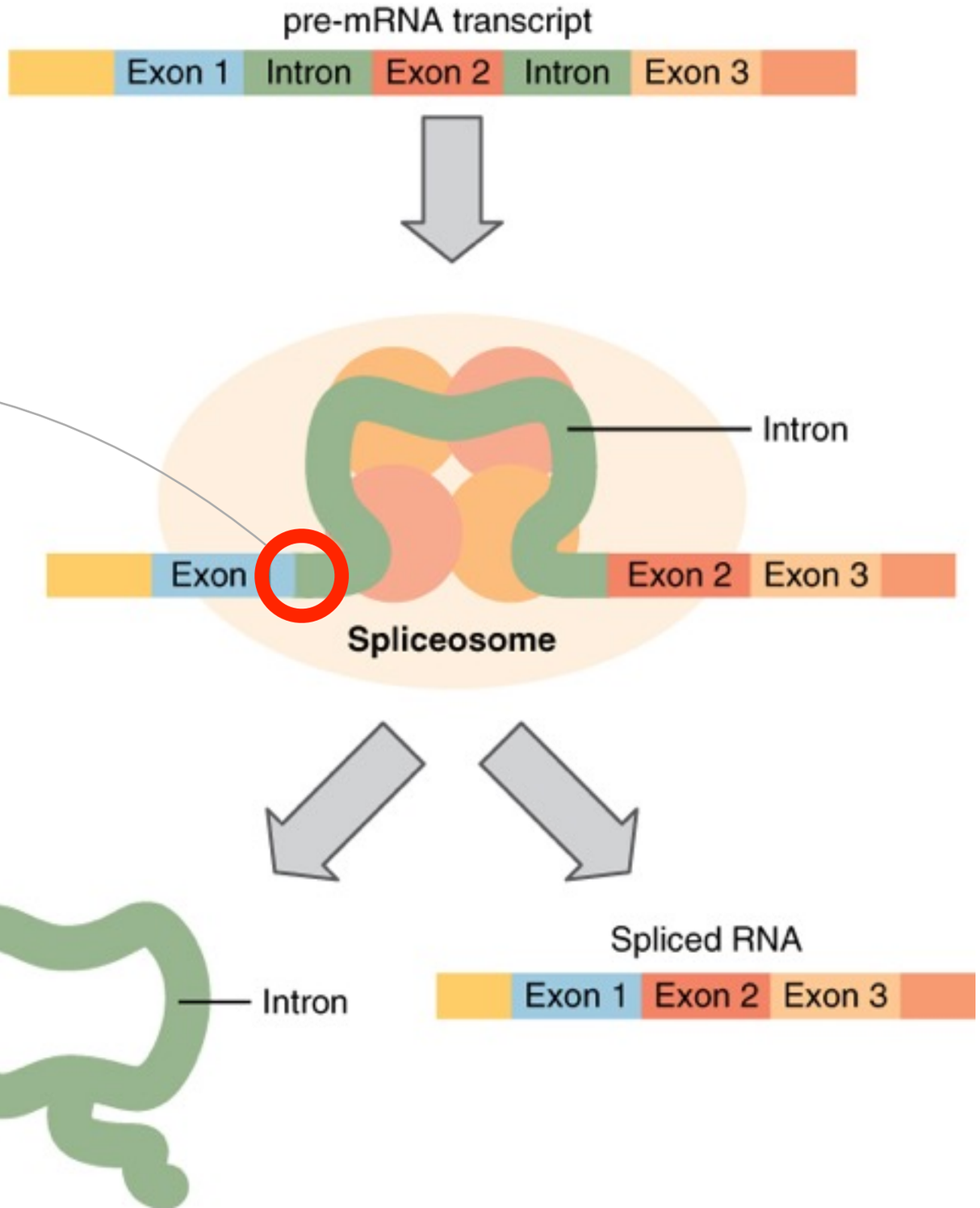
HMM toy model for splicing

Álvaro Abella
Clàudia Fontserè

Introduction

- What is splicing?

