Fine Structural Analysis of the Transcription Start Sites of Human mRNA

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

[1]Department of Virology and [2]Genome Center, Institute of Medical Science, the University of Tokyo, Japan, [3]Genome Science Center, Institute of Physical and Chemical Research (RIKEN), Japan, [4]Intelligent Communication Laboratory, NTT Communication Science Laboratories, Japan, [5]Helix Research Institute, Japan, [6]Institute of Molecular and Cell Biology, the University of Osaka, Japan, [7]Department of Life Sciences, the University of Tokyo, Japan.

**Address for Correspondence*
Department of Virology, Institute of Medical Science, University of Tokyo, 4-6-1, Shirokanedai, Minato-ku, Tokyo 108-8639, *Japan.*
TEL: 81-3-5449-5343
FAX: 81-3-5449-5416
e-mail: ysuzuki@ims.u-tokyo.ac.jp.

Key words: Transcription start site, full-length cDNA, Oligo-capping

Abstract

One-pass sequencing of 100,000 clones from cDNA libraries constructed by "Oligo-capping" enabled us to accumulate 5'-end sequences of 2251 kinds of named genes. As for redundantly isolated genes, we compared the 5'-ends of "Oligo-capped" cDNAs with each other. Unexpectedly the exact 5'-ends were heterogeneous in most cases. We selected 276 genes whose mRNA start sites were represented by more than five "Oligo-capped" cDNAs and mapped 6308 mRNA start sites in total onto the corresponding genomic sequences. Statistical analysis on these mRNA start sites revealed that the start sites were distributed over 61.7 bp with the standard deviation of 19.5 on average. Statistical learning using a learner decision tree algorithm C4.5 revealed that the most discriminative feature of the genes with tightly clustered transcription start sites was the presence of TATA box in the promoter. It was also shown that the Oct-1 and CAAT box have synergistic and inhibitory effect on the presence of TATA box, respectively.

Introduction

In order to elucidate the detailed mechanism of transcription initiation, we accumulated the precise information of the transcription start sites by random

sequencing of full-length enriched and 5'-end enriched cDNA libraries constructed by "Oligo-capping" method (Maruyama and Sugano, 1994; Suzuki *et al.*, 1997).

"Oligo-capping" is a novel method, which we developed previously (Maruyama and Sugano, 1994). This method uses the cap structure, which is the characteristic structure of the 5'-end of a eukaryotic mRNA. "Oligo-capping" replaces the cap structure of mRNA with synthetic oligoribonucleotide by three steps of enzymatic reactions (Fig. 1). Using the cap-replaced oligoribonucleotide as a sequence tag, cDNAs of the mRNAs that originally contained the cap structure were selectively cloned. Full-length cDNAs, which cover the complete sequence of mRNAs from the transcription start sites, were contained with proportion of 50- 80 % in this type of cDNA libraries. ("Oligo-capped" cDNA library; Suzuki *et al.*, 1997).

We have constructed "Oligo-capped" cDNA libraries from 34 kinds of human tissues and cultured cells. 5'-end one-pass sequencing of 100, 000 clones from these cDNA libraries enabled us to accumulate 5'-end sequences of "Oligo-capped" cDNAs for 2251 kinds of genes, each of which should represent the corresponding mRNA start site. We compared redundant "Oligo-capped" cDNAs of the same gene with each other and unexpectedly found that the 5'-ends of the cDNAs were heterogeneous in most cases. In this paper, we describe the statistical analysis of the distribution of the mRNA start sites.

Results

Accumulation of 5'-end sequences of "Oligo-capped" cDNAs

The 5'-ends of redundantly isolated "Oligo-capped" cDNAs for the same genes were compared with each other. Unexpectedly the exact 5'-ends of the cDNAs were heterogeneous in most cases. Figure 2 exemplifies the case of PPH alpha and nucleolin. The diversity of the 5'-ends was observed in both cases although that of PPH alpha cDNAs was small and that of nucleolin cDNAs was significant (Fig. 2).

