Distribution Analysis of 5' Splice Site-Like Sequences in Human and Mouse pre-mRNAs

Sumie Kitamura-Abe1,2 Nobuyuki Takahashi3 Atsushi Sakurai1
sumie@ya2.so-net.ne.jp nobu@cc.hokkyodai.ac.jp ss97446as@sfc.keio.ac.jp

Takanori Washio1 Mineichi Kudo4 Masaru Shimbo5
washy@sfc.keio.ac.jp mine@main.eng.hokudai.ac.jp shimbo@do-johodai.ac.jp

Akihiro Tsutsumi2 Masaru Tomita1
ttsutsu@eng.hokudai.ac.jp mt@sfc.keio.ac.jp

1 SFC Research Institute, Keio University, 5322 Endo, Fujisawa 252-8520, Japan 2 Department of Applied Physics, Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo 060-8628, Japan 3 Hakodate Campus, Hokkaido University of Education, 1-2 Hachiman-cho, Hakodate 040-8567, Japan 4 Division of Systems and Information Engineering, Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo 060-8628, Japan 5 Hokkaido Information University, 5-9-2 Nishinopporo, Ebetsu 069-8585, Japan

Keywords: 5' splice site, splicing, position tree, human, mouse

1 Introduction

The exon definition model accounts for the paring between the 3' splice site and its downstream 5' splice site [5]. It was demonstrated experimentally that 5' and 3' splice-site pairing in metazoan pre-mRNAs occurred in the two distinct steps [1]. In our previous paper, we proposed a subclass method to predict 5' splice sites in mammalian pre-mRNAs, and suggested that several 5' splice site pattern sequences were observed more frequently on the first or second introns than on the other order introns [3]. However, the relation between the 5' splice site sequences and the order of introns is currently under investigation. On the other hand, positional characterization of false positives using computational prediction of human splice sites was presented [6]. In the present study, by using a position-tree method to analyze the distribution of the 5' splice site-like sequences in human and mouse pre-mRNAs, we obtained the 5' splice site sequences whose lengths were minimal but were sufficient for specifying the 5' splice sites. Then we investigated the distribution of the 5' splice site-like sequences that were one nucleotide shorter than the obtained 5' splice site sequences. As a result, it was confirmed that the 5' splice site-like sequences had the tendency of the uneven distribution within pre-mRNAs.

2 Materials and Method

The position tree approach extracts a substring, referred to as a substring identifier, which is unique for a certain position in a long text string and does not appear in the other positions of the text string. Such a substring is seen as one path of a tree, that is, the path from the root to a leaf corresponding to that particular position. Every leaf has a designated position number.

In our experiment, we designate G of the invariant GT dinucleotide of the real 5' splice site as position +1. We define the real 5' splice site sequences as the substring identifiers of the sites, starting at position -3. Because the consensus sequence, (C or A) AG/GT (A or G) AGT (where the stroke '/' indicates the boundary between exon and intron), starts with position -3 [4]. In our experiment, 303 human and 143 mouse pre-mRNAs that include more than three introns each were sampled from GenBank (GenBank, 2000) using the program CLEANUP 1.8.3 which excludes redundancy [2]. The
total number of introns in these human and mouse pre-mRNAs were 2676 and 939, respectively. One position tree was constructed for each of the pre-mRNAs. The sequences of minimal length that identify real 5' splice sites were extracted. Then we investigated the distribution of the 5' splice site-like sequences that were one nucleotide shorter than the obtained 5' splice site sequences within each pre-mRNA. The frequency was obtained by dividing the number of 5' splice site-like sequences by the number of nucleotides within each individual exon or intron.

3 Results and Discussion

One position tree was constructed for each of the 303 human pre-mRNAs and 143 mouse pre-mRNAs. A total of 2676 real 5' splice site sequences starting at position -3 were obtained in 303 human pre-mRNAs. The 5' splice site-like sequence was obtained by omitting one nucleotide at the 3' end of each real 5' splice site sequence. In the same way, a total of 939 real 5' splice site sequences were obtained in 143 mouse pre-mRNAs and the 5' splice site-like sequences were constructed.

We will use the human osteomodulin gene as an example. This gene includes 4 introns in total. The 5' splice site sequence of the first intron of this gene was GAG/GTGGGT (where the stroke '/' indicates the boundary between exon and intron). The sequence contains 9 nucleotides, which is the same length as the consensus sequence (C or A) AG/GT (A or G) AGT [4]. The 5' splice site-like sequence, one nucleotide shorter than the real 5' splice site sequence, was GAGGTGGG. The distribution of this 5' splice site-like sequence within the pre-mRNA was investigated. The 5' splice site-like sequence appeared once in the second intron, once in the last exon and six times in the downstream of the last exon. The frequency was obtained by dividing the number of 5' splice site-like sequences by the number of nucleotides within each particular exon, intron, upstream sequence or downstream sequence. In a similar fashion, the 5' splice site sequences were obtained for the rest of the introns of the pre-mRNA and the distribution of the 5' splice site-like sequences within each pre-mRNA was investigated.

As a result, the 5' splice site-like sequences that were the same as the 5' splice site sequences consisting of 9 or more nucleotides were detected more frequently within the exons and introns at the interior region than within those in the terminal ends of human and mouse pre-mRNAs. As for the distribution of the 5' splice site-like sequences consisting of 6, 7 or 8 nucleotides, we could not find a common tendency between human and mouse pre-mRNAs. It is suggested that the same sequences consisting of 9 nucleotides as the 5' splice site sequences exist less frequently within the exons and the introns at the terminal ends than within those at the interior region. It has not yet been clarified that the properties of the exons and introns at the terminal ends are different from those in the interior regions of pre-mRNAs. However, our approach using the position -tree method seems to be useful for analyzing the distribution of the 5' splice site-like sequences in pre-mRNAs.

References