

Collection and Analysis of Eukaryotic Promoter Regions: DBTSS (DataBase of Transcriptional Start Sites)

Riu Yamashita

ryamasi@ims.u-tokyo.ac.jp

Yutaka Suzuki

ysuzuki@manage.ims.u-tokyo.ac.jp

Toshihisa Takagi

takagi@ims.u-tokyo.ac.jp

Sumio Sugano

ssugano@ims.u-tokyo.ac.jp

Kenta Nakai

knakai@ims.u-tokyo.ac.jp

Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1
Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Keywords: transcription, comparative genomics, full-length cDNA, promoter, *cis*-element

1 Introduction

Recent determination of human and mouse draft genome sequences should be the landmarks for the post-sequencing era. One of the major challenges in this era is the interpretation of the promoter regions. To this end, precise identification of the transcriptional start sites (TSSs) is essential. However, such an information cannot be obtained from usual cDNA or EST data. Although Eukaryotic Promoter Database (EPD) contains reliable data, the number of the data is about 1400, which is not enough for global promoter analysis.

To overcome this problem, we have constructed a database, DBTSS [3], which contains the information of a number of 5'-end sequences produced by the oligo-capping method and mapped onto the genome sequence. The oligo-capping method enables the precise determination of the 5' ends of mRNAs [2, 4]. Here we report a major extension in its ver.3: support of the data of multiple species (human, mouse, and nematode). It contains not only the information of each TSS of these species but also the local sequence similarity between the upstream regions of their orthologous genes.

2 Methods

First, vector regions and low quality regions of each oligo-capping clone were removed. Then, they were compared with reference sequences (RefSeq) [5] using the BLAST program. They were regarded as identical when they were more than 95% identical and less than 10^{-100} in e-value with one of the reference sequences. Sequences that match to multiple RefSeq entries were discarded. Finally, the clones were mapped onto the genome sequence encompassing from -500kbp to $+500\text{kbp}$ with the TSSs annotated in RefSeq using the Sim4 program. From 284,687 human oligo-capping clones, 155,304 were mapped to the genome in this way while 18,986 sequences from 33,720 clones were mapped in mouse. Since the sequences of nematode have not been stored in RefSeq, 7,994 out of 20,340 clone sequences were directly mapped onto the *C. elegans* genome sequence using the BLAT program [1]. For the comparison of upstream sequences, the orthology table between species is needed. From the LocusLink of NCBI, we could obtain the information of 3,470 human-mouse homologs and 143 human-nematode homologs.

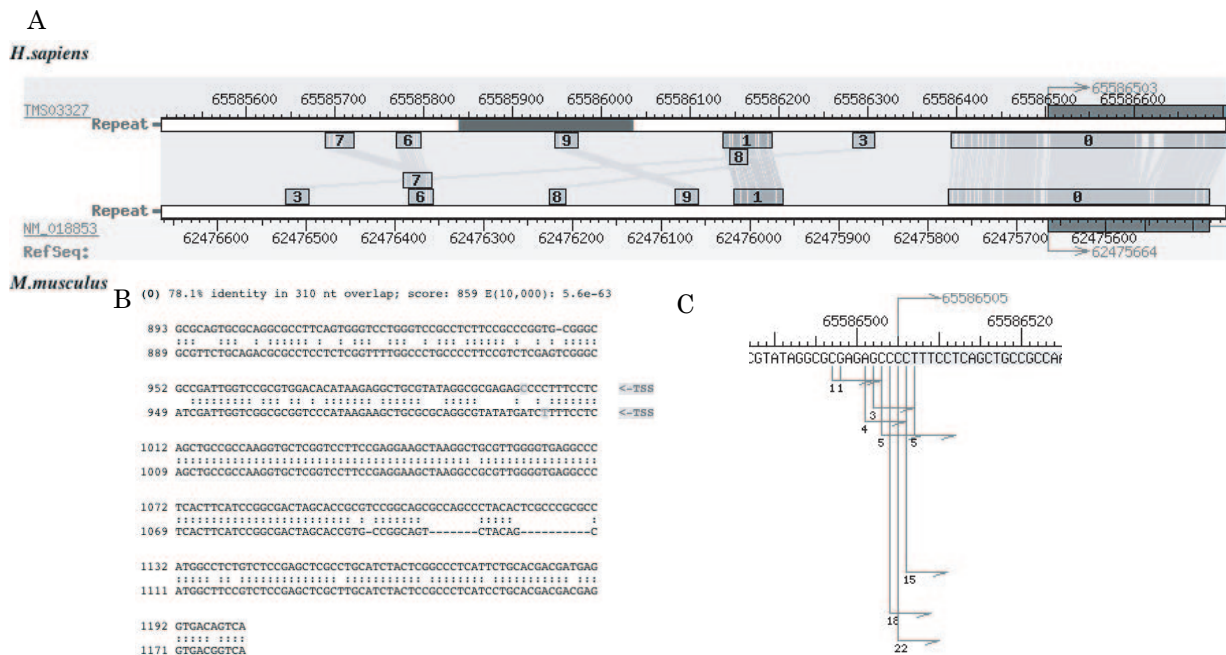


Figure 1: An example of TSS. Human and mouse ribosomal protein, large P1 (RPLP1) (A) Sequence comparison of TSS regions. NM_001003 (human) and NM_018853 (mouse) correspond to the RPLP1 gene. (B) Most conserved region (number 0) detected by the LALIGN algorithm. (C) Distribution of TSSs in NM_001003.

3 Results and Discussion

From our DBTSS web page (<http://elmo.ims.u-tokyo.ac.jp/dbtss/>), users can retrieve the TSS information of a specified gene in several ways: they can use a gene-name search or can directly enter the RefSeq IDs (such as NM_001003), LocusLink IDs, UniGene IDs, or gene symbols. In Figure 1, an example of ribosomal protein large P1 (RPLP1) is shown.

The RPLP1 gene corresponds to human NM_001003 and mouse NM_018853. DBTSS can dynamically compare these human and mouse upstream sequences. In Figure 1-A, local similarity around the TSS region is shown. Each number represents locally similar region identified by the LALIGN program. For example, the number 0 corresponds to the most conserved region (see Figure 1-B for the alignment). In this gene, the TSS region seems to be highly conserved even at its upstream region.

Another advantage of DBTSS is that it allows us to see the distribution TSS for each gene. We found that the genes can be roughly classified into two classes with the threshold value 50bp of the standard deviations of their TSSs; in one class, TSS deviations are likely to be caused non-regulatory (probably due to the slippage of polymerases) while, in the other class, they seem to be regulated by multiple promoters. We show an example of the former class in Figure 1-C.

References

- [1] Kent W.J., BLAT—the BLAST-like alignment tool, *Genome Res.*, 12(4):656–664, 2002.
- [2] Maruyama, K. and Sugano, S., Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides, *Gene*, 138(1-2):171–174, 1994.
- [3] Suzuki Y., Yamashita R., Nakai K., and Sugano S., DBTSS: Database of human transcriptional start sites and full-length cDNAs, *Nucleic Acids Res.*, 30(1):328–331, 2002.
- [4] Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A., and Sugano, S., Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library, *Gene*, 200(1-2):149–156, 1997.
- [5] <http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html>