Splam: a deep-learning-based splice site predictor that improves spliced alignments

Kuan-Hao Chao[1,2,*], Alan Mao[1,2,3], Steven Salzberg[1,2,3,4,*], Mihaela Pertea[2,3,*]

[1]Department of Computer Science, [2]Center for Computational Biology, [3]Department of Biomedical Engineering, [4]Department of Biostatistics, Johns Hopkins University

Johns Hopkins University, Whiting School of Engineering, Department of Computer Science

Johns Hopkins University, Center for Computational Biology (CCB)

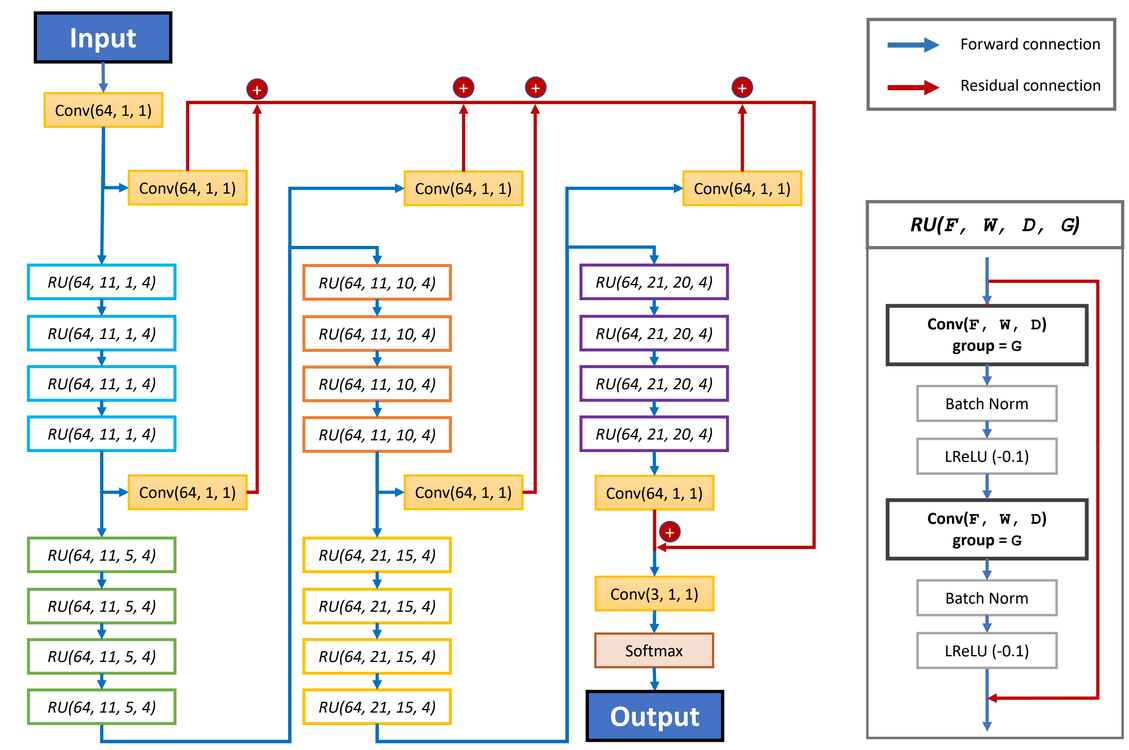
github.com/Kuanhao-Chao/splam ccb.jhu.edu/splam kh.chao@cs.jhu.edu, salzberg@jhu.edu, mpertea@jhu.edu https://khchao.com

Introduction

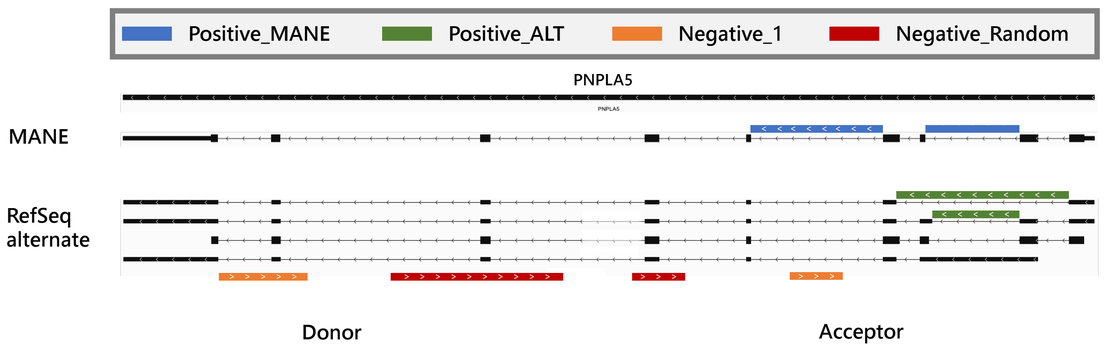
1. Splam introduces the idea of training a neural network on donor and acceptor pairs together, inspired by the splicing machinery itself, which recognizes both ends of each intron at the same time.
2. Splam uses a limited window of 400 bp flanking each splice site, again motivated by the biological process of splicing, which relies primarily on signals within this window
3. Splam recognizes splice sites from genomic sequence alone more accurately than existing methods.
4. Splam can improve the accuracy of transcript assemblies by removing spurious alignments produced by spliced aligners.

