

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

Alexandre Tchourbanov*, Hesham H. Ali* and Jitender Deogun^o

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

^o 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 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 outperforms recent GeneSplicer program for low false negatives. The designed method is of particular value for *ab initio* gene annotation.

Introduction

Biology has entered genomic era. The accuracy and speed with which genes can be predicted is still far from satisfactory. Only 45% exons are predicted right, which requires costly cDNA alignment for further gene detailization. Exact exonic boundaries finding (5' and 3' splice sites) is essential first step for any *ab initio* gene annotation process. Here we introduce a method to improve prediction quality of the splice sites.

Biologically, the precise removal of introns from pre-messenger RNAs (pre-mRNAs) by *spliceosome* is a critical step in expression of most metazoan genes.

Spliceosome is a complex snRNA - protein assembly within which the splicing reactions occurs. An ordinary active spliceosome consists of five small nuclear RNAs (snRNAs) (U1,U2,U4,U5 and U6), more than 50 proteins, a supramolecular assembly that is nearly as complex as ribosome. The spliceosome acts through a multitude of RNA-RNA, RNA-protein and protein-protein interactions to precisely cut out each intron and join the exons in the correct order.

Weakly conserved splice signals are necessary, but not sufficient, for the exact recognition of the exons.

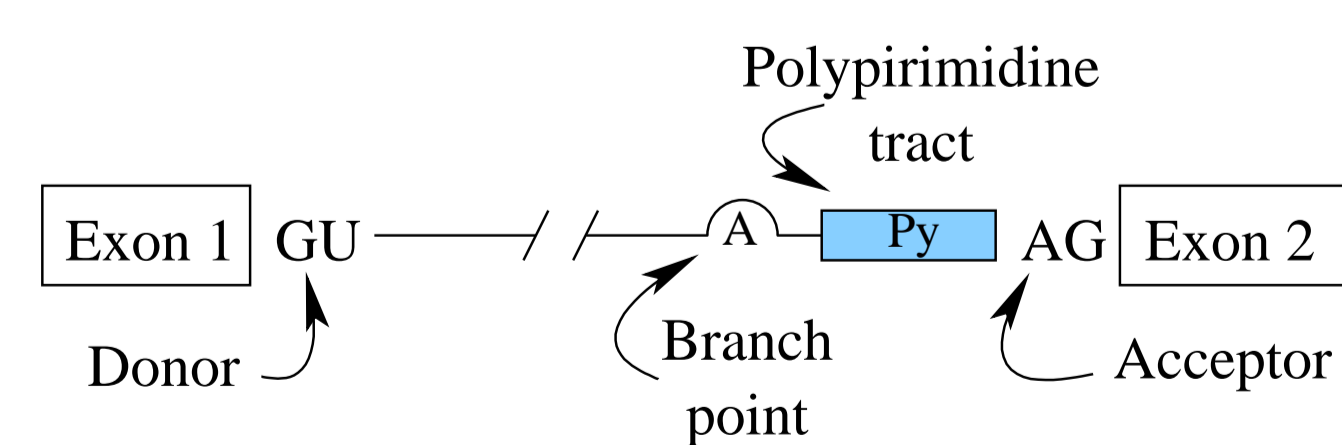
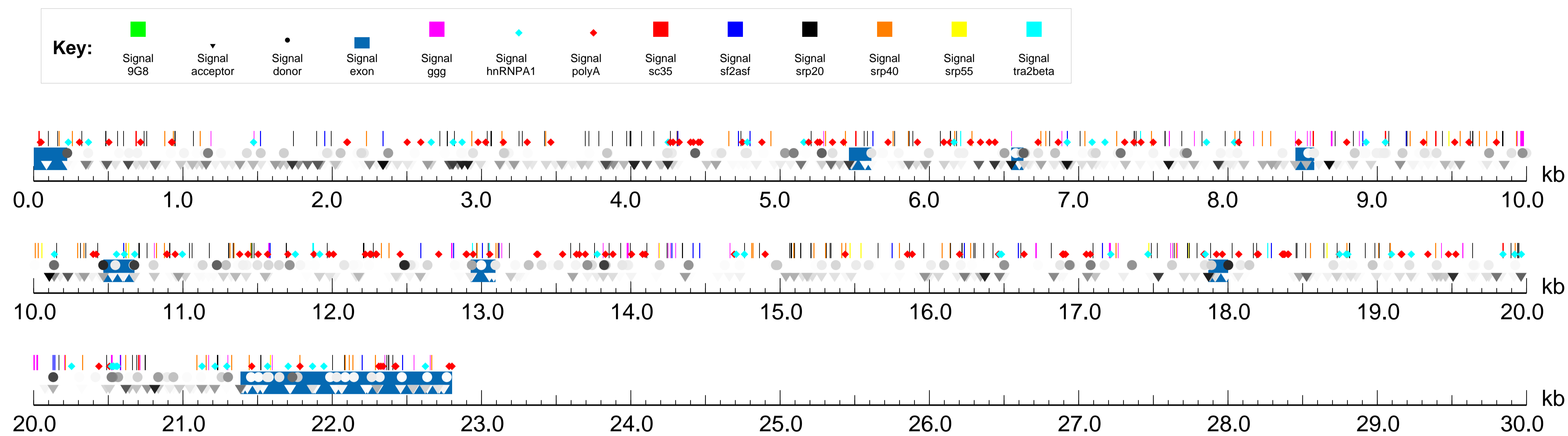


