Splice site prediction with a Hidden Markov Model

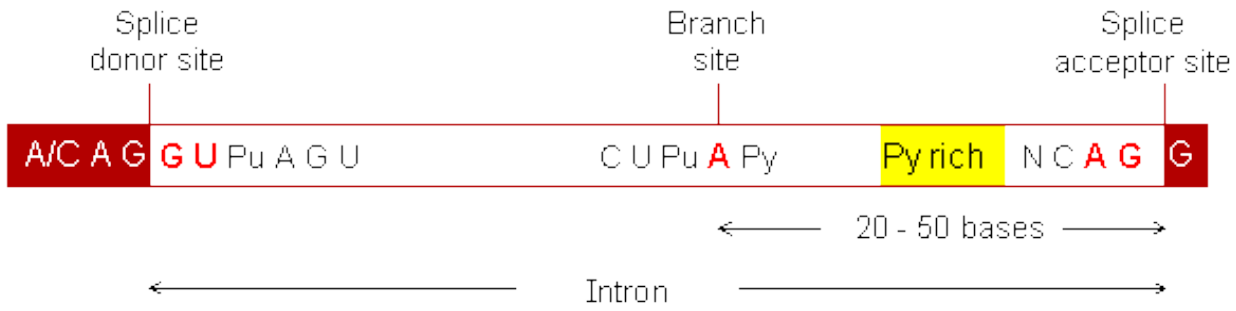
Xinru Qiu

3-19-2019

Background:

In a split gene, before messenger RNA is translated into protein, it undergoes splicing to remove some parts of the RNA. The removed intervals are called introns, and the remaining intervals, which code for a protein in a protein-coding gene are called exons.
The donor site is the region that separates the end of an exon from the beginning of an intron. The acceptor site is the region that separates the end of an intron from the beginning of an exon.

The most conservative part of the donor site is the first two nucleotides of introns (usually GU). The most conservative part of the acceptor site is the last two nucleotides of introns (usually AG). We will refer to the first two nucleotides in intron as Donor1 and Donor2 and the last two nucleotides in introns Acceptor1 and Acceptor2. For example, for canonical donor and acceptor sites GU and AG, Donor1 = G, Donor2 = U, Acceptor1 = A, and Acceptor2 = G.



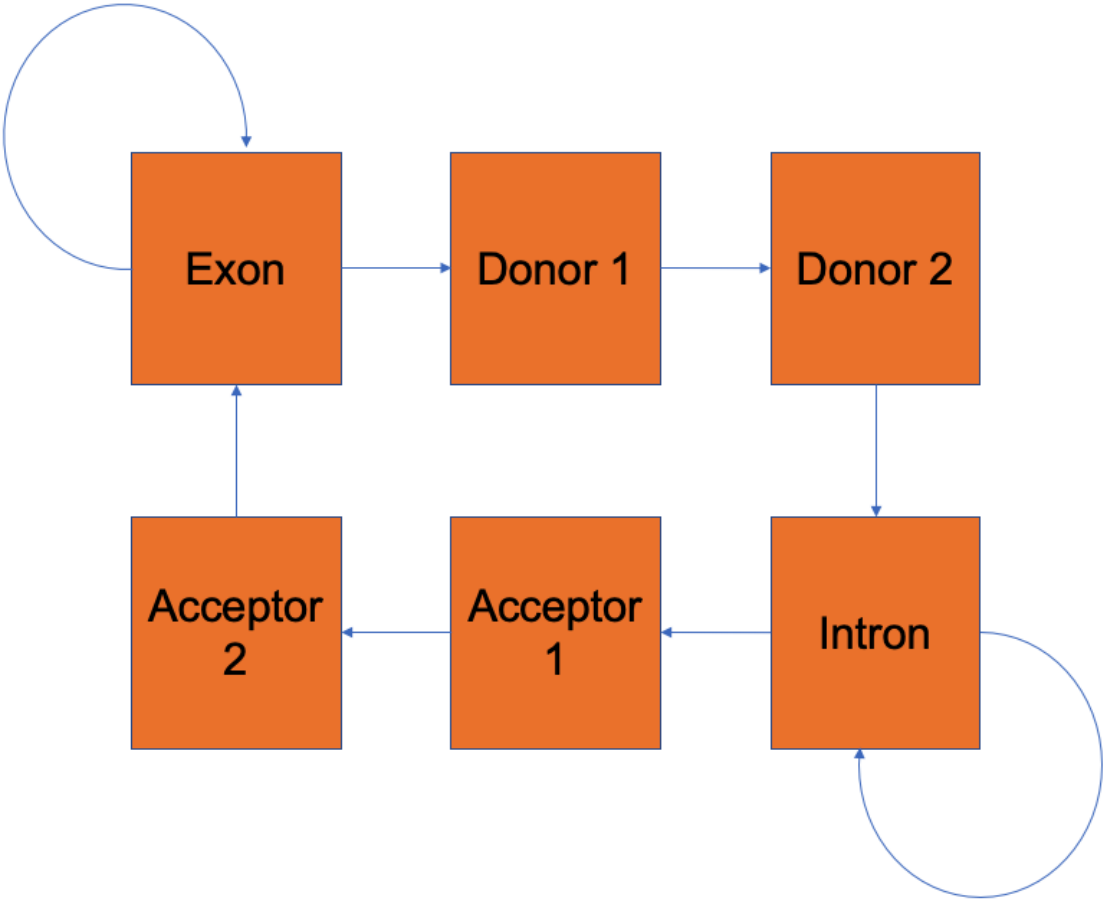
Problem description:

Our goal is to determine whether the genomic region contains exons, and if so, to find them based on finding the donor and acceptor sites.

Description of HMM topology 1:

- Using the entire exon and intron region per gene as training set.

1. Σ: {A, C, G, T}
2. States: Exon, Donor1, Donor2, Intron, Acceptor 1, Acceptor 2



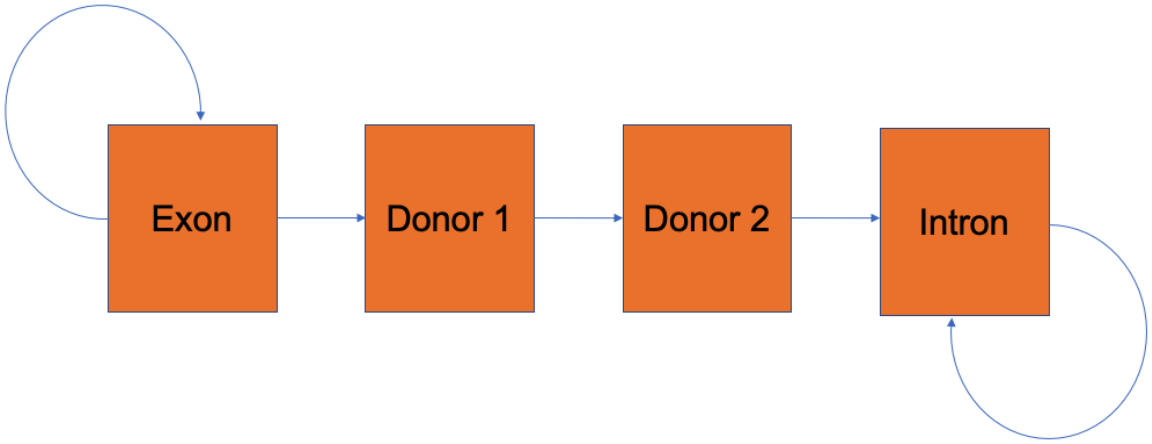
Problem of HMM topology 1:

- The length of Introns are averagely four to five times longer than Exons
- Also, the length of Introns and Exons >> number of Donors and Acceptors.
- The transition state from Intron to Intron is over 99%, so when predicting splicing sites, once get into the Intron state, it is hard to get out.

Description of HMM topology 2:

- Create Donor and Acceptor models.

Donor model



Base	1	2	3	4	5	6	7	8	9
State	Exon	Exon	Exon	Donor1	Donor2	Intron	Intron	Intron	Intron

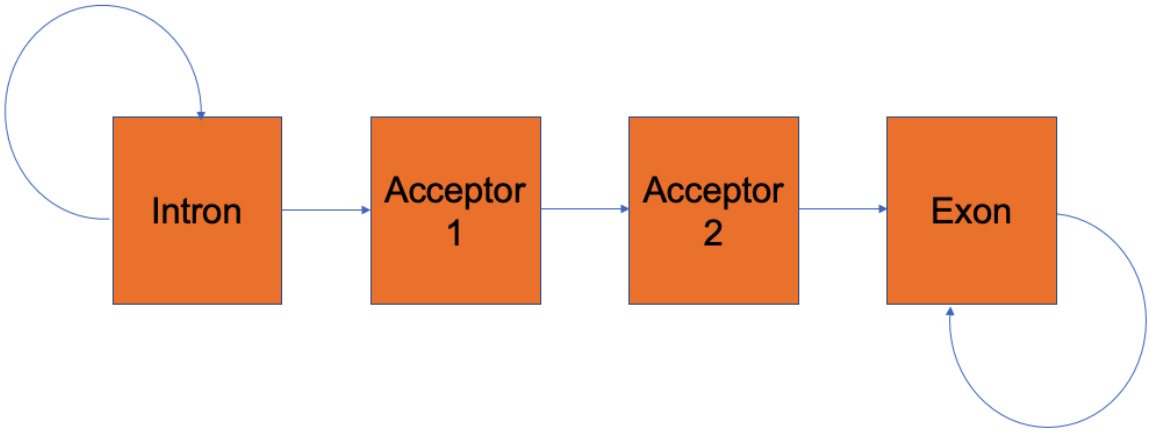
1. Σ: {A, C, G, T}
2. States: Exon, Donor1, Donor2, Intron
3. Transition matrix

