

Genomic Splicing Sites Prediction Algorithm Based on Nucleotide Sequence Pattern

Shu-Hung Lin, Kun-Nan Tsai, Chung-Ming Chen

Institute of Biomedical Engineering, National Taiwan University, Taipei, Taiwan

Keywords: splicing, splicing sites, sequence pattern, influenza virus

1 Introduction.

Splicing sites prediction plays a major role in genomic research of biomedical science. If splicing sites of genomic sequence can be correctly predicted, numerous crucial problems in biomedical area may be resolved naturally. Some typical examples are finding pathogenic factors of serious diseases and developing new gene therapy and so on. The major difficulty in developing the method of predicting the splicing sites lies in the diversity of species, which leads to substantially different sequence patterns of splicing sites among various species. For instance, the existing splicing sites prediction algorithms cannot effectively predict the splicing sites of influenza virus. For this reason, we propose a new genomic splicing sites prediction algorithm in this study, which is based on a data mining technology and many biomedical findings, to discover the candidate patterns of splicing sites in the gene sequences, to accurately predict splicing sites in the gene sequences, and to detect many new candidate splicing sites.

2 Materials and Methods.

The primary idea of the proposed splicing site prediction algorithm, i.e., predicting splicing sites based on nucleotide sequence pattern, originates from several observations on biology. In the splicing process, the splicing factors snRNA and protein constituting the spliceosomes must bind with nucleotide sequences first in order to compose the spliceosomes and carry out a splicing process. To ensure splicing at correct positions, it is assumed that some specific nucleotides co-occur at some specific positions in a nucleotide sequence, which renders sufficient chemical force to help splicing factors recognize and bind with those nucleotides. It is further assumed in this study that the nucleotide sequence may not need specific patterns. As long as a co-occurring nucleotide pattern can provide sufficient chemical force to help splicing factors recognizing the splicing sites and proceeding to the splicing process, it is considered as a valid and effective nucleotide pattern. Based on this assumption, it is believed that if all these patterns can be identified, including the positions and combinations, most splicing sites may be predicted accordingly. Nevertheless, this idea has not been fully exploited in previous splicing sites prediction algorithms. Limited by their underlying ideas, most previous algorithms cannot discover those patterns that are important enough but occur less frequently. As contrast, with a very high sensitivity while maintaining a reasonably high specificity, the proposed algorithm is capable of identifying effective nucleotide patterns with various occurrence frequencies, including those cannot be found by most previous algorithms. To identify the potential patterns more accurately, a technique based on binary transformation of four types of nucleotides, has been proposed to reinforce the idea of predicting splicing sites based on nucleotide sequence pattern. This technique evolves from the biological truth that two classes of nucleotides, i.e., purine and pyrimidine, are involved in the biological system. Moreover, to further improve the performance of the proposed algorithm, several biological properties, e.g., consensus sequence and branch point, etc., have been incorporated to account for the biological facts. Lastly, the correctness of predicted splicing sites is evaluated by translating the nucleotide sequences into amino acid sequences as well as the secondary and tertiary structures of protein. By comparing with the known amino acid sequences and protein structures, the validity of a splicing site may be further assessed.

The proposed algorithm is composed of six major steps as summarized in the following. (1) Multiple sequence alignment is first performed for training data to take account of genomic operations, including insertions, deletions, and replacements. (2) Binary transformation is carried out to transfer the nucleotide sequences of purine (C/T) and pyrimidine (A/G) into Y and R for acceptors. (3) Determine the consensus sequences comprising the nucleotides occurring most frequently in the specific vicinity of the splicing sites of the training data. (4) Construct the mining structure of the nucleotide sequence patterns using the consensus sequences of the training data. Taking a 3-mer consensus sequence, say AGT, as an example, Figure 1 illustrates the binary tree of the mining structure constructed by the proposed algorithm, where \bar{X} denotes "non-X", i.e., all types of nucleotides other than X. The number displayed below each pattern stands for the frequency of the pattern occurring in the training data. If the number of a node in this structure is smaller than a specified minimum support value, denoted as min_sup , the subtree rooted at this node is pruned. (5) Since each \bar{X} has three possibilities, a further analysis is required to determine the actual nucleotides at the position of \bar{X} . (6) Lastly, several biological findings, such as branch points, the topological characteristics of donors and acceptors, and so on, are incorporated to determine the final co-occurring nucleotide patterns as the basis of splicing site prediction.