FIGURE 1: Structure of an intron

ESE and ESS signals

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.

GIGOGene 1.0 predicted gene structure for gi|14211894|ref|NM_032578.1



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. Tra2β 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. 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.

Spliceosome - ESE/ESS interactions

There are several interactions have been reported by researchers:

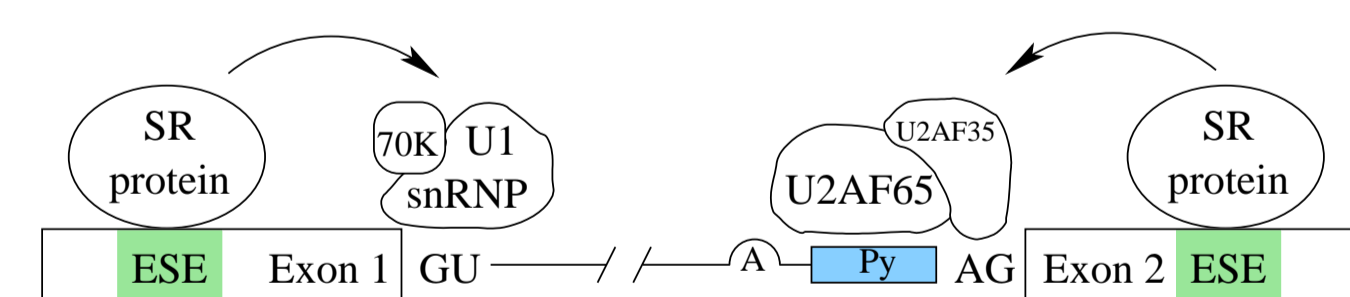


FIGURE 2: Splice sites recognition

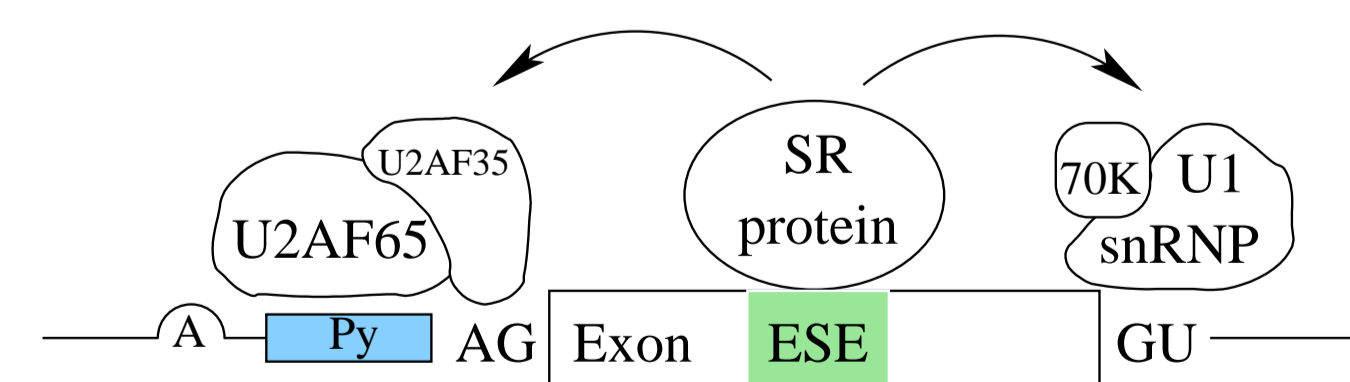


FIGURE 3: Exon definition model

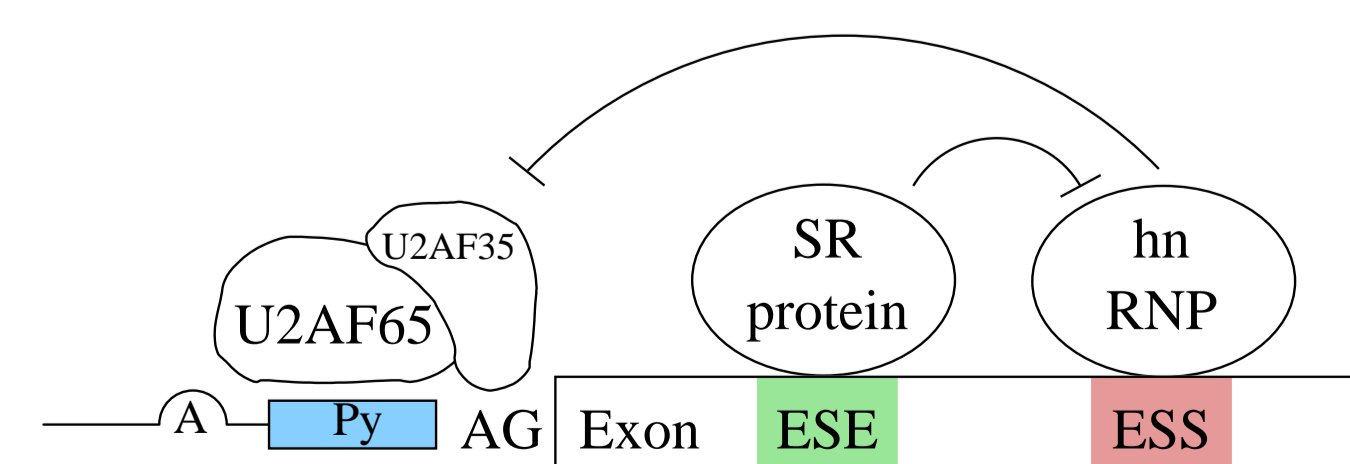


FIGURE 4: ESS-ESE antagonism

Rules for splicing mechanism

- ESEs and ESSes are frequently located in downstream exons [4];
- The precise mechanism by which hnRNP A1 bind ESS in upstream exon and represses splicing of the upstream intron remains unknown [5];
- Most splicing enhancers are located within 100 nucleotides of the 3' splice site and are not active further away [2];
- Each enhancer complex assembles independently for 3' and 5' sites and there is a minor interaction across an intron [3];
- Based on current views of exon definition, each exon should be recognized by the splicing machinery as an independent unit [3];