We mapped the 5'-ends of "Oligo-capped" cDNAs onto the genomic sequences by computational search of Genbank. As a result, promoters (between 500 bp upstream and 100 bp downstream of the mRNA start sites) could be retrieved from Genbank for 1031 genes. Among them, more than five 5'-ends of "Oligo-capped" cDNAs were mapped for 276 promoters, each of which should represent the corresponding mRNA start site. 6308 "Oligo-capped" cDNAs were mapped on these promoters in total (average redundancy= 22.9).

Distribution of the mRNA Start Sites

Using the positional information of each mRNA start site on the promoters for these 276 kinds of genes, we statistically analyzed the distribution of the mRNA start sites. (i) the distance between the most upstream mRNA start site and the most downstream; (ii) the frequency at which each nucleotide in the promoter is used as a transcription start sites (iii) the standard deviation of the distribution of the mRNA start sites were

calculated on each gene. Figure 3 exemplifies the results of the calculation for serum albumin and glypican. The mRNA start sites of serum albumin and glypican were distributed over 6 and 55 bp with the standard deviations of 1.4 and 11.0, respectively. Reported TATA box in the promoter of serum albumin is shown by upper cases and a box. The nucleotides marked with hatched boxes represent the previously identified mRNA start sites. It should be also noted that the previous analysis gave the single start sites in both cases.

After the calculations were performed on each gene, the results were statistically analyzed using all the data of the 276 genes. Figure 4A and B show the statistical results on (i) and (iii), respectively. The mRNA start sites were scattered over 61.7 bp (Fig. 4A) with the standard deviation of 19.5 on average (Fig. 4B).

Relation between the variability of mRNA start sites and TF binding sites

In order to determine whether there is correlation between the presence or absence of the TF binding site in the promoters and the distribution of the transcription start sites, we performed statistical learning, using a learning-based classification program, C4.5 (Quinlan, 1993). Given a data set of cases with attributes, C4.5 generates a decision tree that classifies the cases into discrete classes. The attribute at each branch in the decision tree is selected so that the gain ratio relevant to the classification should be maximized. In brief, the attributes that are located at the upper levels of the decision tree are most discriminative, thus, most reflecting the features of each class.

In the present analysis, we divided promoters into two groups as to whether their transcription start sties were highly variable (class 0) or tightly clustered (class 1). The promoters whose standard deviation of the start site distribution is more than 5 and less than 5 were tentatively categorized as class 0 and class 1, respectively. According to this criterion, 42 promoters were categorized as class 1 and 234 were class 0.

Using all the 205 kinds of matrices of TF binding motifs in TRANSFAC (Rel. 4.0; http://transfac.gbf.de/index.html; Heinemeyer *et al.*, 1999), TF binding sites were predicted on each promoter by TFBIND (Tsunoda and Takagi, 1999). Then, the attribute value is assigned for each promoter as a binary that indicates a TF binding site that appears in the promoter.

Using C4.5, a decision tree that classifies 276 promoters into class 0 or 1 with respect to 205 attributes of TF binding sites was generated. Figure 5A shows the upper 2 levels of the generated decision tree. The attribute located at the top level was that of TATA box. This shows that the presence of TATA box in the promoter was the most discriminative feature in the tested 205 kinds of TF binding sites. Among 42 class 1 promoters, 29 contained TATA box. Conversely, among 102 TATA-containing promoters, 29 were classified as class 1.

At the second level of the tree, CAAT box and Oct-1 were located. All the 19 promoters containing both TATA box and CAAT box were classified as class 0 promoters. Similarly, almost all (137/143) promoters that contain neither TATA box nor

Oct-1 were class 0.

Discussion

In this paper, we described the diversity of the mRNA start sites and analyzed the association between the distribution of the transcription start sites in the promoter and the predicted upstream TF binding sites.

