The Origins of Bi-directional Promoters

-Computational analyses of intergenic distance in the human genome-

*Daiya Takai and Peter A. Jones.

Department of Biochemistry and Molecular Biology,
USC/Norris Comprehensive Cancer Center,
Keck School of Medicine of the University of Southern California,
1441 Eastlake Avenue, Los Angeles, California 90033, USA
*Corresponding author. (dtakai-ind@umin.u-tokyo.ac.jp)

Present address: Department of Respiratory medicine,
Graduate school of medicine and faculty of medicine, the University of Tokyo,
7-3-1, Hongo, Bunkyoku, Tokyo, Japan.
Telephone: 81-3-3815-5411, FAX: 81-3-5684-3987

Keywords: Promoter, human genome, intergenic distance, bi-directional, head-to-head, overlapping promoters.
We have analyzed intergenic distances and searched for the presence of bi-directional genes using the complete sequences and mapping information of human chromosomes 20, 21 and 22 which contain 2,122 known and predicted genes. Intergenic distances between genes with divergent transcripts were distributed in a biphasic manner with a strong peak of 25 kb and a weak peak of 0.3 kb between the divergent transcripts suggesting that the genes might share a common promoter. The weak peak was not observed at the transcriptional ends of genes. Seventy three percent (55/75 pairs of genes, from a total of 150 genes) of these divergent transcripts located within 1kb of one another were CpG islands. Expression of the divergent transcript genes was not concordant in various human tissues suggesting that they were independently regulated. Analyses of the frequency of occurrence of interspersed repeats in the intergenic sequences suggested that these repeats are strongly excluded from the regions of transcriptional starts. This exclusion might be responsible for the existence of these divergent transcript.
In a recent report, Adachi and Lieber (Adachi and Lieber, 2002) suggested that 20% of human genes were located within 1kb of one another and that this frequency was higher in DNA repair genes. To evaluate this phenomena in larger gene subsets, we have analyzed intergenic distances using the complete sequences of human chromosomes 20 (Deloukas et al., 2001), 21 (Hattori et al., 2000) and 22 (Dunham et al., 1999) in conjunction with gene mapping information from the GenBank database. We located the 5’ region of all known and predicted genes on these chromosomes and categorized them into four categories; genes with divergent transcripts, two genes sharing the same strand, genes overlapping with other genes and genes located at the ends of contigs, which are the largest units of continuous sequences available or are adjacent to a sequence gap and cannot be accurately located (Figure 1a). The distributions of these categories were calculated and the proportions indicated in Figure 1b, which also shows the distances between transcriptional start sites of divergent genes. Approximately half of the genes have divergent transcripts whereas the other half of genes are located on the same strand. This proportion is in accordance with the assumption that genes are located randomly in the genome so that for any given gene the direction of a gene located 5’ in the same direction or the opposite direction is 50%.

Lavia et al. (Lavia et al., 1987) and Adachi and Lieber (Adachi and Lieber, 2002) described an association between divergent transcripts and CpG islands, therefore the frequency of CpG islands in these divergent transcripts was also analyzed. Applying the CpG island searcher (Takai and Jones, 2003) at the default parameters of %GC ≥ 55, ObsCpG/ExpCpG ≥ 0.65 and Length ≥ 500 bp, CpG islands were found in 73% of 55 out of a total of 75 pairs of genes which were located within 1kb of one another. This frequency of CpG islands in divergent transcripts appears, at first sight, to be significantly higher than a previous analysis (Takai and Jones, 2002) (46%, 161 CpG islands in 350 genes). However, the value is as expected if CpG islands are not preferentially located in bi-directional promoters since each divergent transcript consists of two genes, and the expected frequency
of CpG islands in the middle of bi-directional transcripts is calculated as $1-(1-0.46)x(1-0.46)=0.71$ so that the observed frequency fits quite well with this expected value. Therefore in contrast to the findings of Adachi and Lieber (Adachi and Lieber, 2002) we do not find that CpG islands are preferentially associated with bi-directional promoters. Perhaps, the reason for this discrepancy lies in the definition of a CpG island which is not clearly specified in the UCSC website. The frequency of overlapping genes on these chromosomes was also evaluated (Figure 1c). Overlapping sequences were seen in 182 genes in which 48 genes shared 5’ regions, 35 genes were located on the same strand and downstream of others, and 41 genes were found within other genes. Genes in these three categories account for 124 “overlapping” genes in Figure 1b. Additionally, 58 genes shared 3’ regions and these were included at the figure based on the location of their 5’ rather than their 3’ ends.

