Combinatorial method of splice sites prediction

Alexander Churbanov, Hesham 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

Abstract

Predicting and proper ranking of splice sites (SS) is a challenging problem in bioinformatics and machine learning communities. Proposed method of donor and acceptor SSs prediction is based on counting oligonucleotide frequencies for splice and splice-like signals. Based on bayesian principle SS sensors were built. We demonstrate advantage of our proposed sensor design compared with existing sensors and tools. In particular, our donor sensor outperforms Maximum Entropy Sensor for several representative test sets of genes when compared on Receiver Operating Characteristic (ROC) curve. We represent combinatorial interaction of SSs and related factors with Logarithm Of oDds (LOD) weight matrices. Based on factor interactions we were able to substantially improve splice signals prediction quality and rank SSs better than `SpliceView`, `GeneSplicer`, `NNSplice` *and* `Genio` *tools. Proposed method is used in our new splicing simulator* `SpliceScan`.

1 Introduction

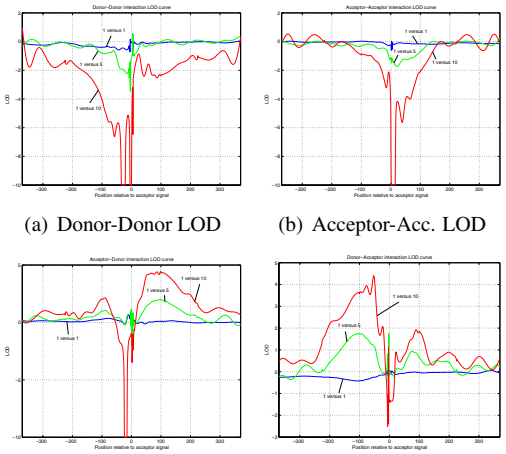
The precise removal of introns from pre-messenger RNAs (pre-mRNAs) by splicing is a critical step in expression of most metazoan genes. The process requires accurate recognition and pairing of $5'$ and $3'$ SSs by the splicing machinery. Inappropriate splicing of a gene may result into the translation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

Weakly conserved SSs are necessary, but not sufficient, for the exact recognition of the exons. Frequently degenerate donor, acceptor, polypyrimidine and the branch point motifs provide insufficient information for the exact SSs detection.

Correct prediction of SSs appear to be the key ingredient to successful *ab initio* gene annotation, since dynamic programming procedures have to see all the exon/intron boundaries in order to find the optimal solution [1]. The most sensitive sensor design predicting the least amount of false positives is preferable. Another good feature of a SS sensor is ability to rank predicted SSs, i.e. assign certain score characterizing importance or strength of a putative site of splicing.

2 Proposed design

There were numerous SS sensor designs proposed, among the best is Maximum Entropy Sensor [3]. Our sensor design is based on 7-mer oligonucleotide counting in splice and splice-like signals, with placement of 7-mers within consensus similar to Maximum Entropy Sensor `http://genes.mit.edu/burgelab/maxent/`.



(a) Donor-Donor LOD (b) Acceptor-Acc. LOD

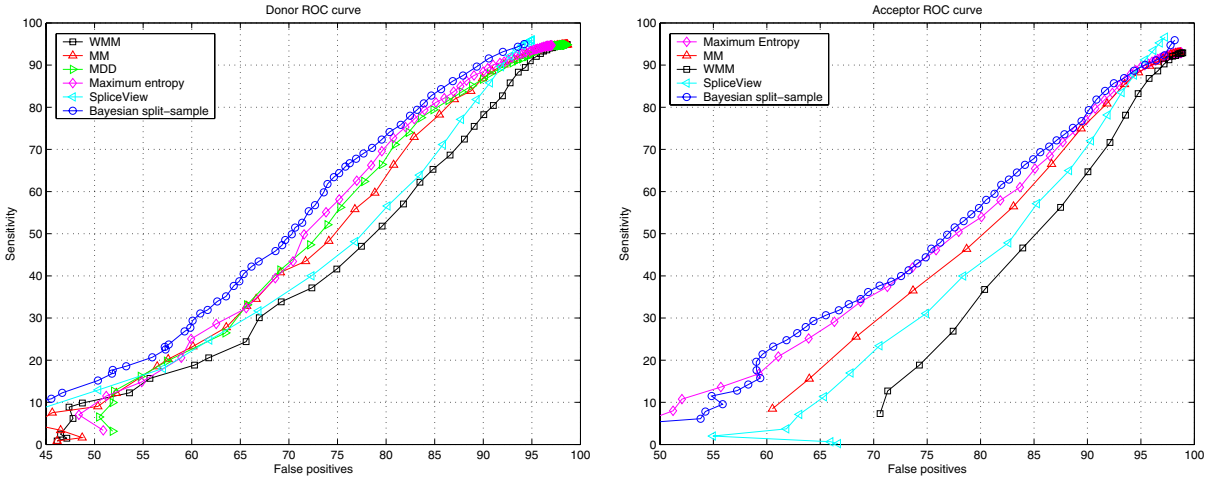
(c) Acceptor-Donor LOD (d) Donor-Acceptor LOD