The extent how widely the transcription start sites were distributed was different between genes in spite that all the mRNAs are transcribed by RNA polymerase II (Orphanides *et. al*, 1996; Lee and Young, 1998). Considering, all the primary information is encoded in DNA sequences, there should be a certain sequence elements that determines the distribution of the transcription start sites. Statistical learning by a learning-based classification program, C4.5 revealed that among the 205 kinds of TF binding motifs in TRANSFAC, presence of TATA box is the most discriminative factor of the genes with tightly clustered transcription start sites (Fig. 5).

However, TATA box was not solely sufficient to define the variability of transcription start sites (Fig. 5). There were many TATA-containing genes with widely distributed start sites and many TATA-less genes with tightly clustered start sites. Additional factors should be involved in complementing or interfering with the function of TATA box. For almost all (137/143) of the promoters containing neither TATA box nor Oct-1, the start sites were widely distributed (stdev. > 5; Fig. 5). On the other hand, in spite of the presence of TATA box, all the 19 promoters containing both TATA box and CAAT box had the widely distributed start sites. Oct-1 and CAAT box may have the synergistic and inhibitory effect on the function of TATA box, respectively.

It is possible that the distribution of the start sites should be reflecting the biophysical features of the mutual interaction between the TFs, the pre-initiation complex and the promoter. The crystal structure of ternary complex of TFIIB-TBP-TATA box has elucidated that the DNA backbone is bended by 90 degree, when TBP is recruited to TATA box (Nikolov *et. al*, 1995). This structural change may cause the pre-initiation complex fixed rigidly on the transcription start site. On the contrary, if there is no TATA box in the promoter, the pre-initiation complex may remain unstable on the promoter.

Similarly when a certain TF is recruited to the promoter, a rigid complex could be also formed between the TF, the pre-initiation complex and the promoter. Once such a rigid complex is formed, the position of the pre-initiation complex relative to the promoter may be strictly determined. In this case the transcription should be initiated from tightly clustered positions. When the complex is loose, it may allow the pre-initiation complex to move back and forth on the promoter, which causes the transcription initiation from widely distributed positions. This dynamics of the pre-initiation complex on the promoter may determine the distribution of the transcription start sites.

One of the ways to understand the biological systems may be the biophysical description of every interaction of biological molecules. To this end, molecular

dynamics should be simulated on every aspect of biological process. When the molecular dynamics of the transcription initiation is analyzed, precise information on how the pre-initiation complex interacts with TFs and DNA and at which nucleotide the transcription is initiated should be indispensable. It is also important to describe whether the interaction of each molecule is static or dynamic. Detailed information on the position and frequency of the transcription start sites described in this paper should lay the groundwork to start the elucidation of biophysical principle that governs the transcription initiation.

References

1. Orphanides, G., Lagrange, T., and Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes Dev.* **10:** 2657-2683.
2. Heinemeyer, T., Chen, X., Karas, H., Kel, A. E., Kel, O. V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F., Wingender, E. (1999). Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res.* **27:** 318-322.
3. Lee, T. I., and Young, R. A. (1998). Regulation of gene expression by TBP-associated proteins. *Genes Dev.* **12:** 1398-1408.
4. Maruyama, K., and Sugano, S. (1994). Oligo-capping: a simple method to replace the cap structure of eucaryotic mRNAs with oligoribonucleotides. *Gene* **138:** 171-174.
5. Nikolov, D. B., Chen, H., Halay, E. D., Usheva, A. A., Hisatake, K., Lee, D. K., Roeder, R. G., and Burley, S. K. (1995) Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature* 377, 119-28.
6. Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
7. Suzuki, Y., Yoshitomo, K., Maruyama, K., Suyama, A., and Sugano, S. (1997). Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene* **200:** 149-156.
8. Tsunoda, T., and Takagi, T. (1999) Estimating transcription factor bindability on DNA. *Bioinformatics* **15:** 622-30.

Figure legends

Figure 1 Scheme of the construction of "Oligo-Capped" cDNA libraries..

Figure 2 Sequence alignment of the 5'-ends of "Oligo-Capped" cDNAs of PPH alpha (A) and nucleolin (B).

Figure 3 Distribution of the mRNA start sites of serum albumin (A) and gypican (B).

