OpenSpliceAI and Splam: Enhancing Splice Site Prediction Across Species to Improve Transcriptome Assembly

Kuan-Hao Chao[1, 2, *, †], Alan Mao[1, 2, 3, †], Anqi Liu[1], Steven L Salzberg[1, 2, 3, 4, *], and Mihaela Pertea[1, 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 Whiting School of Engineering — Department of Computer Science

Johns Hopkins University Center for Computational Biology — CCB

kh.chao@cs.jhu.edu

Genome Biology

Search: Splam

https://ccb.jhu.edu/splam/

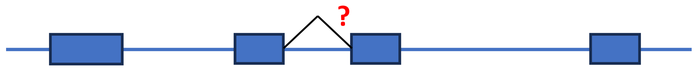
https://github.com/Kuanhao-Chao/splam

https://khchao.com

Introduction

SPLAM

- **Problem 1:** Given a splice junction, is it a good splice junction?



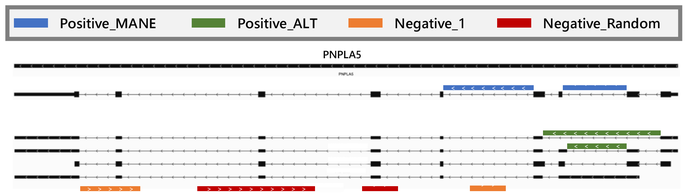
OpenSpliceAI

- **Problem 2:** Given a DNA sequence, where are the splice sites?

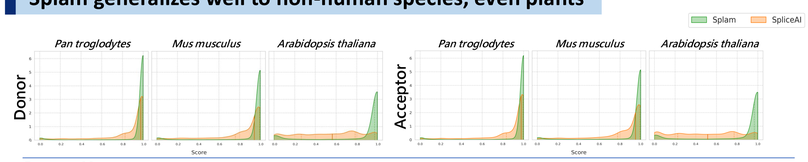
Splam

Training & testing datasets creation

- **Positive-MANE** sites are from the MANE database with ≥100 alignments.
- **Positive-Alt sites** are in RefSeq (not MANE) with ≥100 alignments.
- **Negative**-1 sites occur on the opposite strand with only 1 alignment.
- **Negative-Random** sites are random GT-AG pairs on the opposite strand, non-overlapping with known sites and without alignment support.



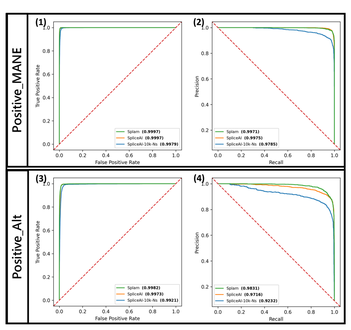
Splam generalizes well to non-human species, even plants

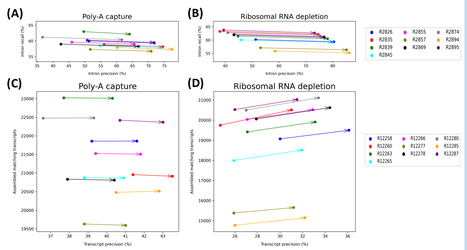


Threshold = 0.8

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





Results comparing RefSeq introns with splice junctions from 10 poly-A capture **(A)** and 10 rRNA depletion **(B)** RNA-seq samples, with original and Splam-cleaned samples connected by arrows. Panels **(C)** and **(D)** compare RefSeq transcripts with assembled transcripts from poly-A capture **(C)** and rRNA depletion **(D)**, also before and after Splam cleanup. The x-axis shows precision (percentage matched), while the y-axis shows intron recall in **A,B**, and matched transcript count in **C,D**.

OpenSpliceAI

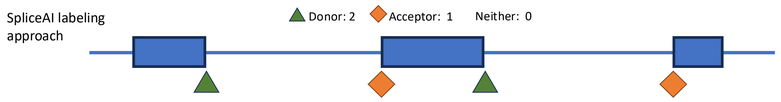
OpenSpliceAI is an open-source version of the SpliceAI program[1], a highly accurate splice site prediction system. OpenSpliceAI uses the newer PyTorch ML package instead of the older, slower Keras package, but otherwise uses the same code and is intended to replicate SpliceAI and allow users to re-train on their own species of interest.