Methods

Model Architecture



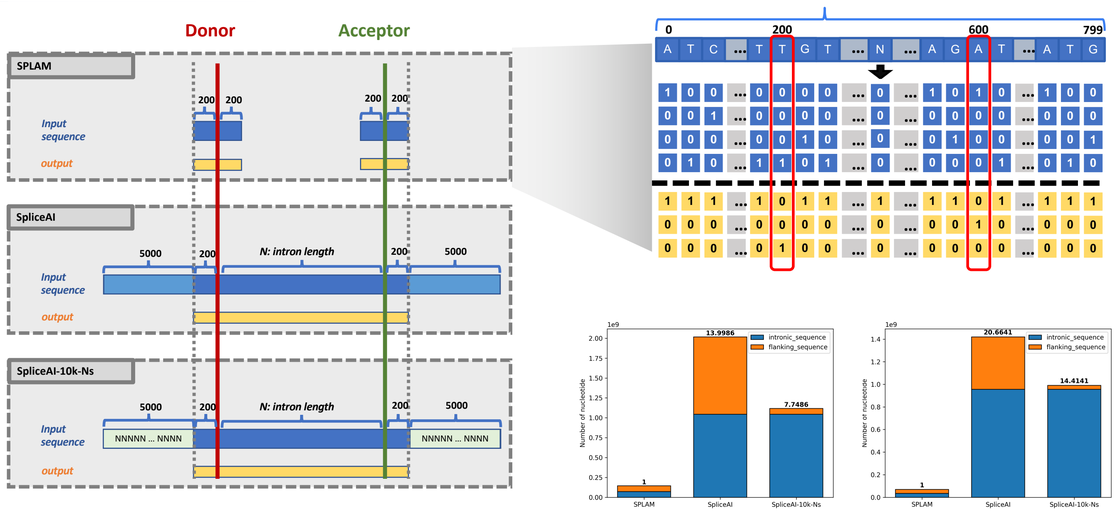
Training & testing dataset creation



Positive-MANE sites are selected from the MANE database with ≥100 alignments. **Positive-Alt** sites are in RefSeq but not in MANE, also with ≥100 alignments.

Negative-1 sites occur on the opposite gene strand, supported by only 1 alignment. **Negative-Random** sites are random GT-AG pairs on the opposite strand, not overlapping with known sites and lacking alignment support.

Data encoding



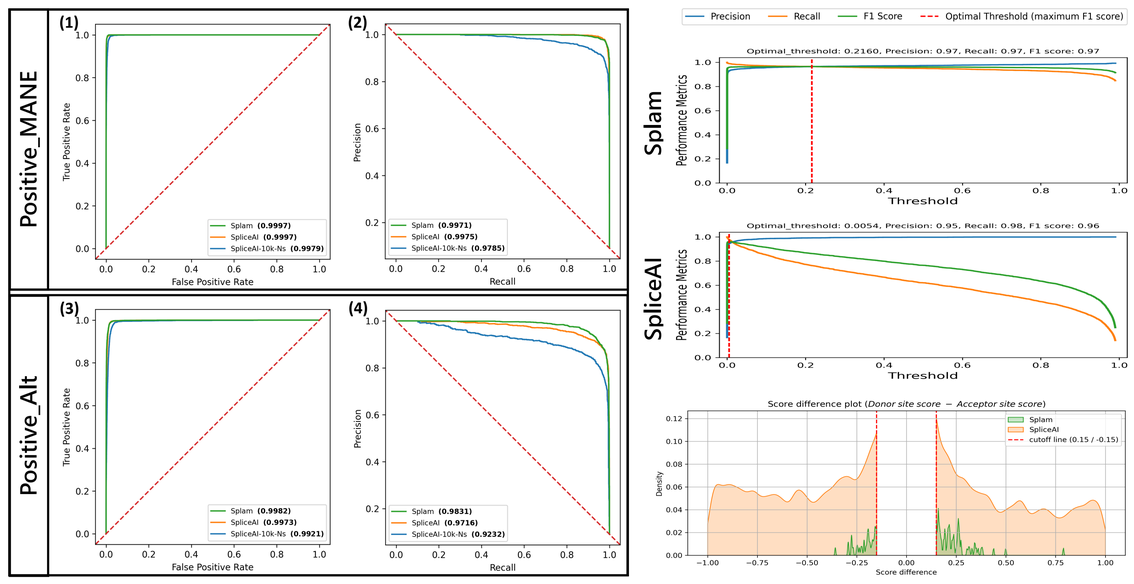
Splam's input (top row) uses 400nt flanking the donor site and another 400nt flanking the acceptor site. The output is labels for the 800nt region (shown in yellow).