Figure 4 The distance between the most upstream and the most downstream mRNA start sites were calculated on each gene and their distribution is shown (A). Standard

deviation of the distribution of the transcription start sites was calculated on each gene. Then the distribution of the standard deviation is calculated (B).

Figure 5 Relation between the distribution of transcription start sites and TF binding sites in the promoter. The ratio of the population between class 1 and class 0 belonging to each group is shown at the bottom margin of each group.

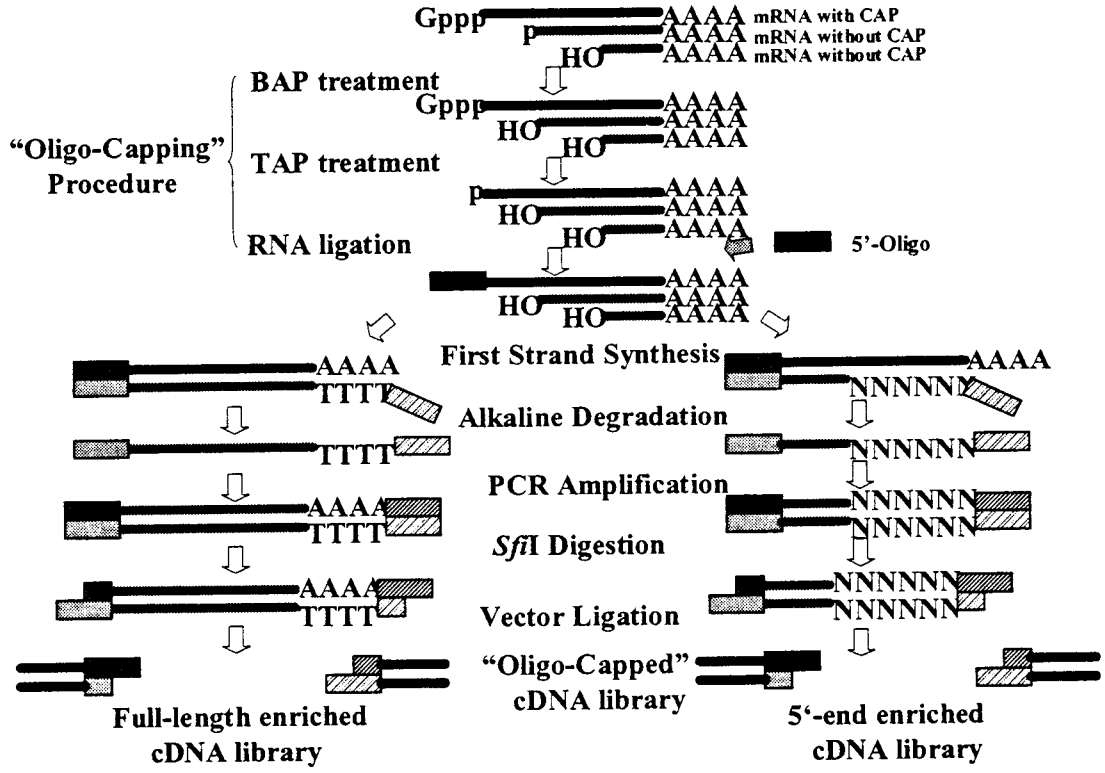

Figure 1

A. 5'-end Sequences of PPHα cDNA

```
GCTCACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
GCTCACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
  TCACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
  TCACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
    ACTTGCAGCAATGGCTTGGATTAGATCCACTTGCATTCTCTTTTTTACCTTGCTTAGATG
```

B. 5'-end Sequences of Nucleolin cDNA

```
TTCCACAGGCGATTACTGGGCAGTCTTTCGCCTCAGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
TTCCACAGGCGATTACTGGGCAGTCTTTCGCCTCAGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                        CTTTCGCCTCAGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                        CTTTCGCCTCAGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                        CTTTCGCCTCAGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                         TTTCGCCTCAGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                                 TCAGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                                 TCAGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                                 TCAGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                                   AGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                                   AGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                                   AGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                                   AGTCTCGAGCTCTCGCTGGCCTTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                                                         TTCGGGTGTACGTGCTCCGGGATCTTCAGCACCCG
                                                                GTACGTGCTCCGGGATCTTCAGCACCCG
```