- Analysis of the experimental data revealed that the splicing efficiency is directly proportional to the calculated probability of a direct interaction between the enhancer complex and the 3' splice site:
 - Strong natural enhancers function at a greater distance from the intron than weak natural enhancers [4];
 - The closer ESE is located to a splice site, the more efficient it is [2];
 - Multiple enhancer sites increase the probability of splicing activating [2];
 - Strong ESS sites may suppress an effect from ESE(s) located upstream [5].

The proposed approach

We use Boltzmann machine for recognition of the splice sites. To get training set we ran our gene annotation utility GIGOGene on the whole RefSeq (Jan. 2003).

The ESE/ESS sites are detected according to the following consensus [1]:

Protein	High-affinity binding site	Functional ESE
SRp20	WCWVC CUKUCY	GCUCUCUCC CCUCGUCC
SC35	AGSAGAGUA GWUWCCUGCUA GGUUAUCUG GAGCAGUAGKS GUUCGAGUA UGUUCSAGWU AGGAGAU	GYYMCRY UGCYGY
9G8	(GAC) ⁿ ACGAGAGAY WGGACRA	
SF2/ASF	RGAAGAAC AGGACRRAGC	CRMSGW
SRp40	UGGGAGCRGUYRGCCGY	YRCRKM
SRp55		YYWCWSG
TRA2β	(GAA) ⁿ	
hnRNP A1	UAGGGW	

TABLE 1: Nucleotide symbols used: M → (A/C), R → (A/G), W → (A/U), Y → (C/U), S → (C/G), K → (G/U).

