Detection of the Splicing Sites with Kernel Method Approaches Dealing with Nucleotide Doublets

Masaki Yamamura masaki@genome.ist.i.kyoto-u.ac.jp
Osamu Gotoh gotoh@i.kyoto-u.ac.jp

Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Keywords: splicing site, kernel methods, SVM

1 Introduction

The RNA-splicing is a very important phenomenon in eukaryotic cells to make pre mRNAs into mature forms. Despite its importance, biochemical details of the mechanism of RNA-splicing have not been understood completely. In this research, using machine learning approaches, we aimed to discriminate the true splice sites from pseudo sites which wouldn't be spliced. Support Vector Machine (SVM) is known as an efficient machine learning method used in many scientific disciplines including bioinformatics [3]. As several researchers have shown, DNA sequence data can be treated with adequate design of kernel functions i.e. core parts of SVM [1, 2, 4]. We focused on patterns of nucleotide doublets of DNA sequence and designed a kernel which might better represent the feature of sequence around splice sites. By examining more than 10,000 human splicing donor (5' ends of introns) and acceptor (3' ends of introns) sites, we found that the accuracies of this methods were 94.49% and 93.94% respectively. These results imply higher efficiency of this method than those observed with other kernel methods and methods based on first- and second-order Markov models. This method would reflect properties of sequence data and would give us insights for the mechanisms of the RNA-splicing.

2 Materials and Methods

The learning and test sets of DNA sequence data of splicing site were collected from Goldenpath database (UCSC Genome Bioinformatics website [5]). The true splicing sites were guaranteed by comparison of human whole genome sequences with cDNA sequences. The sequences of false sites were made up of GT/AG sites that appeared at the nearest positions from the true splicing sites and their neighboring bases of the same length as the true sites.

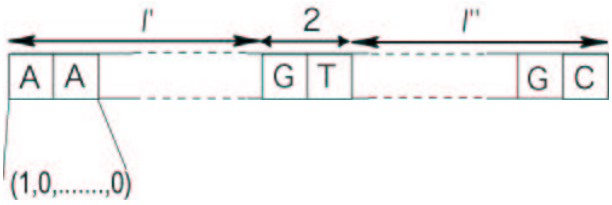

Figure 1: An example of learning and test sets of donor site. Letters in boxes mean a DNA sequence. The total length l is l'+l"+2. Pairs of bases in each position could be converted to 16-dimension binary vectors and therefore total dimension of a sequence is 16(l-1).

The kernel we used was designed with the pattern of nucleotide doublets of each position (Fig. 1). The number of combinations of 4 nucleotides, A, T, G and C, is 16 and the double on each position was expressed by a 16-dimensional binary vector. Compiling them from 5' side to 3' side, a 16(l-1) dimensional binary vector was obtained from a sequence whose total length was l. The numbers of l' and l" was changed from 5 to 15.

3 Results and Discussion

Table 1 shows the results of our method with the kernel mentioned in Material and Methods section. The results indicate that the efficiency of our kernel method clearly outperforms the Markov-model based method, which has been most widely used in gene-prediction programs and precedent research which used kernel methods similar to our methods but with the information of independent one nucleotide in each position [1]. The results would reflect the feature of the mechanism of the RNA splicing and we would get the clues for biological analysis.

Table 1: Comparison of the results of our methods, our precedent research with first- and second-order Markov model and precedent research.

Methods	sensitivity	specificity	accuracy
Our method			
Donor site	96.94%	92.03%	94.49%
Acceptor site	95.10%	92.77%	93.94%
Markov model			
Donor site	91.27%	91.27%	91.27%
Acceptor site	91.40%	91.40%	91.40%
Precedent research			
Donor site	91.63%	94.02%	92.85%
Acceptor site	86.00%	90.96%	88.48%

Accuracy is defined as the average of sensitivity and specificity.

References

[1] Sun, Y.-F., Fan, X.-D., and Li, Y.-D., Identifying splicing sites in eukaryotic RNA: support vector machine approach, *Computers in Biology and Medicine*, 33:17–29, 2003.

[2] Tsuda, K., Kin, T., and Asai, K., Marginalized kernels for biological sequences, *Bioinformatics*, 18:S268–S275, 2002.

[3] Vapnik, V., *Statistical Learning Theory*, Wiley, New York, 1998.

[4] Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T., and Muller, K.-R., Engineering support vector machine kernels that recognize translation initiation sites, *Bioinformatics*, 16(9):799–807, 2000.

[5] `http://genome.ucsc.edu/`