we work here



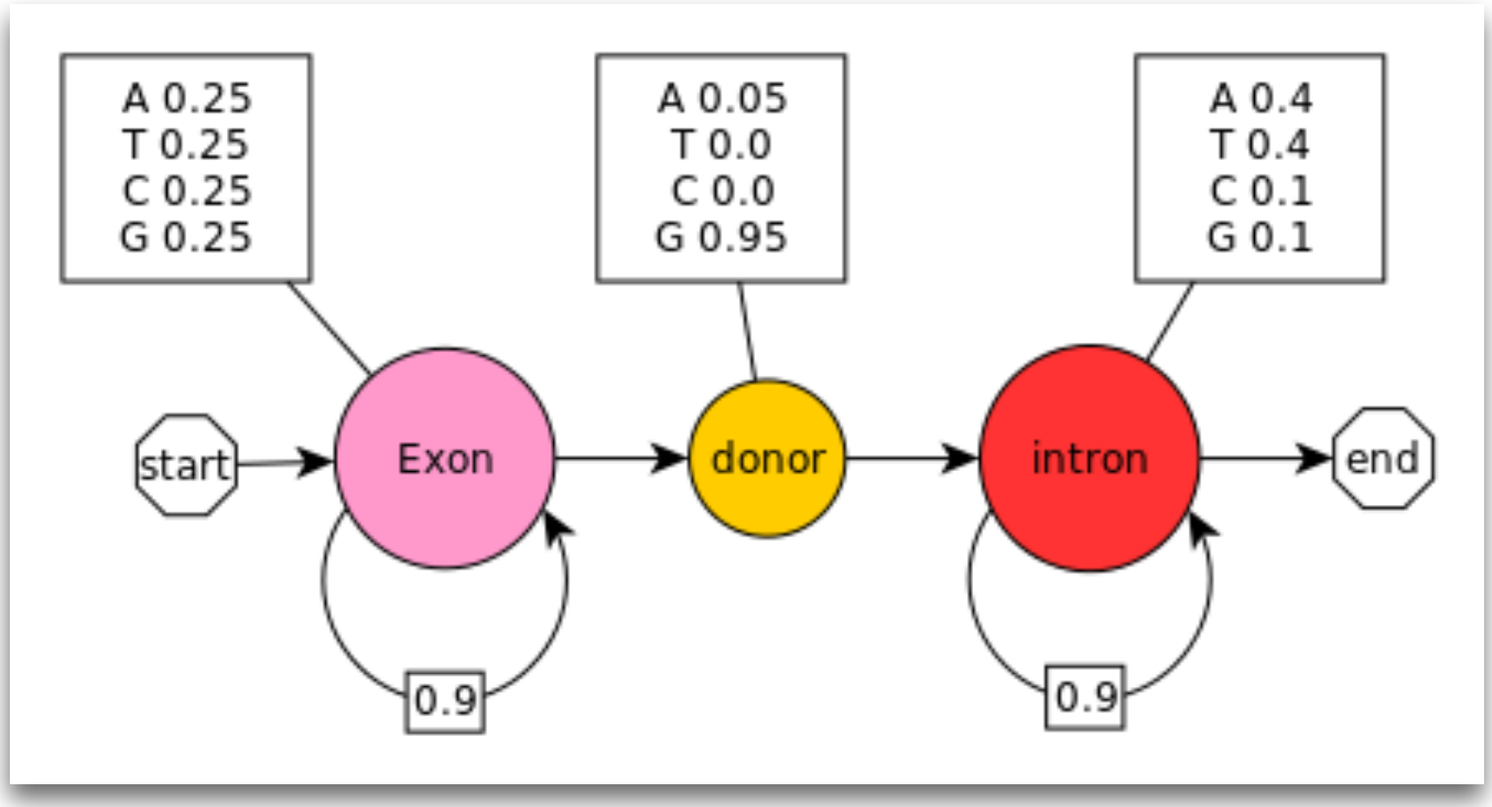
Objectives

- Implement a program to sample sequences from an **HMM**;
- Implement the **Viterbi algorithm** and test it with the toy donor splice site model;
- **Change the model** into a donor site (5' splice site) model that considers the binding of the **U1 snRNP,** by extending the number of states that describe the exon-intron boundary. How many positions should you use for your model? Provide an argument for your answer.
- Incorporate into the previous model a **state describing the presence of a TIA-1** binding site (a Uridine-rich sequence) immediately downstream of the donor site.
- Make an assessment of the performance of the model using **accuracy** measures. Do you find any improvement between models?