	Transition To: **Exon**	Transition To: **Donor1**	Transition To: **Donor2**	Transition To: **Intron**
Exon	0.67	0.33	0.00	0.00
Donor1	0.00	0.00	1.00	0.00
Donor2	0.00	0.00	0.00	1.00
Intron	0.00	0.00	0.00	1.00

4. Emission matrix

	Emit: **A**	Emit: **C**	Emit: **G**	Emit: **T**
Exon	0.34	0.16	0.38	0.11
Donor1	0.00	0.00	1.00	0.00
Donor2	0.00	0.00	0.00	1.00
Intron	0.42	0.05	0.47	0.05

Acceptor model



Base	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
State	I	I	I	I	I	I	I	I	I	I	I	I	I	A1	A2	E	E	E

1. Σ: {A, C, G, T}
2. States: Intron, Acceptor 1, Acceptor 2, Exon
3. Transition matrix

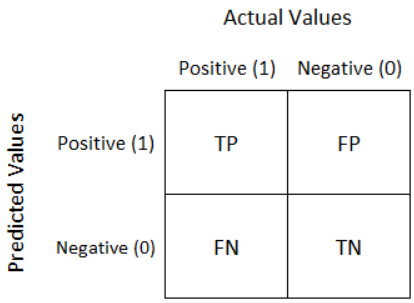
	Transition To: **Intron**	Transition To: **Acceptor1**	Transition To: **Acceptor2**	Transition To: **Exon**
Intron	0.92	0.08	0.00	0.00
Acceptor1	0.00	0.00	1.00	0.00
Acceptor2	0.00	0.00	0.00	1.00
Exon	0.00	0.00	0.00	1.00

4. Emission matrix

	Emit: **A**	Emit: **C**	Emit: **G**	Emit: **T**
Intron	0.09	0.39	0.10	0.42
Acceptor1	1.00	0.00	0.00	0.00
Acceptor2	0.00	0.00	1.00	0.00
Exon	0.24	0.19	0.36	0.21

The measurement of the success of the HMM

Using confusion matrix:



Measurement

- Sensitivity = TP/(TP+FN)
- Specificity = TN/(TN+FP)
- Precision = TP/(TP+FP)
- Accuracy = (TP+TN)/(TP+TN+FP+FN)

Donor Model

	Positive	Negative
TRUE	310	241035
FALSE	11822	1

- Sensitivity = TP/(TP+FN) = 100%
- Specificity = TN/(TN+FP) = 95%
- Precision = TP/(TP+FP) = 3%
- Accuracy = (TP+TN)/(TP+TN+FP+FN) = 95%

Acceptor Model

	Positive	Negative
TRUE	285	250756
FALSE	2101	26

- Sensitivity = TP/(TP+FN) = 92%
- Specificity = TN/(TN+FP) = 99%
- Precision = TP/(TP+FP) =12%
- Accuracy = (TP+TN)/(TP+TN+FP+FN) = 99%

Problem of HMM topology 2:

- Introducing too many false positives by using Donor and Acceptor neighbor bases as training set. If apply this approach to wet lab experiments would cause extremely high laboratory work and cost.

Solution:

1. Decoy Donor & Acceptor model

- Create a Donor and Acceptor model that mainly recognize non-Donor and non-Acceptor sites. Then use the findings exclude high false positives.
2. Increase the Donor and Acceptor model training set sequence length include in more bases from Exon and Intron, this way can lower the false positives. However, this approach may result in introducing more false negatives.
3. Try duration HMM, consider the frequency of intron to intron. For example, although Intron to Intron transition probability is over 99%, the probability of a hundred of them show together is only 37%.