-full or RTPS per pair. Where homology searches gave ambiguous results (with mutually multiple hits), they were excluded from the table, so that the obtained relational table consisted only of the gene pairs of reciprocal best match homologs.

Sequence comparison between promoter pairs of human and mouse genes

Sequences of the promoters were compared between human and mouse homologues. Sequences of the -1000 to +200 bp relative to the TSSs were used and sequence identity was calculated. For the sequence alignment, LALIGN was run with the default parameters. The sequence identities were averaged for the 1200 bp regions. The identity counts of the regions where no alignment was generated using LALIGN were scored as 0.

Search for putative TF binding sites

Putative TF binding sites were searched by using the position weight matrices (PWMs) from TRANSFAC[a] Professional 7.1. Searching was done by Match [Kel *et al.*, 2003], a weight matrix-based tool for searching putative transcription factor binding sites in DNA sequences. Match is closely interconnected and distributed together with the TRANSFAC database. Match applies two cut-offs for the score values of the matrix matches: core cut-off for the 5 core nucleotides and matrix similarity cut-off for the whole match. Match allows usage of different cut-offs for every matrix. We used several sets of cut-offs (so called matrix profiles) provided by TRANSFAC: 1) minFP, to minimize the false positive (over-prediction error) rate, 2) minSUM, to minimize the sum of both errors. For analyzing putative AP-1, NF-κB and NF-AT sites, the PWMs of V$AP1_01 for AP-1, V$NFKAPPAB_01 for NF-κB, and V$NFAT_Q6 for NF-AT in TRANSFAC were used with the core and matrix similarity cut-offs of (0.8, 0.93), (0.8, 0.92), (0.8, 0.97), respectively.

Availability of the Database

From the download site at DBTSS, major resources used for the database construction are available by FTP, including the flat files of the human/mouse one-pass sequences with Genbank accession numbers, retrieved promoter sequences and correlation tables of the promoters. The DBTSS and the data it displays are freely available for academic, nonprofit, and personal use.

Results

Collection and clustering of the human and mouse full-length cDNAs

In total, our database, DBTSS, now records 400,225 full-length cDNA sequences, including an additional 182,823 sequences compared to the previous version (Genbank accession numbers are BP192706-BP383670). This additional data should have improved not only the coverage of genes represented in DBTSS but also the overall reliability of the identified TSSs, since the probability should have greatly increased for a particular TSS being a correctly identified TSS when the redundancy of the supporting full-length cDNAs increased. These cDNAs are isolated from 137 kinds of full-length cDNA libraries, all of which are constructed using the "oligo-capping" method (further details on the library information including the completeness (whether they are full-length) of each of the libraries are presented at http://dbtss.hgc.jp/ in the "Statistics" section).

The "oligo-capped" cDNA sequences were first searched against RefSeqs (as of November 14, 2002) using BLAST [Altschul *et al.*, 1990]. When hits were found, the 5'-ends of them were compared with those of RefSeqs. When the RefSeq was truncated, its 5'-end sequence was complemented to obtain a putative representative full-length cDNA, which we refer to as a Ref-full. As summarized in Table 1, we could generate 9,270 Ref-fulls based on RefSeqs and our full-length cDNA sequences (for further details, see Material and methods). Sequence data obtained by 6,042 Ref-fulls were extended towards the 5'-ends by 71.6 bp on average (see Table 2 and Figure 1A for the distribution of the differences). Some of the extended parts overlapped with the open reading frames (ORFs), thus, were also useful to revise the currently truncated N-terminal sequences of the deduced amino acid sequences in RefSeq. In some of the sequences, upstream ATGs and ORFs are embedded (for further discussion on this issue, refer to our recent papers [Suzuki *et al.*, 2000; Yamashita *et al.*, 2003].

Table 1: Statistics of the collected promoter data.

	Human	Mouse
RefSeq and Ref-full		
Ref-full (promoter retrieval successful)	8,793 (48%)	6,875 (53%)
Ref-full (total)	9,270 (51%)	7,524 (58%)
Ref-full that extended RefSeq	*6,042 (33%)*	*5,018 (38%)*
Ref-full that did not extend RefSeq	*3,228 (18%)*	*2,506 (19%)*
RefSeq that are not covered by Ref-full	8,944 (49%)	5,557 (42%)
RefSeq (total)	18,214 (100%)	13,081 (100%)
One-pass sequences and genome mapping		
Hit to RefSeq (genome mapping successful)	190,964 (48%)	195,446 (34%)
Hit to RefSeq (genome mapping ambiguous)	36,267 (9%)	36,624 (6%)
No hit to RefSeq	172,994 (43%)	348,139 (60%)
One-pass (total)	400,225 (100%)	580,209 (100%)

Statistics of the number of promoters, the redundancies of the supporting full-length cDNAs and the differences between the public data are shown.