Sample sequences from an HMM

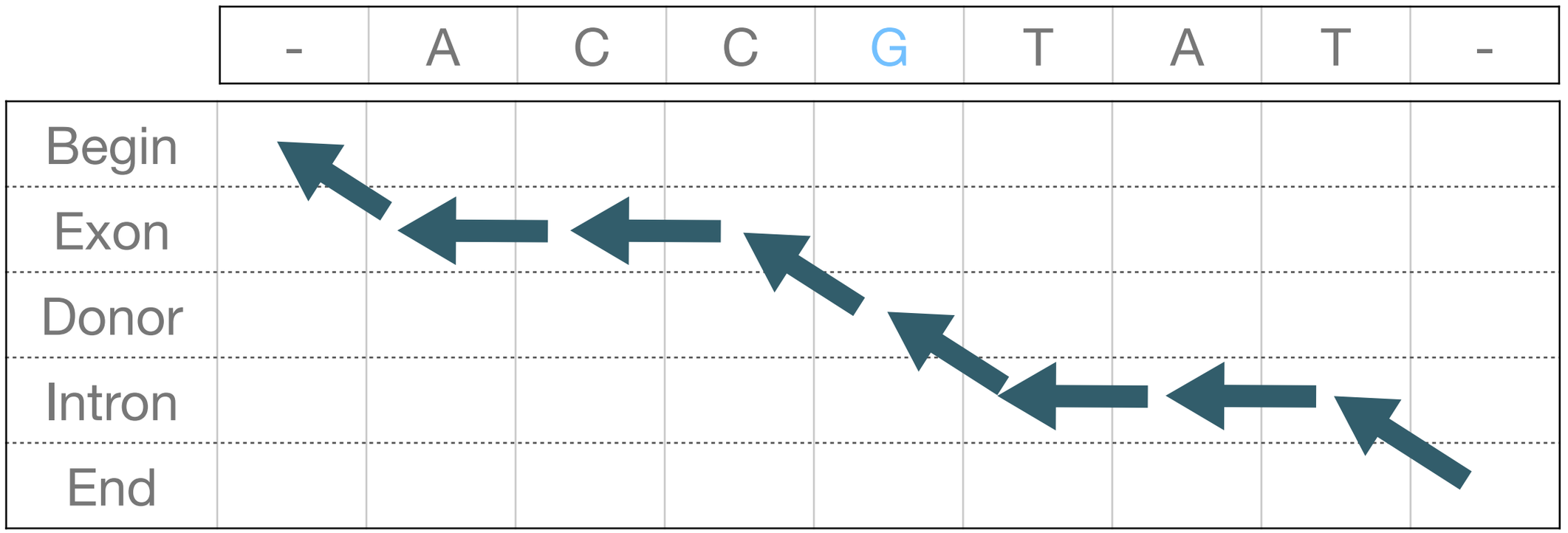
- Script: given an HMM and a required number *n* of observations, it samples a sequence of *n* observations from the given HMM.

- The toy model:



Viterbi algorithm

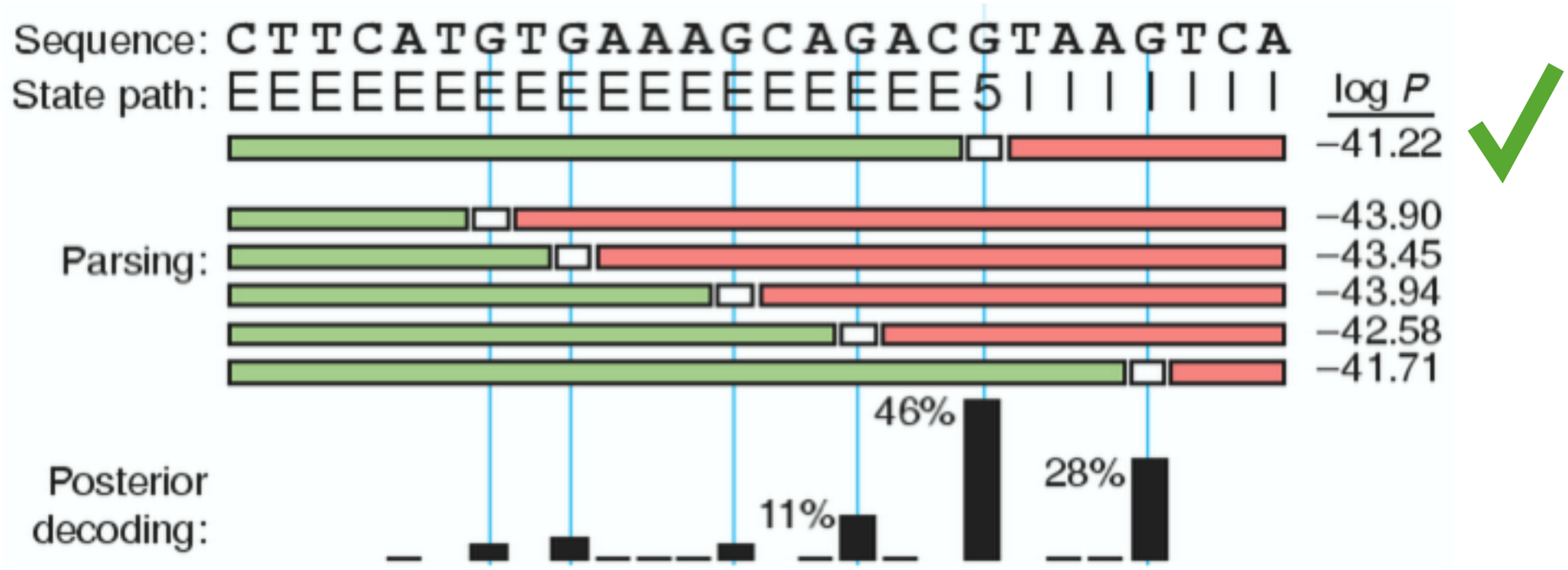
- Script: given an input DNA sequence, it calculates the **state path** (the sequence of hidden states) that maximizes the joint probability of the model and the DNA sequence.



Viterbi algorithm

- DNA sequence of Eddy's article to test whether it works:

-	C	T	T	C	A	T	G	T	G	A	A	A	G	C	A	G	A	C	G	T	A	A	G	T	C	A
B	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	D	I	I	I	I	I	I	I



Testing of accuracy of TOY MODEL

- We have a set of exon-intron sequences:

- To test the accuracy we will compute:
  - **TP** - true positives: number of real donor sites correctly predicted as such
  - **FP** - false positives: number of predicted donor sites that do not correspond to the real ones.
  - **FN** - false negatives: number of real donor sites incorrectly predicted (i.e., missed). Note that in this case, FP = FN.

Testing of accuracy of TOY MODEL

- **SN - sensitivity**. Proportion of real donor sites predicted as such:

$$SN = \frac{TP}{TP + FN}$$

- **SP - specificity**. Proportion of real donor sites among all our predictions:

$$SP = \frac{TP}{TP + FP}$$

Testing of accuracy of TOY MODEL

- Results for a file with real donor sites:

```
Sensitivity: 0.0709844559585492
Specificity: 0.0709844559585492
True positive: 137
False Positive/False negative: 1793
```