For the classification we use Naïve Bayes model:

$$P(SS, ESE_1, D_1, \dots, ESE_n, D_n) \sim P(SS) \cdot \prod_{i=1}^n P(ESE_i | SS, D_i) P(D_i)$$

Here N is the number of signals in the context (± 400), SS is a splice site strength and D_i is distance between the splice site and a certain Enhancer/Silencer ESE_i .

This factoring is equivalent to the sum of LODs.

Results

Some of the biases we use:

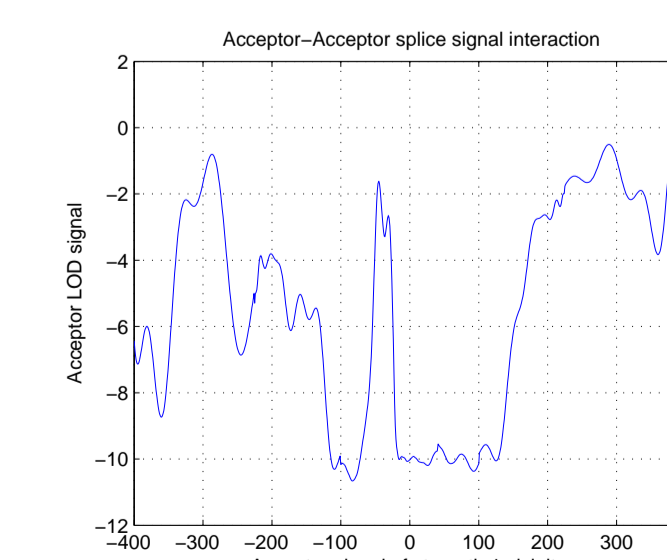


FIGURE 5: Acceptor-Acceptor interaction

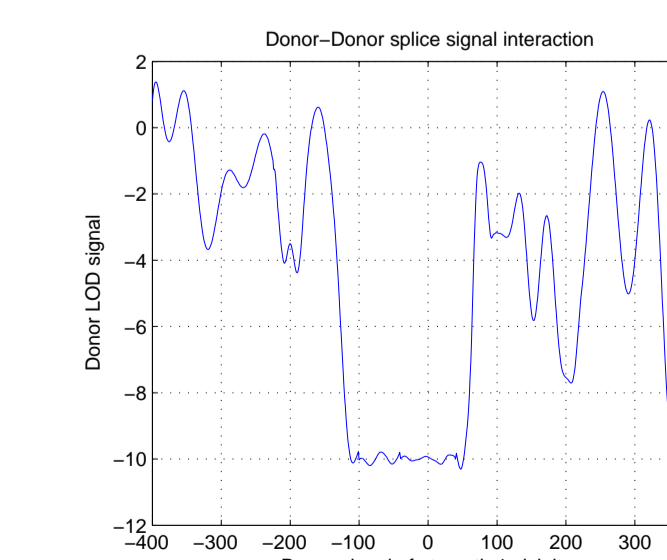


FIGURE 6: Donor-Donor interaction

Weak Donor and Acceptor sites prefer not to have strong cryptic neighbors. The biases also point at direction of scanning by commitment complex from inside an exon.

