Characterization and Classification of Splicing Patterns

Yumi Tamada **Toshio Shimizu**
gs98613@si.hirosaki-u.ac.jp slsimi@si.hirosaki-u.ac.jp
Department of Information Science, Graduate School of Science, Hirosaki University
3 Bunkyo-cho, Hirosaki, Aomori 036-8561, Japan

1 Introduction

We analyzed the splicing patterns to distinguish the true splicing patterns from pseudo patterns.

The data set of 506 true donor patterns of human genes were finally extracted from DDBJ entries with the descriptions of "complete cds." in the DEFINITION field and "prim_transcript" without the word "putative" in the FEATURES field.

The sequence of length of 12 bps, i.e., 4bps upstream (denoted $-4\sim-1$ positions) and 8bps downstream ($+1\sim+8$ positions) around a donor site, is defined as a donor pattern.

62,922 pseudo donor patterns with the dinucleotide, "GT" at the positions of +1 and +2, were obtained from 100 random nucleotide sequences of length 10,000 bps generated by using random numbers.

We first obtained the weight matrix from the true donor patterns to calculate the scores of the donor patterns [1].

Next, we calculated the information contents of the donor patterns by using Shannon's information measure, and examined the correlations among individual positions [2].

We also tried to make some combinations of nucleotide positions to investigate the importance of each position.

2 Results and Discussion

2.1 Calculation of the Score

The scores are distributed around 0.9-1.0 for true donor pattern and around −0.1-0.0 for pseudo pattern (Fig. 1).

However, both the distributions overlap each other in the range of 0.0 to 1.2.

And, there seem to be three more peaks other than the major peak (with the highest score) in the distribution curve for real donor pattern.

It suggests that at least four kinds of donor patterns are existing.

2.2 Information Contents

Information contents of positions −1, +3, +4, +5 and +6 are much larger than other positions except +1 and +2.

When the nucleotide G doesn't appear at +5 position, information content at −1 position increases largely; the nucleotide A appears at −1 position with a higher probability than the other nucleotides.

This result means the position −1 covers a lowering of the information content at +5 position.

Thus, we obtained (−1,+4&+5&+6) and (+3,+5) as the combinations of positions coupled each other.

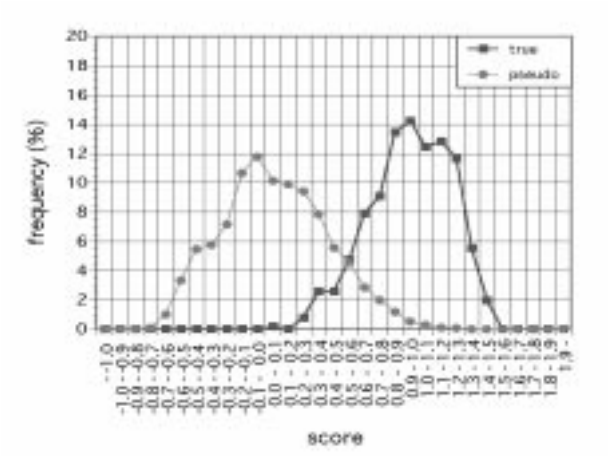

Figure 1: The score distribution.

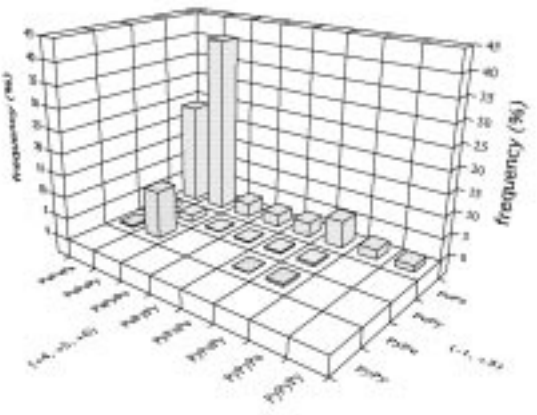

Figure 2: Pattern distribution.

2.3 Clustering of the Donor Patterns

By grouping four the nucleotides A, T, C and G into two categories purine (denoted by Pu) and pyrimidine (Py), the donor patterns are clustered into four groups (Fig. 2) : (Pu,Pu,Pu/Py), (Pu,Py,Pu/Py), (Py,Pu,Pu/Py) and (Py,Py,Pu/Py) at (+4,+5,+6) positions.

This result corresponds to the score distribution curve with four peaks in Fig. 1.

We are also trying to classify of the donor patterns by using k-weight matrix model [3, 4].

References

[1] Senapathy, P., Shapiro, M.B., and Harris, N.L., Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project, *Methods in Enzymology*, 183:252–278, 1990.

[2] Tolstrup, N., Rouze, P., and Brunak, S., A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites, *Nucleic Acids Research*, 25:3159–3163, 1997.

[3] Burge, C.B., Modeling dependencies in pre-mRNA splicing signals, *Computational Methods in Molecular Biology*, Elsevier, 129–164, 1998.

[4] Murakami, K. and Takagi, T., Clustering and detection of 5' splice sites of mRNA by k weight-matrices model, *Genome Informatics 1998*, Universal Academy Press, 282–283, 1998.