Analysis of the Context of 5' Splice Site Sequences and 5' Splice Site-Like Sequences in Mammalian mRNA Precursors by a Position Tree

Sumie K. Abe[1] sumie@ya2.so-net.ne.jp
Nobuyuki Takahashi[2] nobu@cc.hokkyodai.ac.jp
Mineichi Kudo[3] mine@main.eng.hokudai.ac.jp
Masaru Shimbo[3] shimbo@huie.hokudai.ac.jp
Akihiro Tsutsumi[1] atsutsu@eng.hokudai.ac.jp

[1] Department of Applied Physics, Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo 060-8628, Japan
[2] Hakodate Campus, Hokkaido University of Education, 1-2 Hachiman-cho, Hakodate 040-8567, Japan
[3] Division of Systems and Information Engineering, Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo 060-8628, Japan

Keywords: 5' splice site, genome sequence, position tree

1 Introduction

The pre-mRNA splicing takes place in the spliceosome composed of several small nuclear ribonucleoprotein particles (such as U1, U2, U4, U5, and U6 snRNPs) and many proteins. To analyze 5' splice site sequences, several methods have been proposed [1, 3]. In our previous paper, we proposed a subclass method, and could predict correctly 5' splice sites from unknown sequences with a discrimination rate of over ninety percent [2]. However there were some sequences that were not correct 5' splice site sequences but were similar to 5' splice site sequences. In the present study by a position-tree method, we obtained the sequences, we called such sequences pattern sequences, whose lengths were minimal but were sufficient for specifying 5' splice sites. Then we investigated the distributions of the 5' splice site-like sequences that were one nucleotide shorter than the obtained pattern sequences.

2 Method

The position tree approach extracts a substring, called gsubstring identifiersh, that 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 at every position is expressed as one path of a tree, that is, the path from the root to a leaf corresponding to that position.

In our experiment, the nucleotide sequences that contained 100 introns of pre-mRNAs of mammalian pre-mRNAs were joined to form a string. A position tree for the string was constructed. We made a positive set, selecting the substring identifiers of the sites, starting within the range of 11-nucleotides (6 in the exon and 5 in the intron) around 5' splice sites. We constructed a negative set, selecting the substring identifiers of the other sites except 5' splice sites. The following types of the sequences of minimal lengths that identify 5' splice sites were extracted in order by the position-tree method: Pattern sequences that were common among as large a number of the substring identifiers of the positive set and did not agree with the sequences of the negative set.

Then we investigated the distributions of the 5' splice site-like sequences that were one nucleotide shorter than the obtained pattern sequences.

3 Results and Discussion

We obtained 60 pattern sequences to cover 100 5' splice sites of the positive set. The average length of the pattern sequences was 8.1, which was a little shorter than the 9 of the consensus sequence (C or A)AG/GT(A or G)AGT (where the stroke '/' indicates the boundary between exon and intron) [3]. Each of the pattern sequences agreed with only a small portion of the whole set of the positive set. For example, even the first pattern sequence G/GTGAGTC includes only 6 nucleotide sequences of the positive set. By using the position tree method, some genes that were related to each other, were extracted together according to 5' splice site sequences. For example, the first pattern sequence G/GTGAGTC agreed with 4 of globin family genes. The ribosomal protein gene and the oxytocin/neurophysin gene, however, were also extracted according to the first pattern sequence. With the 5' splice site data gathered here, for a test case, we are able to foresee possible new data steps to classify genes, according to the 5' splice site sequences, into a database.

On the other hand, we analyzed the distribution of the 5' splice site-like sequences of the first 30 pattern sequences. The 5' splice site-like sequences were detected within the first introns and the last introns more than the internal introns, though the total length of all the internal introns was similar with that of all the first introns and that of all the last introns. We are conducting analysis of the tendency of the uneven distribution of the 5' splice site-like sequences within pre-mRNAs.

References

[1] Iida, Y., DNA sequences and multivariate statistical analysis. Categorical discrimination approach to 5' splice site signals of mRNA precursors in higher eukaryotes' genes, *Comput. Applic. Biosci*, 3(2):93–98, 1987.

[2] Kudo, M., Kitamura-Abe, S., Shimbo, M., and Iida, Y., Analysis of context of 5'-splice site sequences in mammalian mRNA precursors by subclass method, *Comput. Applic. Biosci.*, 8(4):367–376, 1992.

[3] Mount, S.M., A catalogue of splice junction sequences, *Nucleic Acids Res.*, 10(2):459–472, 1982.