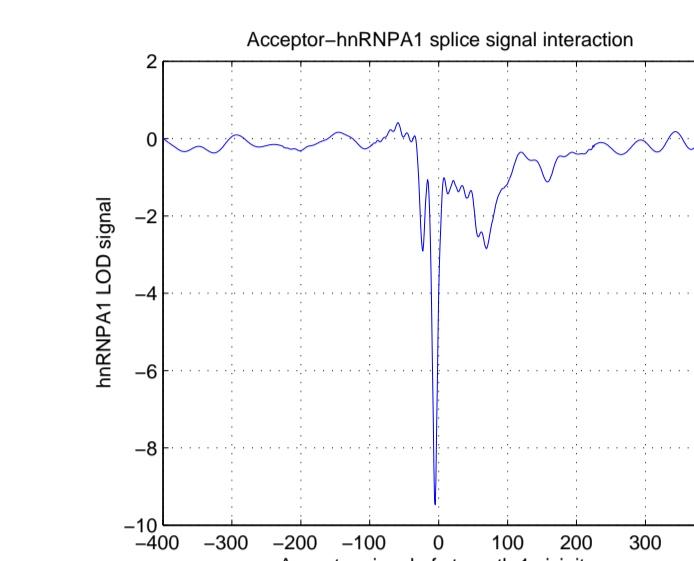


FIGURE 7: Acceptor-hnRNP A1 interaction

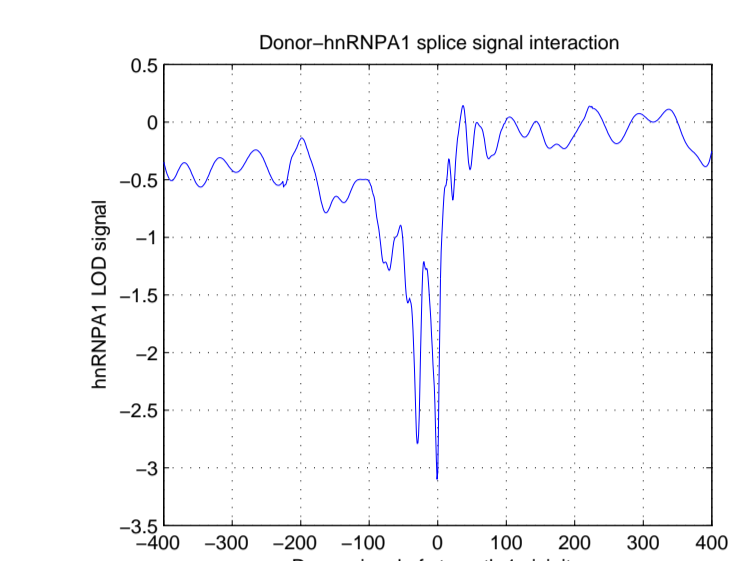


FIGURE 8: Donor-hnRNP A1 interaction

Compare to cryptic sites, real sites vicinity is depleted of hnRNP A1 silencers.

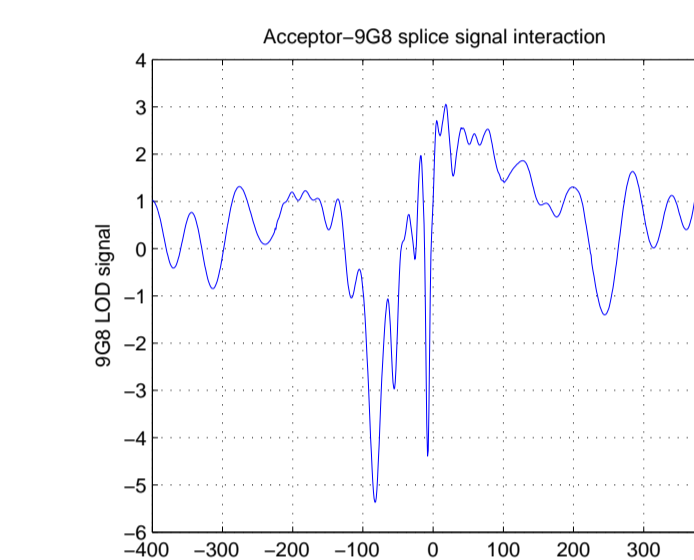


FIGURE 9: Acceptor-9G8 interaction

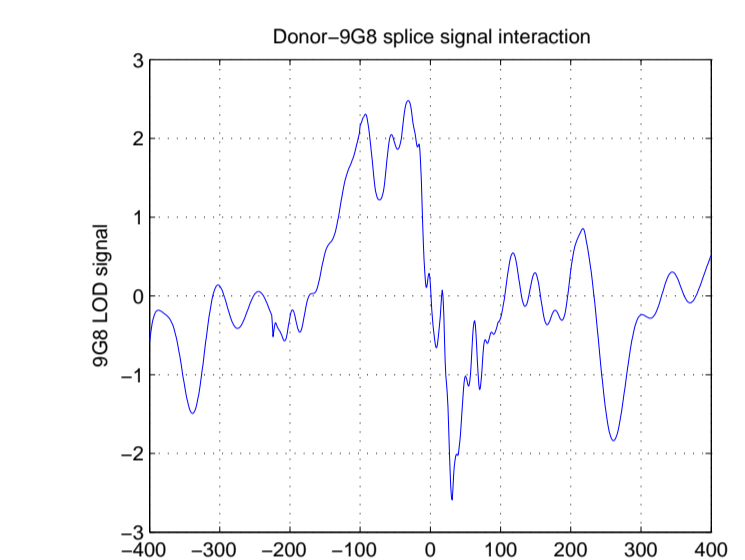


FIGURE 10: Donor-9G8 interaction

Compare to cryptic sites, real sites vicinity is rich of 9G8 enhancers.