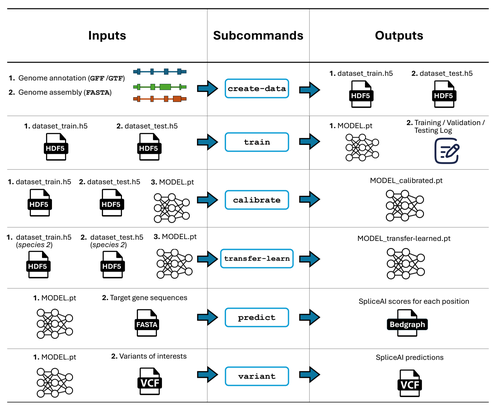
1. Jaganathan, Kishore, et al. "Predicting splicing from primary sequence with deep learning." Cell 176.3 (2019): 535-548.

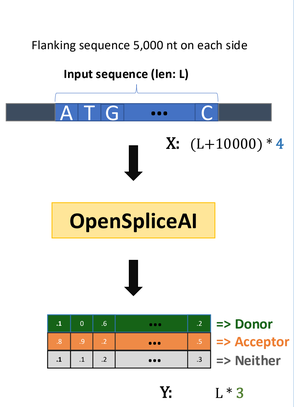
OpenSpliceAI

OpenSpliceAI design & model training

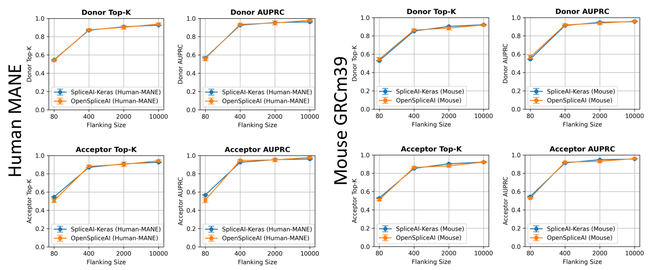
SpliceAI labeling approach





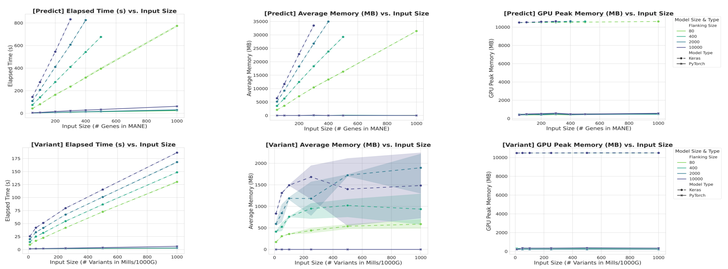


Testing OpenSpliceAI-MANE on Human MANE and Mouse GRCm39 splice sites



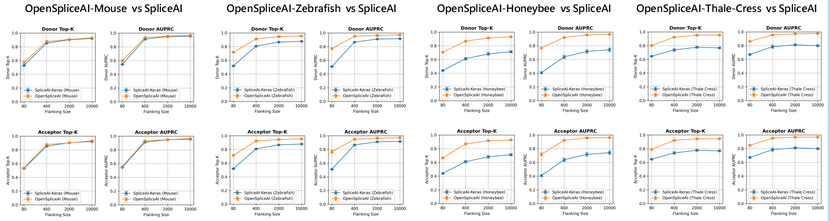
Comparison of OpenSpliceAI-MANE and SpliceAI-Keras on donor and acceptor sites trained with 80nt, 400nt, 2,000nt, and 10,000nt flanking sequences. Metrics include top-1 and AUPRC. Blue curves show SpliceAI-Keras; orange curves show OpenSpliceAI-MANE. Each dot represents the average score with ± one standard deviation.

