A Deep Neural Network Approach to Splice Site Prediction

Tilman Hinnerichs

Knowledge Mining Lab – KAUST

September 19, 2019
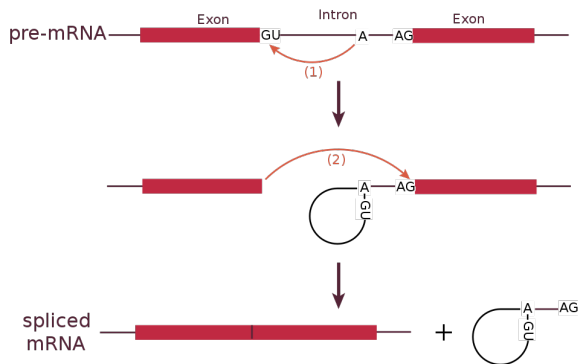
Outline

Problem Description

Figure: RNA splicing reaction

Problem Description

Splice side prediction on Arabidopsis thaliana genome

Problem Description

Splice side prediction on Arabidopsis thaliana genome

▶ Acceptor side:
. . . CGTATCTAGATGAGCA. . .

▶ Donor side:
. . . ATGATTTGTGCAGTCA. . .

Problem Description

Splice side prediction on Arabidopsis thaliana genome

▶ Acceptor side:
  . . . CGTATCT AG ATGAGCA. . .

▶ Donor side:
  . . . ATGATTT GT GCAGTCA. . .

Problem Description

Splice side prediction on Arabidopsis thaliana genome

▶ Acceptor side:
  . . . CGTATCT AG ATG AG CA. . .
▶ Donor side:
  . . . ATGATTT GT GCA GT CA. . .

Dataset description

Example file, e.g., acceptor side

$$
\begin{bmatrix}
CT \ldots \text{300 nt} \ldots AG \ldots \text{300 nt} \ldots GC \\
AG \ldots \text{300 nt} \ldots AG \ldots \text{300 nt} \ldots TT \\
GA \ldots \text{300 nt} \ldots AG \ldots \text{300 nt} \ldots AA \\
\vdots \\
TT \ldots \text{300 nt} \ldots AG \ldots \text{300 nt} \ldots CC
\end{bmatrix}
\begin{array}{c} \text{100,000 records} \end{array}
$$

Classification results: Baseline

▶ Build on label encoded data

▶ All models are validated by 10-fold cross validation

Approach	Samples	Acceptor acc.	Donor acc.
Naive Bayes	40,000	79.32	80.76
Naive Bayes	200,000	79.42	81.12
SVM	4,000	79.21	80.23

Figure: Baseline results

Classification Results: NN – binary classification

▶ Models built on one-hot-encoded data
▶ Dense networks with dropout

Approach	Samples	Depth	Acceptor acc.	Donor acc.
NN	40,000	1	62.24	63.15
NN	40,000	2	92.29	93.50
NN	200,000	2	93.62	93.62
DNN	40,000	4	92.21	93.53
DNN	20,000	6	92.49	93.87
DNN	20,000	7	92.38	93.43
DNN	200,000	7	93.34	93.34
DNN	40,000	8	92.50	93.82
DNN	200,000	8	93.25	93.20
DNN	40,000	15	91.78	92.94

Figure: Binary classification results

Next Steps

▶ Adapt convolution techniques from other papers with same goal

▶ Vary pre and post marker nt sequence length

▶ Use DiProDB[Friedel et al., NAR, 2009] for input

▶ Utilize electron io interaction potential (EIIP) for prediction

Sources

Images:

- https://en.wikipedia.org/