Figure 2

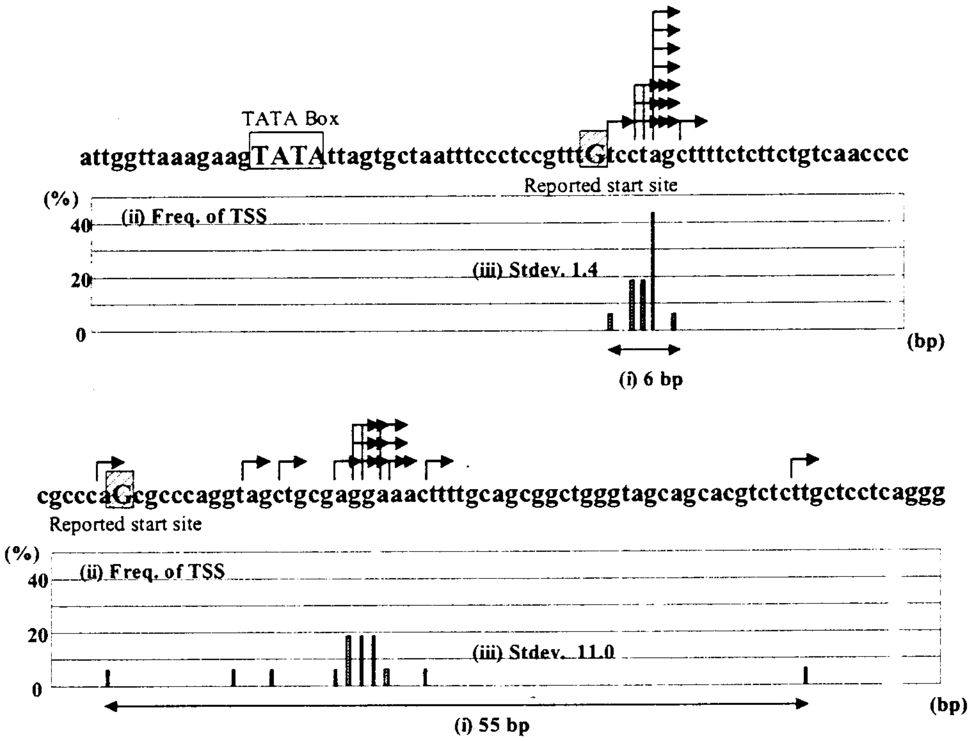

Figure 3