Table 2: Statistics of full-length cDNA sequences used for the retrieval.

	Number of registered genes (average redundancy)	Average length difference from RefSeq (mRNA level)	Average length difference from RefSeq (genomic level)
Human	8,793 (21.7)	71.6	4,396
Mouse	6,875 (28.4)	76.0	4,027
Human/mouse pairs	3,324 (25.2/38.0)	63.3/68.8	3,998/3,380

Statistics of the full-length cDNAs used for the database construction is shown.

The remaining 3,228 cDNAs in which the 5'-ends of the Ref-fulls were almost consistent with the RefSeq 5'-ends and were used to confirm that the RefSeqs had originally represented the full-length cDNAs. In the present study, we excluded the cDNAs that did not correspond to RefSeqs, as the one-pass sequences without the RefSeq supports are singletons in many cases. Among them a number of spurious cDNAs, such as cloning artifacts and other kinds of aberrant transcripts, might be included. Besides, our recent analyses suggested that sporadic transcription from non-genic region regions are inherent in human and mouse genomes (Sakakibara *et al.*, in preparation). Since it was a concern that incorporating this part of the data could make the dataset confusing, we did not include it to the current dataset.

As for the mouse genes, full-length cDNAs were obtained from the FANTOM database (http://fantom.gsc.riken.go.jp/) and processed by a similar procedure as that used for the human cDNAs. Starting from 580,209 one-pass sequences of the 5'-ends of the full-length cDNAs, 7,524 Ref-fulls were obtained of which 5,018 extended pre-existing RefSeq sequences by 76.0 bp on average (Figure 1A).

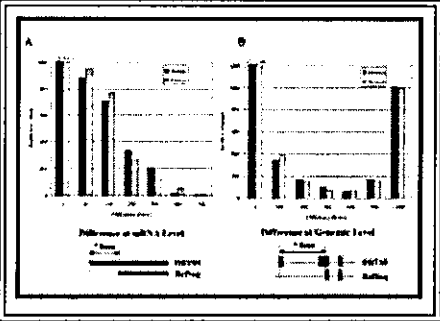


Figure 1: Comparison between Ref-fulls and RefSeqs. The distributions of the differences between Ref-fulls and RefSeqs are presented, when compared at the mRNA level (A) and genomic level (B). Black and gray bars represent the cases for human and mouse genes, respectively.

Retrieval of promoter sequences based on Ref-fulls

The one-pass sequences of 190,964 and 195,446 human and mouse cDNAs corresponding to 8,793 and 6,875 Ref-fulls were precisely mapped onto the human and mouse genomes, using strict criteria described in Materials and Methods. Exact positional information of the TSSs could be determined on each of the genomes (Table 1). The average redundancy, that is, the number of full-length cDNAs supporting TSS of each of the genes, was 21.7 and 28.4, respectively. Although 1,980 human TSSs and 691 mouse TSSs were determined by single full-length cDNA data (singletons), the others were supported by multiple full-length cDNA data (Figure 2). As the average frequency of the full-length cDNAs

in the cDNA libraries is more than 80%, the probability should be low that the truncated erroneous full-length cDNAs happened to be mapped closely so as to lead to misidentification of the promoters.

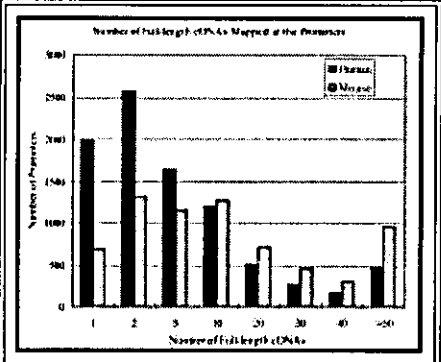


Figure 2: Distribution of the numbers of mapped full-length cDNAs. The distribution of the numbers of mapped TSSs is shown for human and mouse genes by black and gray bars, respectively.