We demonstrate the prediction improvement compared to GeneSplicer; a recent splice-site detection program reported to perform favorably when compared to NetPlantGene, NetGene2, HSPL, NNSplice, GENIO and SpliceView.

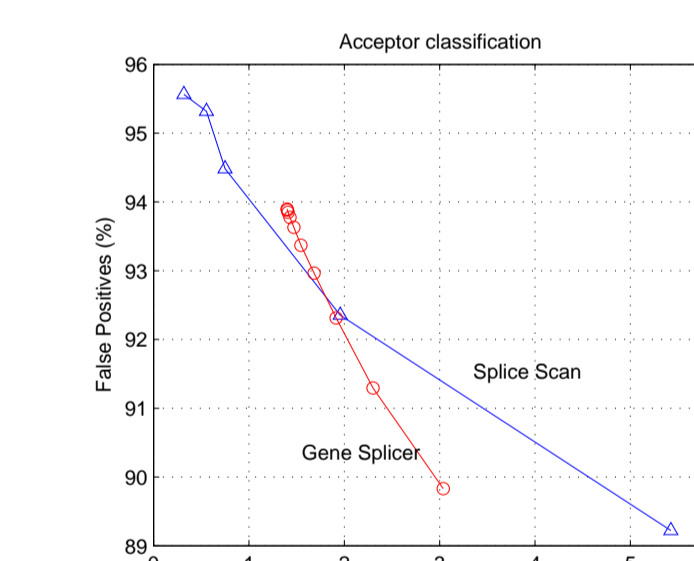


FIGURE 11: Acceptor classification improvement

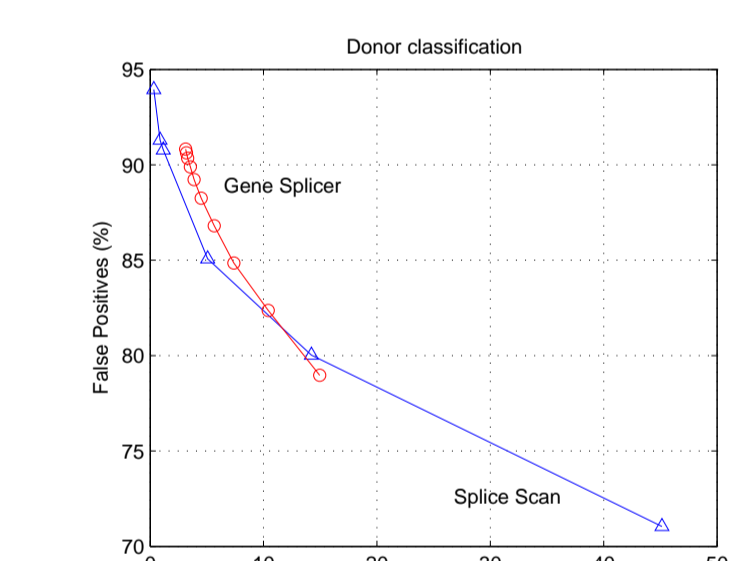


FIGURE 12: Donor classification improvement

Future work

We plan to improve our method by considering *de novo* motifs combined with digrammatical analysis of true exons vicinity.

References

- [1] Luca Cartegni, Shern L. Chew, and Andrian R. Krainer, *Listening to silence and understanding nonsense: exonic mutations that affect splicing*, Nature Genetics Reviews 3 (2002), 285–298.
- [2] Brenton R. Graveley, Klemens J. Hertel, and Tom Maniatis, *A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers*, The EMBO journal 17 (1998), no. 22, 6747–6756.
- [3] Bianca J. Lam and Klemens J. Hertel, *A general role for splicing enhancers in exon definition*, RNA 8 (2002), 1233–1241.
- [4] Akila Mayeda, Gavin R. Sreaton, Sharon D. Chandler, Xiang-Dong Fu, and Andrian R. Krainer, *Substrate specificities of SR proteins in constitutive splicing are determined by their RNA recognition motifs and composite pre-mRNA exonic elements*, Molecular and cellular Biology 19 (1999), no. 3, 1853–1863.
- [5] Jun Zhu, Akila Mayeda, and Andrian 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 (2001), 1351–1361.