A Deep Neural Network Approach to Splice Site Prediction

Tilman Hinnerichs

Knowledge Mining Lab – KAUST

October 2, 2019

Outline

Problem Description

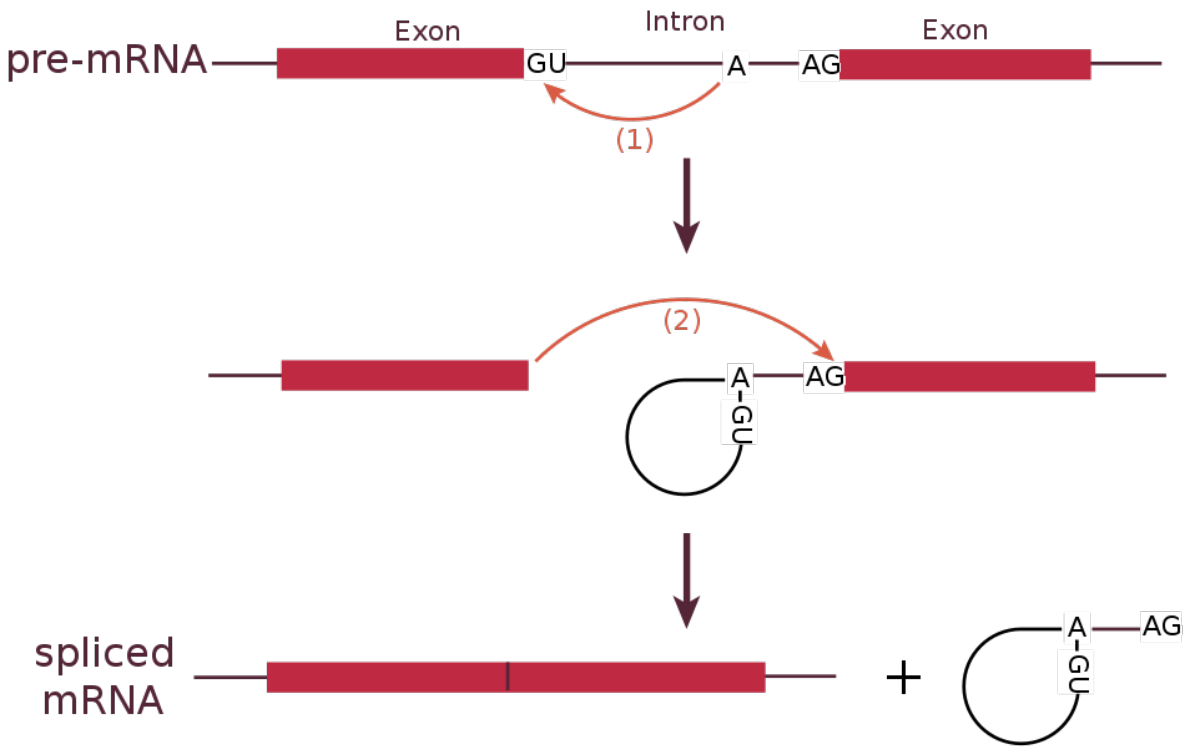

Figure: RNA splicing reaction (en.wikipedia.org)

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

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

Next Steps from last week

- 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

Application of NN to the DiProDB data

- DiProDB is database for the physicochemical properties of dinucleotides (127 entries)
- Applied PCA yielding 15 dimensions

Application of NN to the DiProDB data

- DiProDB is database for the physicochemical properties of dinucleotides (127 entries)
- Applied PCA yielding 15 dimensions

Approach	Samples	Depth	Acceptor			Donor		
			Acc.	Prec	Rec.	Acc.	Prec	Rec
DiProDB	40,000	7	92.38	92.1	93.5	93.6	92.9	93.6
DiProDB	40,000	9	92.28	90.4	94.2	93.4	92.0	93.9

DiProDB BINARY CLASSIFICATION APPROACH
Data shape: (40000, 601, 15)

Layer (type)	Output Shape	Param #
input_10 (InputLayer)	(None, 601, 15)	0
flatten_10 (Flatten)	(None, 9015)	0
dense_46 (Dense)	(None, 150)	1352400
dropout_37 (Dropout)	(None, 150)	0
dense_47 (Dense)	(None, 80)	12080
dropout_38 (Dropout)	(None, 80)	0
dense_48 (Dense)	(None, 30)	2430
dropout_39 (Dropout)	(None, 30)	0
dense_49 (Dense)	(None, 10)	310
dropout_40 (Dropout)	(None, 10)	0
dense_50 (Dense)	(None, 1)	11

Application of CNN to the DiProDB data

- Convolutional NN to adapt ideas from other papers

Application of CNN to the DiProDB data

- Convolutional NN to adapt ideas from other papers

Approach	Layers		Acceptor			Donor		
	Conv.	Others	Acc.	Prec	Rec.	Acc.	Prec	Rec
CNN DPDB	4	5	94.4	95.4	94.6	94.9	94.4	94.7
CNN DPDB	4	7	93.5	93.3	94.5	94.0	94.0	93.3
CNN DPDB	6	5	94.0	93.9	94.9	94.2	95.4	91.6
CNN DPDB	6	5	94.4	97.0	93.8	95.2	96.5	93.7
SpliceRover	4	2	96.1	93.9	97.4	95.4	95.6	96.7

SpliceRover[Zuallaert et al., 2018]

DiProDB: BINARY CLASSIFICATION APPROACH
Data shape: (40000, 601, 15, 1)
Epochs: 5, Batch size: 500
Model: "model_10"

Layer (type)	Output Shape	Param #
input_10 (InputLayer)	(None, 601, 15, 1)	0
conv2d_28 (Conv2D)	(None, 599, 1, 32)	1472
max_pooling2d_28 (MaxPooling	(None, 299, 1, 32)	0
conv2d_29 (Conv2D)	(None, 297, 1, 64)	6208
max_pooling2d_29 (MaxPooling	(None, 148, 1, 64)	0
conv2d_30 (Conv2D)	(None, 146, 1, 128)	24704
max_pooling2d_30 (MaxPooling	(None, 48, 1, 128)	0
flatten_10 (Flatten)	(None, 6144)	0
dense_28 (Dense)	(None, 64)	393280
dropout_19 (Dropout)	(None, 64)	0
dense_29 (Dense)	(None, 64)	4160
dropout_20 (Dropout)	(None, 64)	0
dense_30 (Dense)	(None, 1)	65

Next steps

- Follow improvements of SpliceRover
- repDNA (Liu et al., 2015) for different embeddings of the sequences
- Train CNN on all embeddings
- Freeze convolutional layers and funnel them into DNN

Citations

- Liu B, Liu F, Fang L, Wang X, Chou K-C.repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics 2015;31(8):1307-1309.
- Jasper Zuallaert, Fréderic Godin, Mijung Kim, Arne Soete, Yvan Saeys, Wesley De Neve, SpliceRover: interpretable convolutional neural networks for improved splice site prediction, Bioinformatics, Volume 34, Issue 24, 15 December 2018, Pages 4180–4188, https://doi.org/10.1093/bioinformatics/bty497