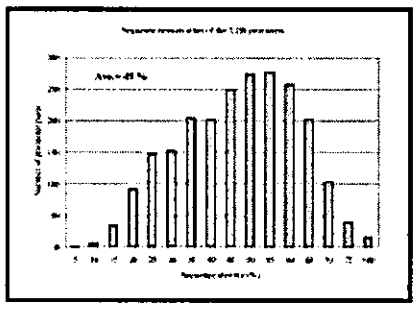
The average distances between the 5'-ends of the RefSeqs and the Ref-fulls calculated at the genomic level were 4,396 bp and 4,027 bp for human and mouse, respectively (Figure 1B). In this dataset, 62% and 56% of the mapped TSSs of the human and mouse genes are located in the CpG islands, respectively. As large introns were observed just downstream of the exact TSSs in many cases, the distances between the 5'-ends of RefSeqs and Ref-fulls were much greater than those calculated at the mRNA level. In these cases it was impossible to identify real promoters based solely on the RefSeqs, even if the differences calculated at the mRNA level were small.

Relating the promoters of human and mouse gene counterparts

In order to compare the retrieved human and mouse PPR sequences with each other, we wished to relate the human genes to the mouse gene counterparts. We started from RefSeqs and the RTPS dataset, which are the representative sets from mouse created in FANTOM mouse full-length cDNA annotation meetings (for further details see the reference Okazaki *et al.*, 2002). We compared their sequences both at the nucleotide and amino acid level so that all of the related gene pairs should be 1:1 reciprocal best hit homologs. In total, we could correlate 8,185 human and mouse genes in total.

Using the obtained relational table, we could define 3,324 human and mouse gene pairs among our PPR dataset, supported by 83,708 (redundancy: 25.2) human and 126,326 (redundancy: 38.0) mouse full-length cDNA data. Of these, 2,256 promoter pairs were supported by more than three full-length cDNA sequence data of both human and mouse (in total more than six cDNAs were mapped). The PPR pairs were aligned with each other using a sequence alignment program, LALIGN [Huang *et al.*, 1992]. On average, the overall sequence conservation between the promoter pairs was 45%, when evaluated in the regions from -1000 to +200 (TSS was designated as 0) of the 2,256 dataset. The average length of the aligned upstream sequences was 510 bp. However, the size and patterns of the sequence alignment were quite different between promoters (for further details, refer to Suzuki *et al.* in preparation). Figure 3 represents the extent to which the sequences were conserved between PPR pairs.

Figure 3: Distribution of the sequence conservation between human and mouse promoters Sequence alignments were performed using LALIGN with the default parameters. The sequence identity was evaluated as the number of aligned nucleotides in the regions of -1000 to +200 (TSS: 0).



Construction of DBTSS for comparative studies

All of the created data is made publicly available through our newly developed database, DBTSS. The schematic of the user interface is illustrated in Figure 4 and details of the database description are presented in Supplementary Information.

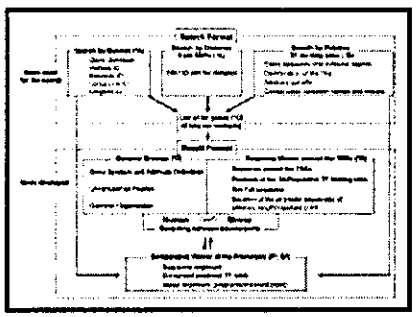


Figure 4: Schematic of the user interface of DBTSS. The boxes that are marked with asterisks (A~G*) correspond to the respective forms illustrated in Supplementary Figures S1 and S2.

It should be noted that this version of DBTSS has implemented the search for the PPRs by putative TF-binding sites that are conserved between human and mouse genes. For this search, arbitrary combinations/positions of the putative TF-binding sites can be set. For example, it is possible to search "TATA-plus PPRs containing NF-κB binding site(s) and either NF-AT site(s) or AP-1 site(s), all of which are conserved between human and mouse within 500 bp of the TSSs" (this combination of the TFs is frequently observed in the promoters responsible for inflammatory responses) [Baeuerle and Baltimore, 1996; Ho and Glimcher, 2002; Praz *et al.*, 2002].