To further investigate the underlying meaning of the values of intergenic distance, we applied the Kernel Density Estimation method (Simonoff, 1996), which can be thought of as giving smoothed histograms with small Gaussian curves centered at each different value (Figure 1d-g). These plots, shown on logarithmic scales, show almost same biphasic shape in ch.20, ch.21 and ch.22, a major peak at 25 kb and a minor peak centered at 0.3 kb consisting of bi-directional transcripts within 1kb of one another. These plots reveal that the distributions of intergenic distances of ch.21 and ch.22 is not so different each other as previously reported median values of intergenic distances of ch.21 and ch.22 (Chen et al., 2002). In addition to these small chromosomes, a plot of chromosome 14, which has relatively low gene density to compare with chromosome 20 and 22, is indicated (Figure 1g). This minor peak is not observed between the ends of genes (Figure 1h) or between the ends and starts of genes (Figure 1i) suggesting that the close proximity might have been maintained for some functional reason. These bi-directional transcripts make up approximately 10% of genes, however, the actual first exons might have been missed in many genes, especially in predicted ones, so that our estimate should be considered a lower
limit. On the other hand, Adachi and Lieber (Adachi and Lieber, 2002) used ESTs to define the 5’ regions of genes in many cases. This might possibly lead to an overestimation of the frequency of bi-directional transcripts, because of the way that ESTs are used to define genes.

The biphasic nature of the distribution of genes on human chromosomes 20-22 raises interesting questions regarding its significance. One possibility is that genes sharing promoters might be co-ordinarily regulated in different tissues. However, this does not seem to be the case since genes with divergent promoters within 1kb of each other and genes share 5’ ends did not show a higher degree of concordant expression than other divergent transcript in the cancer genome anatomy project (CGAP) database (Figure 2). An alternative explanation might stem from the evolution of the human genome from a more compact genome. A large part of this expansion is thought to have been due to the radiation of repetitive elements (Smit, 1996). Perhaps the bi-directional genes have been maintained because interference with the shared promoter, caused by insertion of a repetitive element, would result in the simultaneous disruption of the regulation of two genes at once.

We therefore assessed how often repetitive elements were present in the vicinity of start sites by analyzing the content in 100 bp windows at various distances from where transcription begins. Analysis of these windows in 578 genes, which did not have divergent promoters within 10 kb of each other, showed that repetitive elements were strongly excluded from the first 300 bp of such start sites (Figure 3a). Repetitive elements were also excluded, although less strongly, from the ends of transcripts (Figure 3b). Figure 3c shows that divergent promoters, which had less than 1kb between start sites, were virtually devoid of repetitive elements. Since the average length of these promoters was 300 bp (Figure 1d), we also analyzed the content of repetitive elements in the 300-bp upstream regions of divergent transcripts separated by more than 10 kb (Figure 3c). Also the content of repetitive elements in the 150-bp upstream regions of divergent transcripts separated by more than 10 kb is indicated in Figure 3c. These data show that divergent promoters within
1kb of each other are distinct in that they contain low levels of such repetitive elements. Perhaps these observations can explain the persistence of close divergent transcripts in the human genome after its presumed expansion from a more compact form. If promoters, or, essential sequences necessary for regulation of gene expression (Zhang, 1998) like CpG islands, are inherently refractory to invasion by transposable elements (Figure 4) then it follows that two overlapping refractory zones might superimpose on each other thus preventing their separation as has happened in the majority of the genome.
Materials and Methods