Figure 1. LOD diagrams for splice sites interactions

We used our `GIGOgene` [2] tool to collect extensive learning set of predicted human and mouse gene structures.

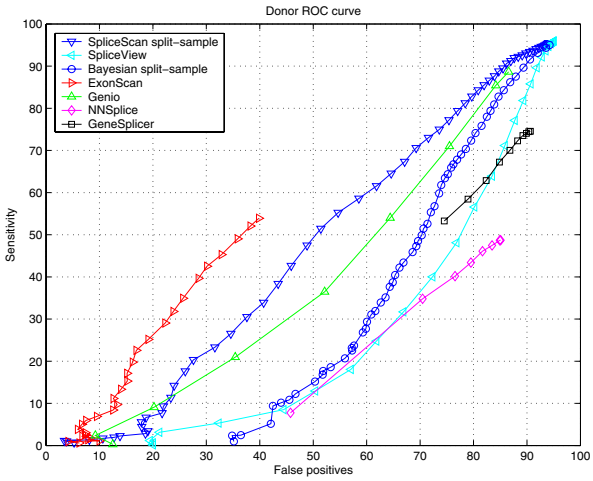
Based on collected oligonucleotide frequencies, we can evaluate probability of a SS given an oligonucleotide.

$$P(ss|oligo) = \frac{P(ss) \times P(oligo|ss)}{P(ss) \times P(oligo|ss) + P(\neg ss) \times P(oligo|\neg ss)}$$

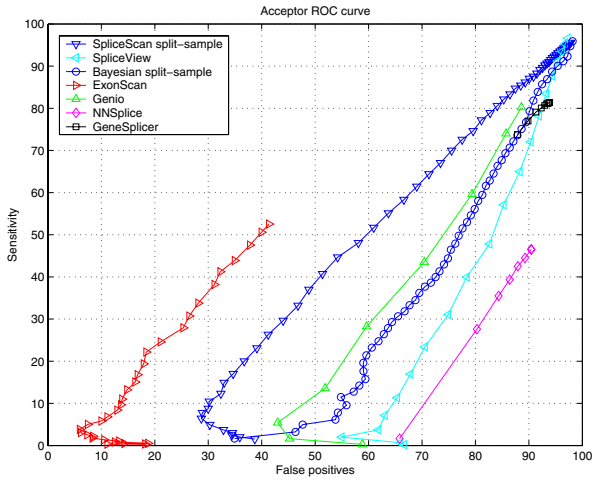


(a) Sensor ROC diagram for $5'$ splice site

(b) Sensor ROC diagram for $3'$ splice site



(c) Applications ROC diagram for $5'$ splice site



(d) Applications ROC diagram for $3'$ splice site

Figure 2. ROC diagrams for Donor and Acceptor signals

Using the learning set, we evaluated SS interactions for signals of different strengths (on scale 1-10) and interpolated normalized signal concentration ratios $\log_2\left(\frac{\text{splice}}{\text{splice-like}}\right)$ to get LOD diagrams, as shown in Figure 1. We incorporated biases found into our new `SpliceScan` tool.

3. Results

We tested performance of our *Bayesian* sensor and `SpliceScan` on 250 multi-exon annotated human genes that were specifically excluded from the learning set. We use Receiver Operating Characteristic (ROC) to compare performance of different sensors and tools, as shown in Figure 2. Program, learning set and test results are available at `http://bioinformatics.ist.unomaha.edu/~achurban/`.

References

[1] A. Krogh. Gene finding: putting the parts together. In M. J. Bishop, editor, *Guide to Human Genome Computing*, chapter 11, pages 261–274. Academic Press, San Diego, CA, 2 edition, 1998.

[2] A. Tchourbanov, D. Quest, H. Ali, M. Pauley, and R. Norgren. A new approach for gene annotation using unambiguous sequence joining. In *Proceedings of the Computational Systems Bioinformatics (CSB'03)*, pages 353–362. IEEE Computer society, Aug. 2003.

[3] G. Yeo and C. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*, 11(2):377–394, 2004.