Practically, when the PPRs were searched using TRANSFAC [Matys *et al.*, 2003] with strict parameters (minFP64.prf; see also Supplementary Information), 183,712 and 170,926 hits were detected from human and mouse promoters, respectively, in total. However, we were concerned that these matches might include a lot of false positive hits. To decrease the number of false predictions we used the comparative PPR data following the assumption that among the detected putative TF binding sites evolutionary conserved ones may have functionally relevance. Consistently, confidential data elucidated that most functionally relevant TF binding sites are conserved throughout evolution (between 64-75%; Hannenhalli and Levy, 2002; Sauer and Wingender, in preparation). Using the promoter alignment data as a filter for selecting the conserved TF binding sites, DBTSS could pick up 22,794 putative TF binding sites in human promoters which are conserved between human and mouse. By doing this, it is possible to select the TF binding sites that should have first priority for experimental validation. Results of the search for representative TFs are presented in Supplementary Information Table.

We temporarily focused on evolutionarily conserved TF binding sites. Actually, we observed that about

85% of the predicted conserved-TF binding sites are located in the conserved regions of the promoters. However, this does not imply that non-conserved predicted TF binding sites always should have no functional relevance. Some of the TF binding sites which are not conserved between human and mouse might play roles in a species-specific manner. This always should be kept in mind whenever this kind of search is attempted.

The so-called "phylogenetic foot printing" approach is the most powerful when the combination(s) of the TF binding sites is taken into account as well. For example, when promoters containing putative binding sites of NF-κB were searched using the standard cut-offs (for further details see Material and Methods), 1,491 and 983 sites were detected in the human and mouse promoters, respectively. However, when the hits were restricted to the conserved ones, the number of hits decreased to 36. When a similar search was performed for promoters containing putative NF-AT or AP-1 binding sites, the numbers of hits were 7,368, 5,545 and 652 for human, mouse and conserved, respectively. When searching for promoters containing both of the conserved NF-κB and NF-AT/AP-1, the number of hits was 22. These should be primary targets for initiating the experimental characterization of promoters as to whether they really respond to an inflammatory stimulus [Kel *et al.*, 2003].

Discussion

In this paper, we described the large-scale collection of initial comprehensive comparative analyses of promoters of human and mouse genes. The dataset generated in this study as well as the newly developed database are unique, based on the experimentally identified TSSs. Although there are a number of databases which enable genome-wide comparison between human and mouse genes, such as HGB at UCSC, Ensembl at EBI, Map Viewer at NCBI [Clamp *et al.*, 2003; Karolchik *et al.*, 2003; Wheeler *et al.*, 2003], they are mainly focused on the global alignments of the genomes, and are intended for finding exonic regions rather than for the characterization of promoters. To our knowledge, rVISTA) and GALA are rare exceptions, mainly focusing on promoter comparison [Loots *et al.*, 2002; Giardine *et al.*, 2003; Ureta-Vidal *et al.*, 2003]. However, in all of these pre-existing databases, most of the "5'-flanking regions" are not defined by experimentally determined TSSs; therefore, it has been difficult to distinguish which part should correspond to exons and which should be regarded as promoters, even if conserved regions were identified. Actually previous observations reported that the average sequence identity of the "upstream regions" of human and mouse genes was approximately 70-75% [Waterston *et al.*, 2002], which is apparently higher than our calculation (45%; Figure 3). This may have been caused by the fact that they used upstream 200 bp regions. The degree of the sequence identity might be lower at more upstream regions. Consistently, a previous report indicated that frequency of the alignable sequences becomes lower relatively rapidly in the upstream regions [Jareborg *et al.*, 1999]. Since we used the entire -1000 bp to +200 bp regions in the present study, the calculated sequence identity might be lower than the previous result. Further extensive analyses of the sequence alignments generated from various global/local alignment programs should reveal how the sequences in the upstream regions of the TSSs are conserved between human and mouse.

Taking advantage of the large-scale collection of the full-length cDNAs, we could focus on the limited regions of the genomic sequences for the analysis of promoters. Also, we could take into account the positions of the predicted TF-sites relative to the TSS for the search of the analysis of the putative TF-binding sites. Recent reports described that the TF-sites predicted kilo bases apart from the TSS should have less probability of having biological consequences [Liu *et al.*, 2003]. In order to expedite the experimental analyses of the promoters by minimizing the false positives, the target regions that should be used for the primary searches have to be defined for each of the TFs. Implementing this feature, DBTSS should be the first database which makes the most use of the promoter data for the practical requirements of experimental biologists.