3 Results and Conclusions

The proposed algorithm has been tested by leave-one-out cross validation on the human and drosophila data from BDGP and applied to the splicing sites prediction for influenza virus A from ISD (Influenza Sequence Database). With the characteristics of frequent mutation, it is reasonable to expect that new splicing sites might occur from time to time in influenza virus due to mutation. The consequence is obvious that new species of influenza virus might evolve and bring about a horrible prevalence threatening the health and safety of human beings. While none of the existing splicing sites prediction algorithms may effectively predict the splicing sites of influenza virus, the proposed algorithm has achieved prediction accuracy as high as 94%-100% for the splicing sites on the seventh and eighth segments of influenza virus A. Tables 1 and 2 summarize the prediction efficiencies of the proposed algorithm for these three species. Overall speaking, the performances of the proposed algorithm on Drosophila and Human are comparable with NNSPLICE and outperform several well-known algorithms, including NNSPLICE, SPLICEPREDICTOR, GENSCAN, and so forth.

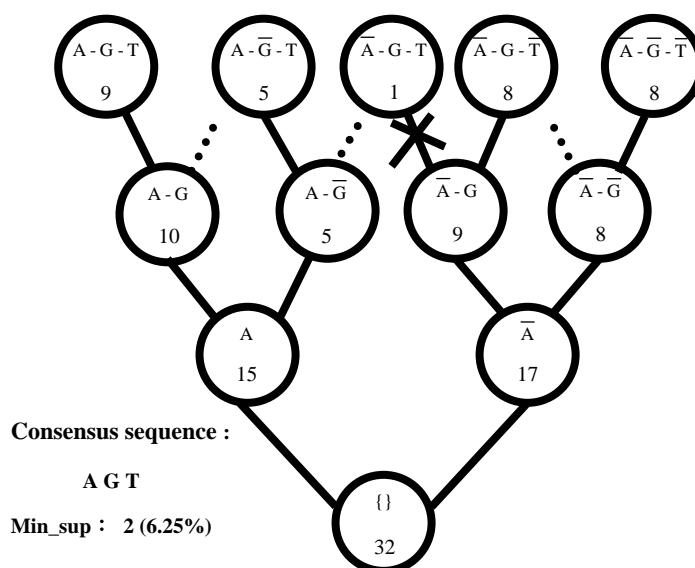


Figure 1: Mining structure of the proposed splicing sites prediction algorithm

Species		Sen(%)	Spe(%)	Acc(%)
IFA7	Donor	95.1	97.8	97.8
	Acceptor	100	94.2	94.3
IFA8	Donor	99.6	99.7	99.7
	Acceptor	100	95.3	95.3

IFA7,8: Influenza A segment7 and segment8
Sen: Sensitivity Spe: Specificity Acc: Accuracy

Species		Sen(%)	Spe(%)	Acc(%)
Dro	Donor	90.5	95.8	94.7
	Acceptor	85.5	87.2	86.8
Human	Donor	92.8	93.0	92.9
	Acceptor	89.4	88.2	88.4

Dro: Drosophila
Sen: Sensitivity Spe: Specificity Acc: Accuracy

4 References

- [1]Burge, C.B. et al. 1999. Splicing of precursors to mRNAs by the spliceosome. *In the RNA world II*, Cold Spring Harbor Laboratory Press, pp. 525-560.
- [2]Helen, P., Jiawei, H., Jian, P., Ke, W., Qiming, C. and Umeshwar, D. 2001. Multi-dimensional Sequential Pattern Mining. *CIKM*, pp. 5-10.
- [3]Parthasarathy, S., Zaki, M.J. 1999. Incremental and Interactive Sequence Mining. *CIKM*, pp. 251-258.
- [4]Rakesh, A. and Ramakrishnan, S. 1994. Fast Algorithms for Mining Association Rules. *Proceedings of the 20th VLDB Conference*
- [5]Reese, M.G., Eeckman, F.H., Kulp, D. and Haussler, D. 1997. Improved splice site detection in Genie. *J Comput Biol.*, 4, pp. 311-23.
- [6]Senapathy, P., Shapiro, M.B., Harris, N.L.1990.Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods Enzymol.*183, pp.252-278.
- [7]Staley, J.P. and Guthrie, C. 1998. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell*, 92, pp. 315-326.