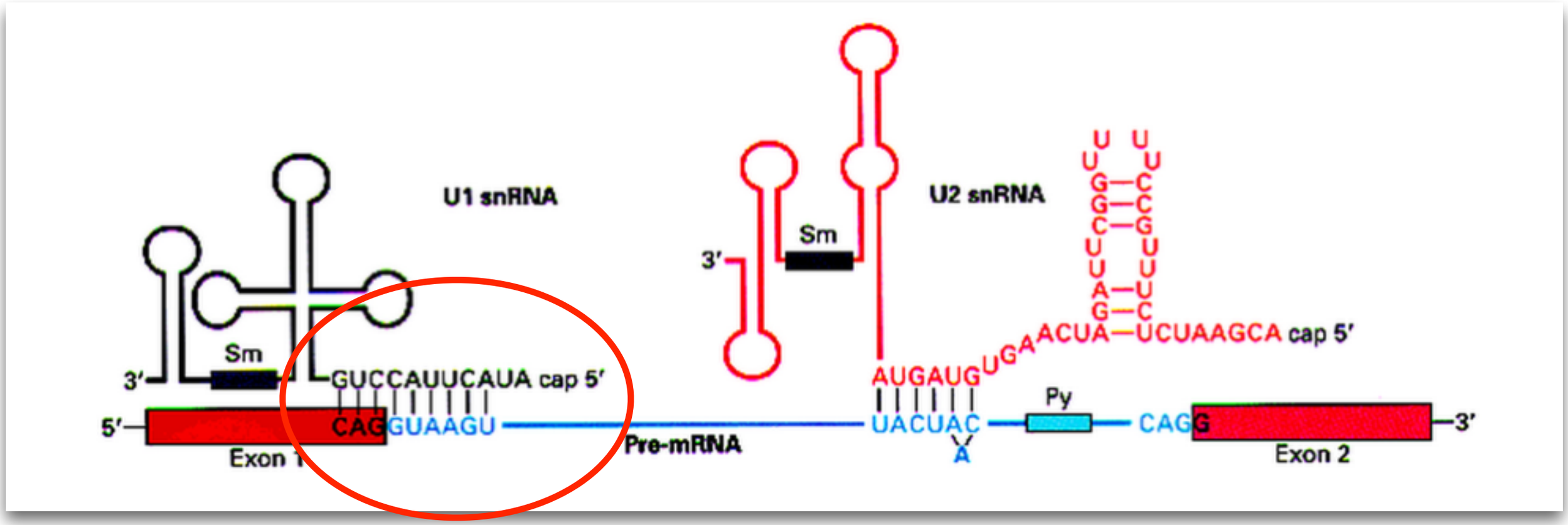
SN = 7%

SP = 7%

This model is not
good enough

We must improve it!

Improving the toy model:
binding of the U1 snRNP



U1snRNP initiates spliceosome assembly by binding to the 5' splice site through base pairing between the single stranded terminal sequence of the U1 RNA molecule and the loosely conserved stretch of nucleotides at the 5' splice site (**CAG/GURAGU**) marking the exon-intron boundary.

Improving the toy model: binding of the U1 snRNP

- We have two datasets:

Real donor sites	False donor sites
CATGTAAGT	CCTGTTTGT
CAGGTAAGC	GCTGTTCAT
CAGGTAGGG	CGGGTCGGC
GTGGTAAGG	TCGGTGAAG
GAGGTGAGT	TCTGTATTG
CAAGTAAGT	GCAGTGATC
AATGTAAGA	TCTGTATTG
AGAGTAAGG	GCAGTGATC

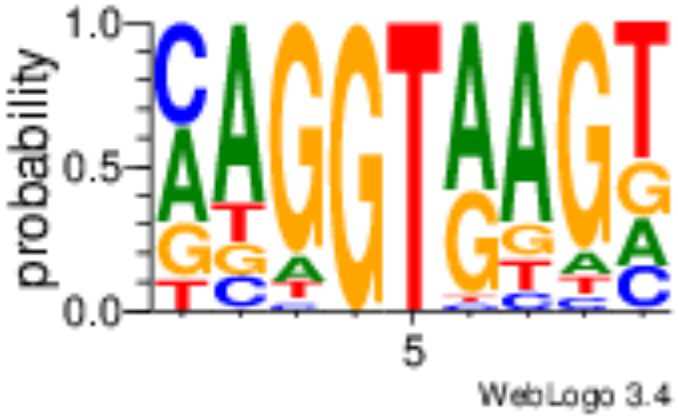
CAG/GURAGU

We want to find which positions are relevant and extract the emission probabilities

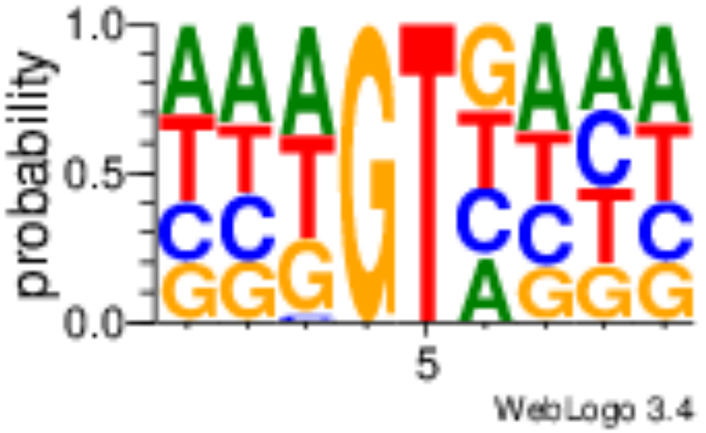
Improving the toy model: binding of the U1 snRNP

- Counting frequencies

real donor sites



false donor sites



- Which are the most relevant positions?

