

# Using enhancing signals to improve specificity of *ab initio* splice site sensors

Alexandre Tchourbanov, Hesham H. Ali

Department of Computer Science, College of information science and technology,  
University of Nebraska at Omaha, Omaha, NE 68182-0116  
achurbanov|hali@mail.unomaha.edu

Jitender Deogun

Department of Computer Science and Engineering, University of Nebraska-Lincoln  
Lincoln, NE 68588-0115, deogun@cse.unl.edu

## Abstract

*In this paper, we describe a new approach to improve the precision of splice site annotation in human genes. The problem is known to be extremely challenging since the human splice signals are highly indistinct and frequent cryptic sites confuse signal sensors. There is a strong evidence that Exonic Splicing Enhancers (ESE) and Exonic Splicing Silencers (ESS) influence commitment to splicing at early stages. We propose the use of a Naïve Bayesian Network (BN) combined with Boltzmann machine splice sites sensor, to improve the specificity of splice site prediction. The SpliceScan program is implemented to demonstrate feasibility of specificity enhancement based on ESE/ESS signals interactions. SpliceScan is more sensitive than GeneSplicer and NNSplice for the same specificity. The designed method is of particular value for ab initio gene annotation.*

*Key words:* ab initio methods, donor, acceptor, splicing enhancer, splicing silencer

## 1. Introduction

The precise removal of introns from pre-messenger RNAs (pre-mRNAs) by splicing is a critical step in expression of most metazoan genes. The process requires accurate recognition and pairing of 5' and 3' splice sites by the splicing machinery. Inappropriate splicing of a gene may result into the translation of a non-functional protein.

Considerable progress has been made in the understanding of spliceosome assembly [2]. Weakly conserved splice signals are necessary, but not sufficient, for the exact recognition of an exons. Frequently degenerate donor, acceptor,

polypyrimidine and the branch point motifs provide insufficient information for the exact splice sites detection [3].

## 2. Splicing Enhancers and Silencers

Specificity in the splicing process derives partly from sequences other than splice-site signals, including Exonic Splicing Enhancer (ESE) and Exonic Splicing Silencer (ESS) signals [2]. There are 10 Serine/aRginine-rich (SR) Splicing Enhancer proteins known today: SRp20, SC35, SRp46, SRp54, SRp30c, SF2/ASF, SRp40, SRp55, SRp75, 9G8 and approximately 20 hnRNP Splicing Silencing factors, among them, the most studied, hnRNP A1 complex. Tra2 $\beta$  is reported being SR Splicing regulator.

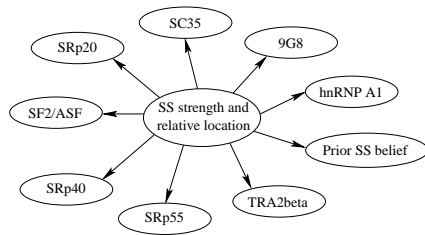
Together, with inefficient splice-site signals, the appropriate balance of ESE and ESS elements somehow allows fine tuning of the splicing mechanism [7].

However, it is most likely that in any system the exonic and intronic definition elements are functioning together for specific splicing [6].

The complexity of constitutive and alternative splice site recognition suggests multiple layers of regulation, with each layer being the result of combinatorial arrays of elements and factors.

## 3. The proposed approach

In order to classify splice sites being cryptic or real, in our approach we use Boltzmann machine for splice site finding. We use the Naïve Bayes Model to combine ESE/ESS evidence collected from the context of a putative donor or acceptor, since the dependencies seem to be linear [1]. The network topology is shown in Figure 1.



**Figure 1. Combining the evidence behind a splice site**

This factoring is equivalent to the sum of LODs. We can change threshold value to adjust number of False Negatives (FN) and true positives (TP) in posterior result.

#### 4. Implementation results

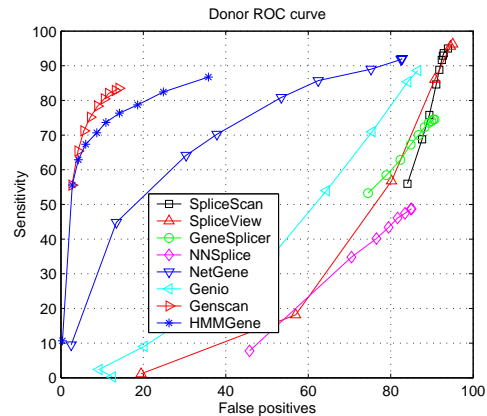
Based on the discussed principles, we implemented a new splice sites prediction algorithm called *SpliceScan*. We compiled the test set out of 250 human genes shorter than 100,000 nt. The results of the comparative study of *SpliceScan* and other programs are shown in Figure 2

We demonstrate clear prediction improvement of our approach in terms of Sensitivity for the same rate of false positives when compared to *GeneSplicer* [4] and *NNSplice* [5]. It performs very similar to *SpliceView* program, the most sensitive program we have evaluated. For high specificity values *GenScan* performs best up to 85% sensitivity, according to our tests. Since 15% of the splice sites are not recoverable with *GenScan*, we have to rely on other software, including our *SpliceScan* program, to spot the remaining splice sites. Currently, we are employing additional filtering mechanisms to improve on false positives ratio.

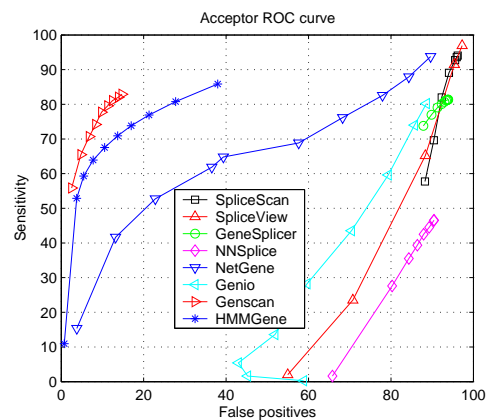
The *GeneSplicer* program, the test set and the results are freely available from <http://bioinformatics.ist.unomaha.edu/~achurban/>.

#### References

- [1] B. R. Graveley, K. J. Hertel, and T. Maniatis. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. *The EMBO journal*, 17(22):6747–6756, 1998.
- [2] M. L. Hastings and A. R. Krainer. Pre-mRNA splicing in the new millenium. *Current opinion in Cell Biology*, 13:302–309, 2001.
- [3] L. P. Lim and C. B. Burge. A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences*, 98(20):11193–11198, 2001.



(a) Performance comparison of *SpliceScan* versus other *ab initio* programs on Donor Splice Sites



(b) Performance comparison of *SpliceScan* versus other *ab initio* programs on Acceptor Splice Sites

**Figure 2. *spliceScan* performance versus other programs in terms of Receiver Operating Characteristic (ROC) curves**

- [4] M. Pertea, X. Lin, and S. L. Salzberg. *GenSplicer*: a new computational method for splice site prediction. *Nucleic acids research*, 29(5):1185–1190, 2001.
- [5] M. Reese and F. Eeckman. Splice sites: A detailed neural network study. In D. Bentley, E. Green, and P. Hieter, editors, *Proceedings of the 1996 Genome Mapping and Sequencing Meeting*. Cold Spring Harbour, New York, 1996.
- [6] V. Rooke, Nanette ans Markovtsov, E. Cavagi, and D. L. Black. Roles for SR proteins and hnRNP A1 in the regulation of c-src exon N1. *Molecular and cellular biology*, 23(6):1874–1884, Mar. 2003.
- [7] J. Zhu, A. Mayeda, and A. R. Krainer. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Molecular and cellular biology*, 8:1351–1361, 2001.