SpliceAI's input (second row) follows its standard configuration, using 200nt upstream and downstream of the donor and acceptor sites, the entire intron, and 10Kb of flanking sequences.

SpliceAI-10k-Ns (third row) uses a modified input, replacing the 5Kb flanking sequences with Ns.

Results

Accuracy comparisons on human splice junctions in test dataset

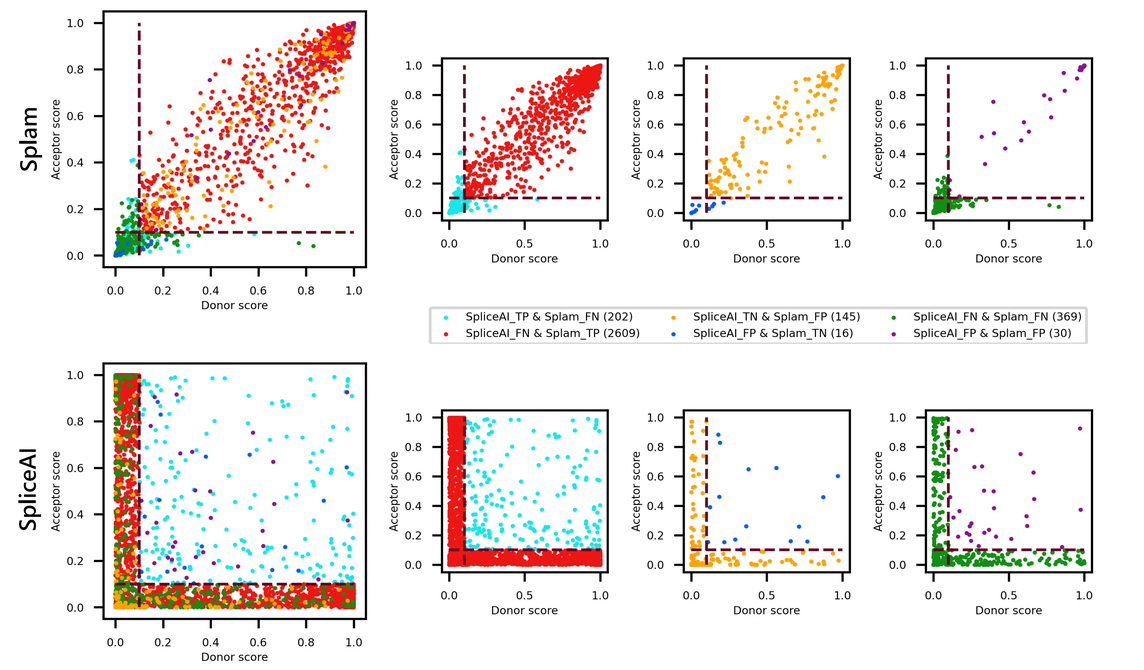


ROC / PR curves results are shown at the junction level, where the junction score is determined by the minimum of its donor and acceptor scores.

Discrimination threshold plots show the precision (blue curve), recall (orange curve), and F1 score (green curve) calculated at different thresholds. The optimal threshold (maximum F1 score) is indicated by a red dashed line.