Kullback-Leibler divergence

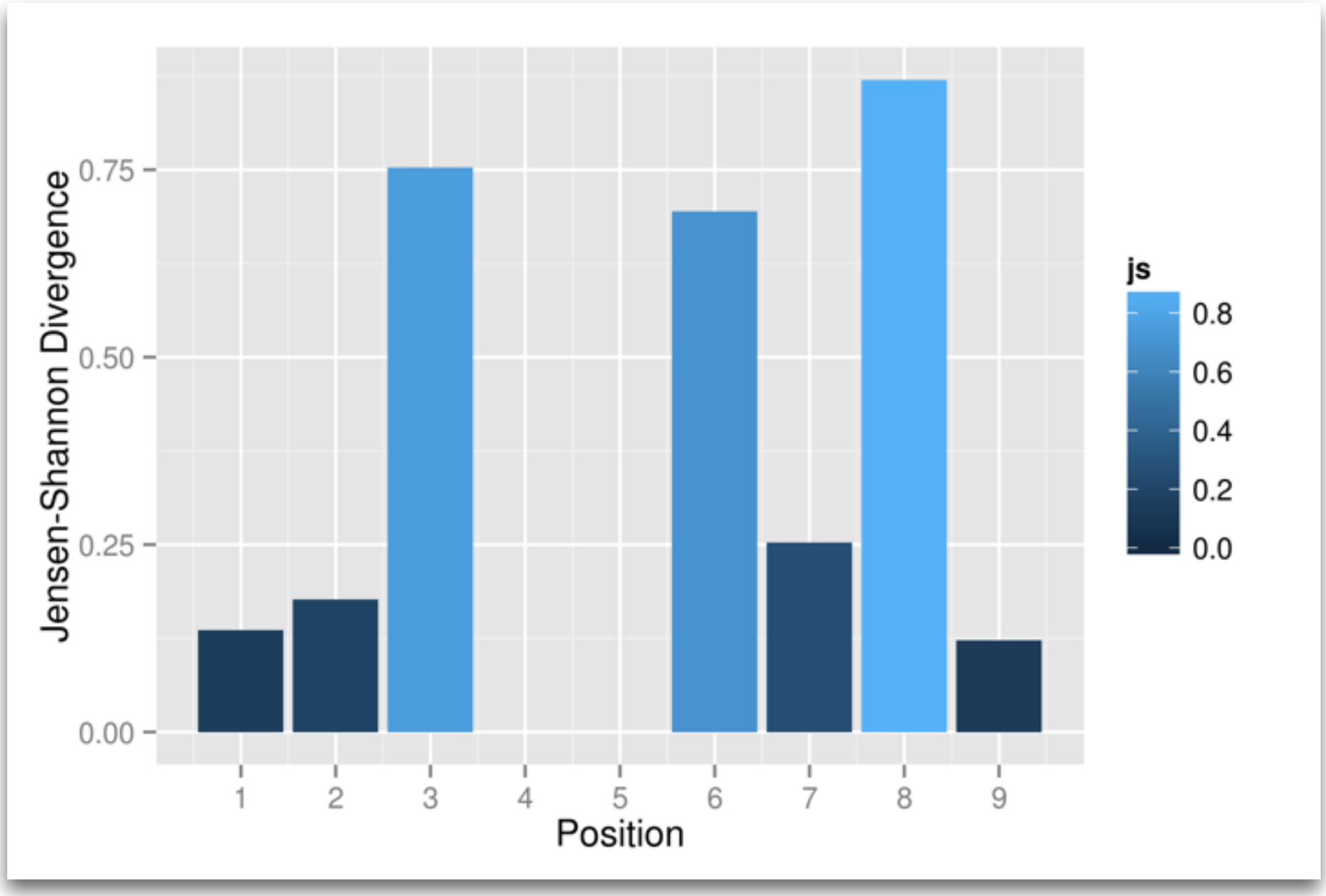
$$D_{KL}(P \parallel Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