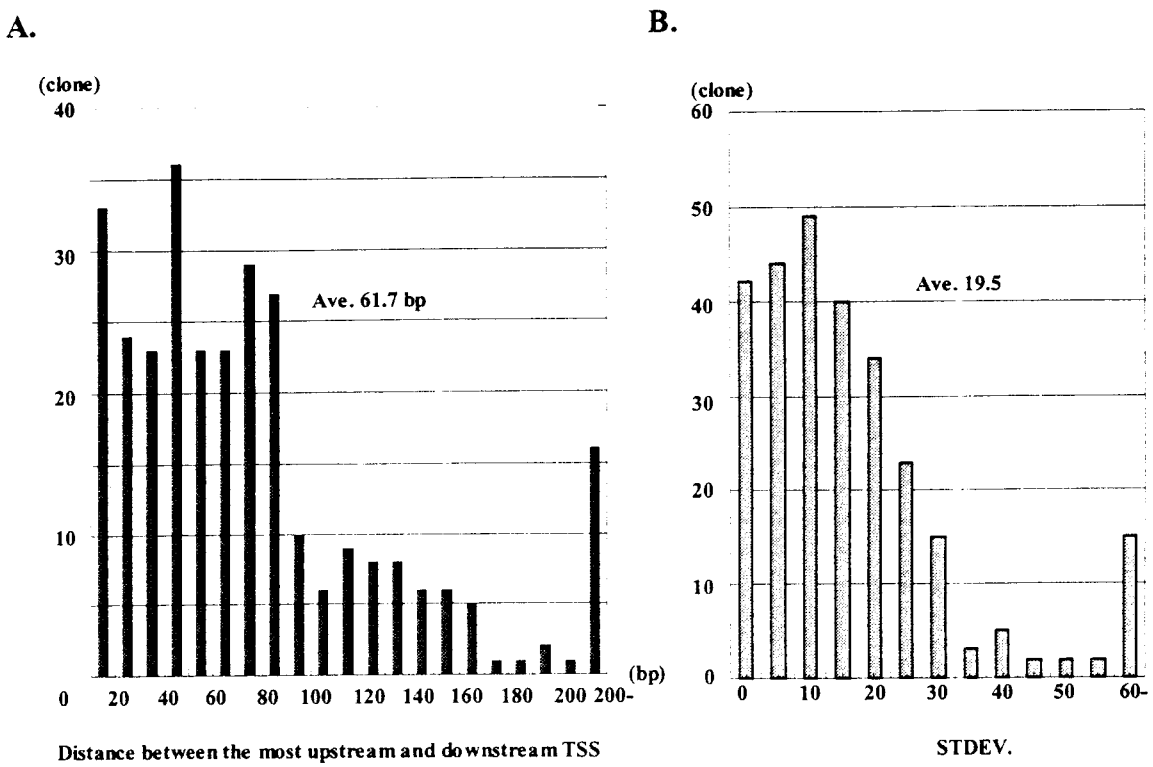

Figure 4

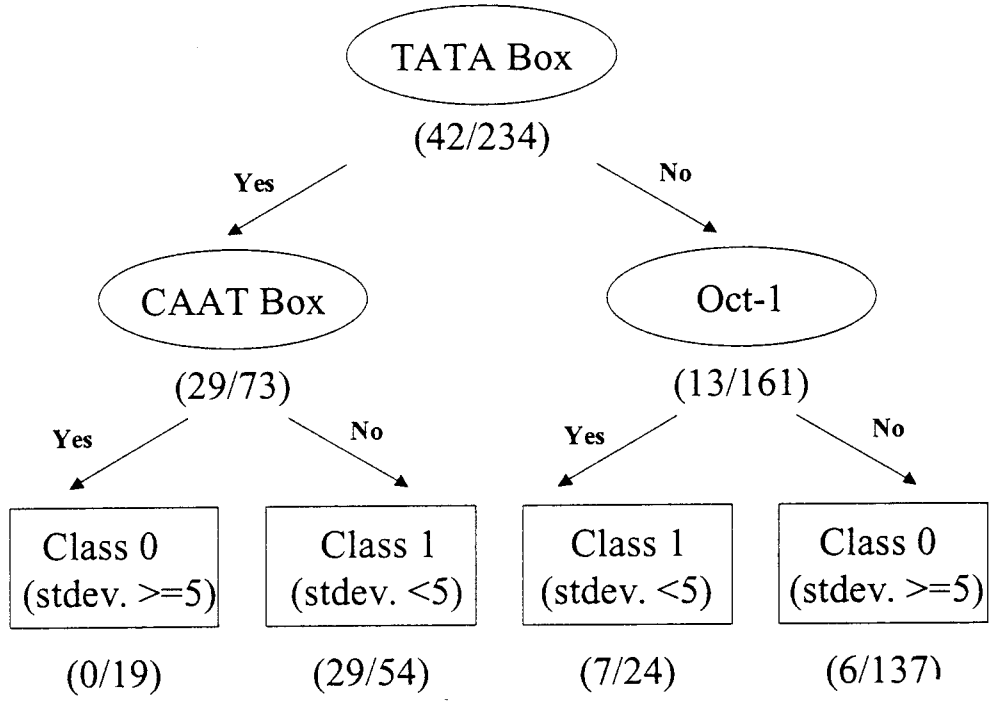

Figure 5