# Comparison of Statistical Algorithms for Predicting Splice Junctions in mRNA Precursors of Mammalian Genes

Yukiyasu Ogawa
ogawa@sunta.csse.muroran-it.ac.jp

Tomomasa Nagashima
nagasima@csse.muroran-it.ac.jp

Sirajuddin Khawaja

Department of Computer Science and Systems Engineering,
Muroran Institute of Technology,
27-1 Mizumoto-cho, Muroran, Hokkaido 050, Japan

## Abstruct

It is well known that many eukaryotic genes are interruped by introns, which are removed from mRNA precursors by the RNA splicing mechanism. While it is well accepted that the consensus sequences of exon-intron boundaries in mRNA precursors are important for specifying splice sites, the signals that govern the excision of introns are not well understood yet, because actual splice site sequences are more or less different from the consensus sequences.

So far several statsistical methods for predicting actual splice site sequences ( splice junctions)in pre-mRNAs of mammalian genes have been proposed. Shapiro-Senapathy gave a method by weight matrix; Iida has proposed a method based on the quantification analysis(categorical discriminant analysis). He applied his method for analyzing the 3'-splice site sequences as well as 5'-splice site and discussed mutational problems in beta-globin genes.

However, while the statistical methods proposed so far have some ability for predicting the splice site sequences in statistical tests, it seems to be far from sufficient whenn applied to actual problems i.e., predicting the actual splice junctions in complete mammalian genes.

We propose here a new algorithm which can be applicable to predict actual splice junctions in complete mammalian genes and demonstrate its ability by comparing the predicting performances between the algorithms by ours, Shapiro-Senapathy and Iida.

小川幸泰，長島知正，クワジャ・シラジュディン：室蘭工業大学情報工学科，〒050 北海道室蘭市水元町 27-1

Our algorithm is described as an extention of the categorical discriminant analysis(CDA) by Iida into the hierarchical form i.e., at the first level, we start from the ordinary categorical discriminant anlysis and determin the classes of sample sequenses which do not fall into the overlapping region of sample scores. For the samples which fall into the overlapping region of the sample scores are treated in the next level. For these data, we apply the categorical discriminant analysis. This process is repeated till the number of the samples which fall into the overlapping region of sample scores becomes to be negligible.

By applying this algorithm to the 3'-splice junctions as well as 5'- splice junctions of Rat Chymotrypsin B gene, we have obtained fairy well predicting peformance compared to Shapiro-Senapaty and Iida. In tab.1, we have presented a comparision of the predicting ability obtained by ours and CDA. As shown in tab.1, our algorithm gives fewer potential sequences(candidate) for the splice junctions than CDA. As for Shapiro-Senapathy algorithm, we have also obtained that the algorithm gave more potential sequences than CDA, that implies our algorithm peforms well than CDA and Shapiro-Senapathy method.

As a summary, we have develloped the algorithm which acts as a filer for selecting the splice signals among a huge number of unknown sequences. We will discuss the details of our algorithm in the symposium.

| *No. of Introns* | *(a)* | *(b)* | | |
|---|---|---|---|---|
| | *No. of active seq.* | I | II | III |
| 6 | 165 | 1 | 4 | 2 |
| | | 2 | 0 | 6 |
| | | 3 | 1 | 5 |
| | | 4 | 1 | 4 |

Table 1: Prediction of authentic 5'-splice site sequences of Rat Chymotrypsin B gene by CDA method(a) and proposed algorithm(b)

I. level number
II. No. of authentic splice site sequences
III. No. of non-authentic splice site sequences

# References

[1] M. B.Shapiro and P. Senapathy, "RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression," *Nucleic Acids Reserch*, Vol. 15, pp. 7155-7175, 1987.

[2] Iida,Y., "DNA sequences and multivariate statistical analysis. Categorical discrimination approach to 5' splice site signals of mRNA precursors in higher eukaryotes' genes," *CABIOS*, Vol. 3, pp. 93-98, 1987.

[3] S. Khawaja, T. Nagashima and K. Ono, "A new algorithm for predicting splice site sequence based on an improvement of categorical discriminant analysis," *CABIOS*, Vol. 11, pp.349-359, 1995.