Elasped time, GPU Memory, and main memory benchmark



OpenSpliceAI

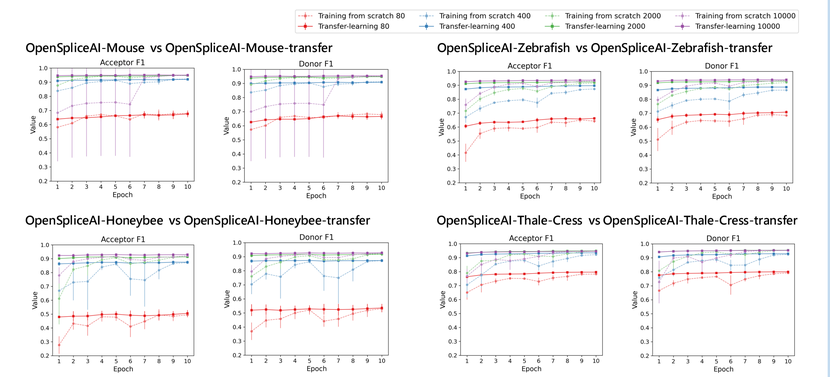
Retraining OpenSpliceAI on Mouse, Zebrafish, Honeybee, and Thale Cress



Performance of OpenSpliceAI trained on species-specific genomes and annotation files (mouse, zebrafish, honeybee, thale cress) vs. SpliceAI-Keras trained on the human genome and the MANE annotation. Orange curves show OpenSpliceAI metrics; blue curves show SpliceAI-Keras. Each subplot covers Top-K accuracy and AUPRC for donor and acceptor sites.

SpliceAI
OpenSpliceAI

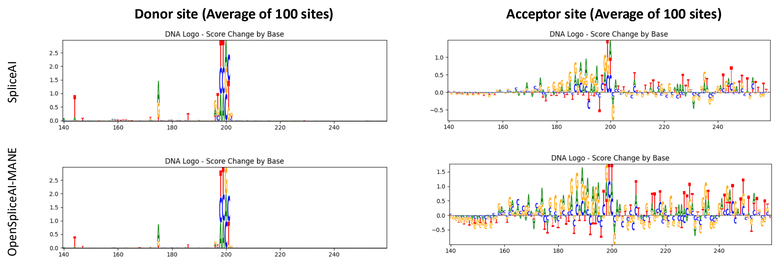
Transfer learning: Transfer OpenSpliceAI-MANE to four species



Transfer learning improves model performance by leveraging knowledge from related domains. We tested if OpenSpliceAI, trained on human splice annotations, could adapt to predict splice sites in other species. Using five pre-trained OpenSpliceAI-MANE models, we fine-tuned species-specific models for mouse, honeybee, zebrafish, and thale cress, then compared them to scratch-trained models. Figures show F1 score for donor and acceptor sites in 80 nt, 400 nt, 2,000 nt, and 10,000 nt models over epochs 1–10 on the test dataset, comparing scratch-trained and transfer-trained variants."

In Silico Mutagenesis (ISM) analysis

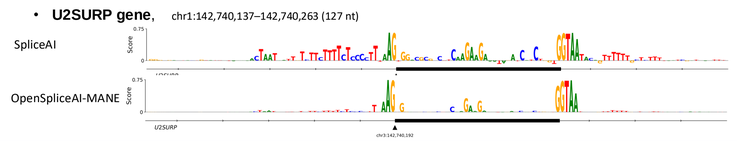
Perturbation-based forward propagation determines the relevance of a feature via mutating each nucleotide to all three possible alternatives. The **"importance score"** of a nucleotide for a splice acceptor is calculated by taking the reference acceptor score, S_{ref}, and recalculating it with the nucleotide replaced by A, C, G, and T to obtain S_A, S_C, S_G, S_T. This procedure is often referred to as in-silico mutagenesis (ISM).



- **DST gene**, exon 2 chr6:56,735,192-56,735,344 (153 nt)

$$S_{ref} = \frac{S_A + S_C + S_G + S_T}{4}$$

- **U2SURP gene**, chr1:142,740,137–142,740,263 (127 nt)



Acknowledgements: supported by NIH under grants R01-HG006677, R35-GM130151, and by NSF under grants DBI-2412449