For analyses of chromosomes 20, 21 and 22, we obtained sequence and mapping information from the GenBank database. We used the contigs (build 30), NT_011387, NT_025215, NT_028392, NT_011362, NT_030871, NT_35608, NT_011333 (chromosome 20), NT_029490, NT_011512, NT_030187, NT_030188, NT_011515 (chromosome 21), NT_011516, NT_028395, NT_011519, NT_011520, NT_011521, NT_011522, NT_011523, NT_030872, NT_011525, NT_019197, NT_011526 (chromosome 22), NT_026437.10 (chromosome 14). The potential start position of a gene was based on GenBank mapping information and possible multiple transcriptional start sites were not assessed. To obtain intergenic distances and status of direction, genes which located in the 5'end of contigs and genes which were included in other genes were excluded. Then the status of 5'end of genes was evaluated.

For CpG islands identification, the sequences between divergent promoters were extracted or centered 500bp sequences were extracted if distances between divergent promoters were less than 500 bp. Then we applied the CpG islands searcher (http://www.uscnorris.com/cpgislands/) and its command line version (Takai and Jones, 2003) with the criteria %GC ≥ 55 ObsCpG/ExpCpG ≥ 0.65 and Length ≥ 500 bp. The algorithm to search for CpG islands is described in the previous report (Takai and Jones, 2002). Other analyses were also done with programs coded by D.T.

For expression information, we used the “tissue” section of the cancer genome anatomy project web site (http://cgap.nci.nih.gov). Correlation coefficient values were determined from the SAGE data and EST data for up to 50 different tissues. For each combination of tissue, expression was computed by dividing the number of ESTs or SAGE tags representing the gene divided by the total number of ESTs or SAGE tags in all libraries with the given tissue. This ratio was then multiplied by 200,000, giving the number of ESTs or SAGE tags per 200,000. Then Pearson’s coefficient value was calculated for each combination of divergent or overlapping genes. If both two genes has zero as the number of
ESTs or SAGEs tags in a tissue, that values were not used to avoid that the coefficient becoming biased by excessive data at origin. Genes with less than 5 informative expression data points of tissues were also excluded. As a control, pairs of genes, located on different chromosomes, were randomly selected from genes which makes divergent transcript or genes share 5’ends.

\[
f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right),
\]

where kernel function \( K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \).

For kernel density estimation (Simonoff, 1996) we have used distribution function given where kernel function \( K \) is given

As the bandwidth \( h \), we have used

Repetitive elements were detected by the RepeatMasker (University of Washington Genome Center, Seattle, http://ftp.genome.washington.edu/cgi-bin/Repeat Masker).
Acknowledgements

We thank to T. Takai, G. Coetzee, D. Shibata, P. Laird and M. Lieber for critical reading of the manuscript. D.T and P. A. J are supported by National Cancer Institute Grants R01 CA82422 and R01 CA 83867.
References


Figure Legends

Figure 1
(a) Schematics of categories of the status of 5’ end of genes. When two genes make bi-directional transcripts they are categorized as “divergent transcript”. When two adjacent genes share the same strand, the 5’ nearest neighbor gene for the downstream one is categorized as “same strand”. When the 5’ end of gene overlaps with other gene, it is categorized as “overlapping”. When a gene apparently has no gene on the 5’ side, it is categorized as “end of contig”. (b) Bar graphs of the status of 5’ end of genes. The proportions of 5’ neighbor genes on each chromosome and a summation of all 3 chromosomes are shown. With “divergent transcript”, the proportions of subcategories based on the distances between transcriptional start sites are also indicated. (c) Schematics of categories for overlapping genes and numbers on human chromosomes 20-22. Genes with opposite directions and sharing 5’ ends were categorized into “Genes share 5’ ends”. Genes with the same direction and with the 5’ ends is included another, are categorized as “Genes share same strand”. In the case in which both ends of one gene were included within another, such a gene was categorized into “Gene within gene”. These three categories account for the “overlapping genes” in Figure 1b. From analysis of the status of the 3’end of genes, genes with opposite directions and sharing 3’ ends were categorized as “Genes share 3’ ends”. (d)-(i) Kernel estimates of the distribution of distances between divergent transcripts for of nearest neighbor genes of ch.20 (d), ch.21 (e), ch.22 (f), ch.14 (g), 3’ nearest neighbor genes and intergenic distances between starts (h) and ends of genes (i). Each tick along the bottom of the plot gives the distances of the nearest neighbor genes.