Jensen-Shannon divergence

$$JS(P,Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M)$$

Improving the toy model: binding of the U1 snRNP

- Results of JS divergence



Only most informative: 3-8

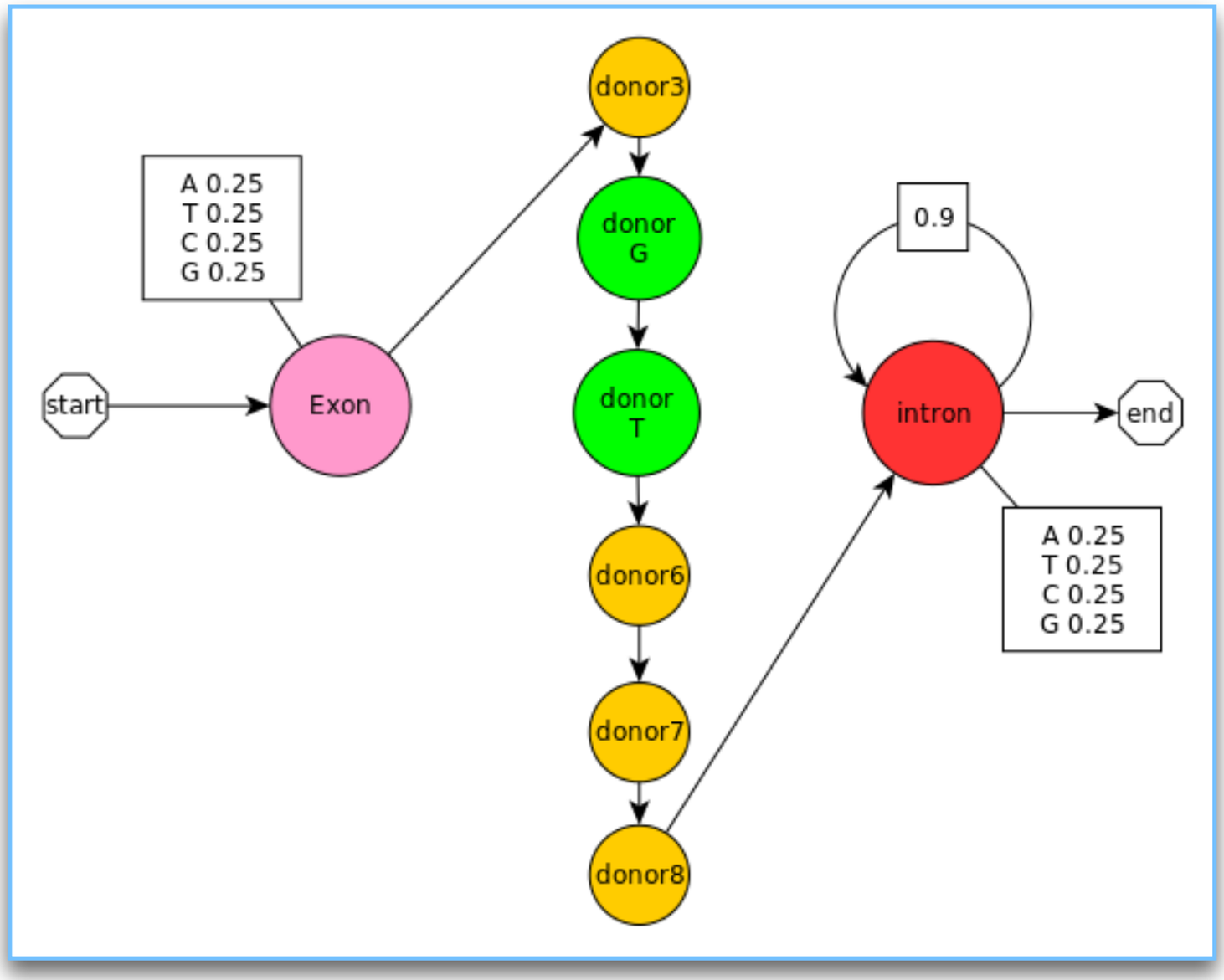
All positions: 1-9

Improving the toy model: binding of the U1 snRNP

- Modify the toy model adding more states!