Kernel density plot visualizes the differences between donor and acceptor scores (donor score − acceptor score)

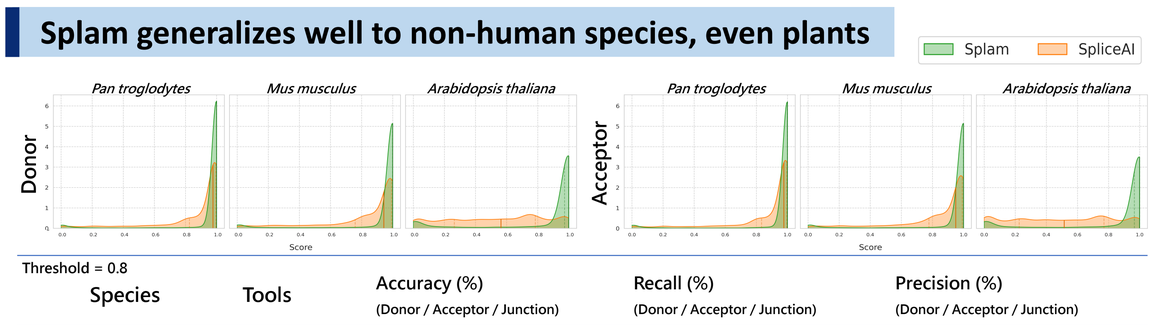
Comparing the score distributions of Splam and SpliceAI



Each dot represents a splice junction. The red dashed lines indicate the 0.1 cutoff threshold for labeling splice sites as true positives (TPs), true negatives (TNs), false positives (FPs), or false negatives (FNs).

Subplots in the second & third columns show cases where one program was correct while the other was incorrect (TP and FN, or FP and TN); the fourth column shows cases where both programs made incorrect predictions (FP and FN).

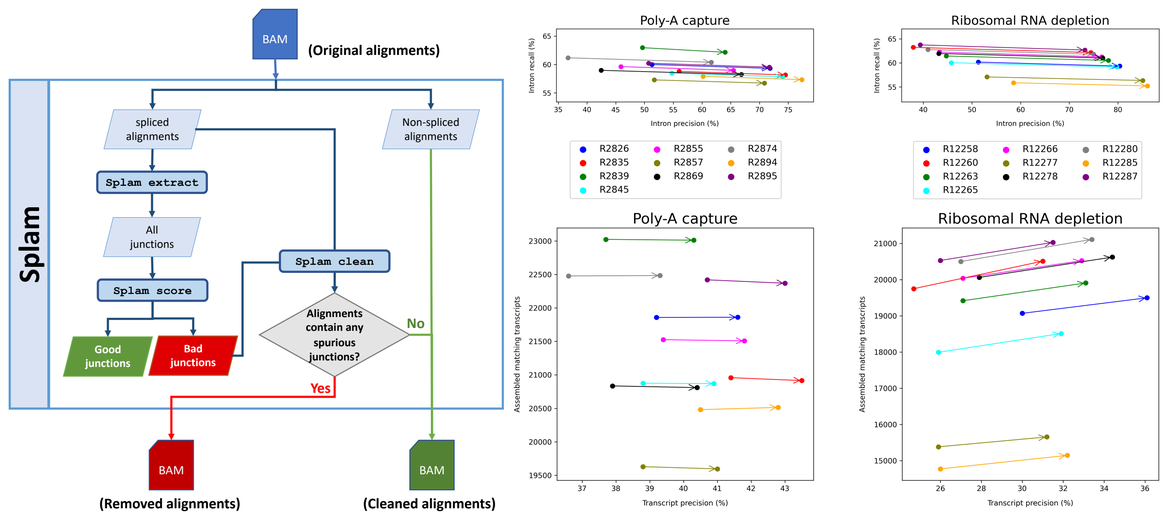
Splam generalizes well to non-human species, even plants



Threshold = 0.8

Species	Tools	Accuracy (%) (Donor / Acceptor / Junction)	Recall (%) (Donor / Acceptor / Junction)	Precision (%) (Donor / Acceptor / Junction)
Pan troglodytes (chimpanzee)	Splam	95.7 / 95.7 / 95.5	91.4 / 91.4 / 91.0	99.9 / 99.9 / 99.9
	SpliceAI	84.6 / 86.0 / 78.7	69.2 / 72.0 / 57.3	99.9 / 99.9 / 99.9
Mus musculus (house mouse)	Splam	93.6 / 93.7 / 93.3	87.3 / 87.4 / 86.6	99.9 / 99.9 / 99.9
	SpliceAI	80.0 / 82.1 / 73.6	59.7 / 64.2 / 47.3	99.9 / 99.9 / 99.9
Arabidopsis thaliana	Splam	90.7 / 90.4 / 90.0	81.4 / 80.9 / 80.2	99.9 / 99.9 / 99.9
	SpliceAI	68.2 / 67.8 / 60.0	36.3 / 35.7 / 20.0	99.8 / 99.9 / 99.9

Filtering low-scoring spliced alignments improves transcriptome assembly



Acknowledgements: supported by NIH under grants R01-HG006677 and R35-GM130151 and by NSF under grants IOS-1744309 and DBI-1759518.