Figure 2
Correlation coefficient (r) of expression level of two genes which make divergent transcripts or which share 5’ ends. Expression data used for analyses were obtained from the “tissue” section of the cancer genome anatomy project (CGAP) web site. As a control,
pairs of genes, which were located on different chromosomes, were randomly selected from genes which were analyzed as “divergent transcript” or “genes share 5’ends”. In the CGAP database, 29, 25, 61, 11 and 23 pair of genes were available for “divergent transcript <1kb”, “divergent transcript 1~10 kb”, “divergent transcript >10 kb”, “genes share 5’ends” and “pairs of randomly selected genes” respectively. Informative data are depicted. Mean value of \( r_{EST} \) and \( r_{SAGE} \) are also indicated as horizontal bars in the plots. The highest value of \( r_{EST} \) in “Divergent transcript 1~10kb” came from a combination of crystallin beta B1 (CRYBB1) and crystallin beta A4 (CRYBA4); \( r_{EST} \) was 0.9998. Both two of genes were expressed in lens of eye specifically. The highest value of \( r_{SAGE} \) came from a combination of apolipoprotein L1 (APOL1) and apolipoprotein L2 (APOL2); \( r_{EST} \) was 1. Both genes were expressed in mammary gland and placenta.

**Figure 3**

Presence of interspersed repeats in the vicinity of transcriptional start sites of genes on chromosomes 20-22. The percentage of sequence characterized as “interspersed repeats” by “RepeatMasker” was calculated in various window sizes of sequences submitted to the RepeatMasker server. (a) Distribution of repetitive elements in 100 bp windows of sequences at the indicated distances from the transcriptional start sites of 578 genes which were separated by > 10 kb from the nearest adjacent start site. (b) Distribution of interspersed repeats in 100 bp windows at the indicated distances from the 3’ ends of 436 genes which make 3’end-to-3’end structure and separated by >10 kb from the nearest neighbor. (c) Presence of interspersed repeats in all 75 bi-directional promoters in which the transcripts started within 1kb of each other. The data is compared to a 150 bp or a 300 bp windows of sequence immediately upstream of 578 start sites more than 10 kb from the nearest divergent transcript and the average values for chromosome 20-22 (13,000 sequences of 100 bp each randomly selected from these chromosomes, which account for 1% of these chromosomes) are indicated.
Figure 4

Schematics for a model of exclusion of interspersed repeats in divergent promoters. Transcriptional starts and upstream regions exclude the insertion of interspersed repeats as a function of distance (see Figure 3a). If the distance between two divergent transcripts is sufficiently large, interspersed repeats can insert in the region between the two genes. If the distance is sufficiently small (<1 kb), both promoters synergically exclude targeting by interspersed repeats.
Figure 1

(a) Diagram showing chromosome contigs, divergent transcript, same strand, and overlapping contigs.

(b) Table showing distribution of divergent transcripts vs. same strand and end of contig lengths.

(c) Diagram illustrating gene relationships: shared 5' ends, same strand, shared strand, within gene, shared 3' ends.

(d-e-f) Density plots for chromosome regions:
- Ch.20 (Divergent) 436 genes
- Ch.21 (Divergent) 204 genes
- Ch.22 (Divergent) 324 genes

(g-h-i) Density plots for chromosome regions:
- Ch.14 (Divergent) 498 genes
- Ch.20-22 (3' end to 3' end) 959 genes
- Ch.20-22 (5' end to 3' end) 1009 genes
Figure 2: Correlation coefficient of expression level of two genes.

- Divergent transcript <1kb
- Divergent transcript 1~10kb
- Divergent transcript >10kb
- Genes share 5' ends
- Pairs of randomly selected genes

Diamonds represent EST data, and triangles represent SAGE data.
Figure 3
Targeted by interspersed repeats

Strength of exclusion

>1kb

Resistant to interspersed repeats

<1kb

Figure 4