Only most informative: 3-8

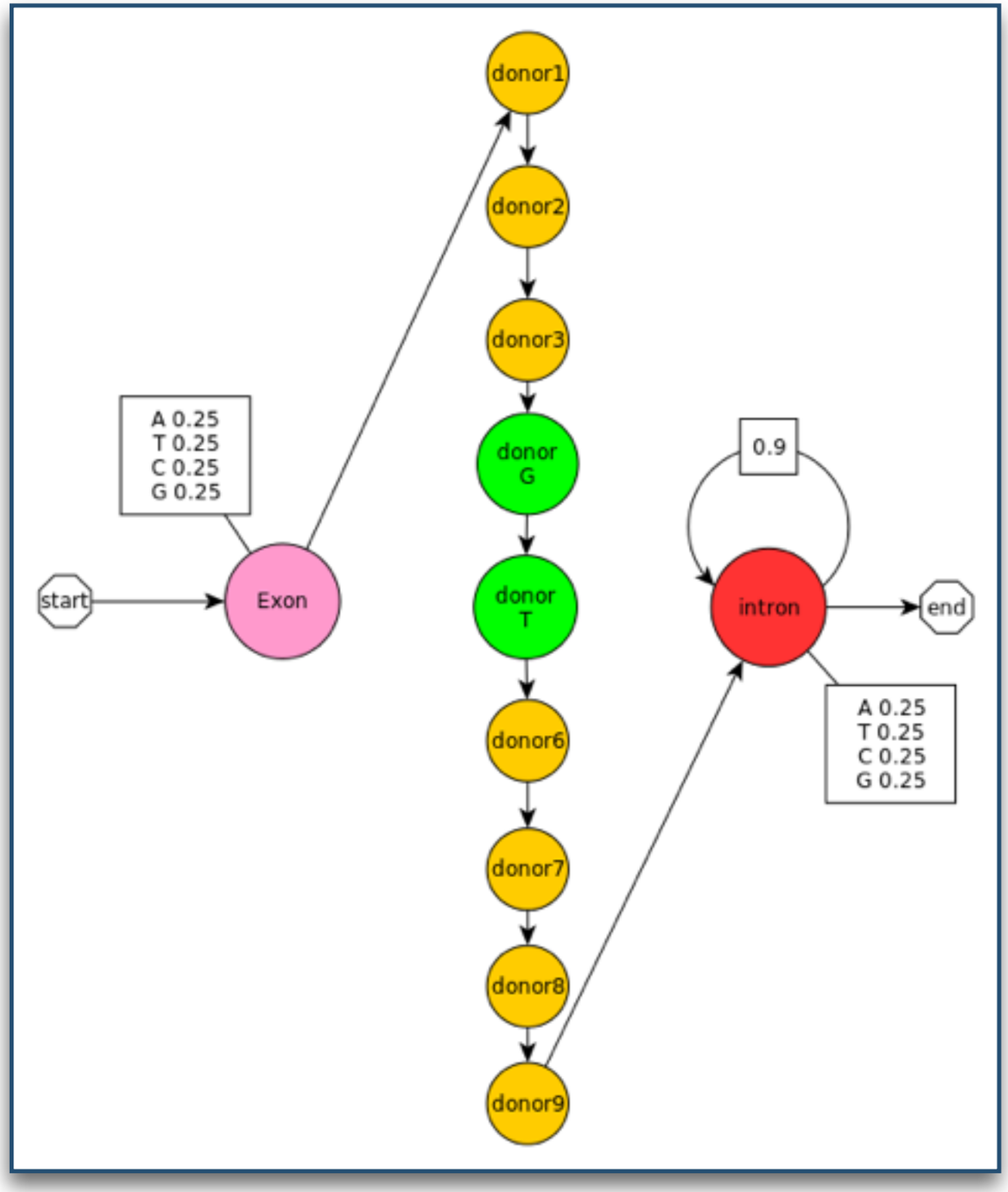
All positions: 1-9



Improving the toy model: binding of the U1 snRNP

Only most informative: 3-8

All positions: 1-9



Improving the toy model: binding of the U1 snRNP

- Testing accuracy of both new models:

Only most informative: 3-8

Big test set

```
Sensitivity: 0.86580310880829
Specificity: 0.86580310880829
True positive: 1671
False Positive/False negative: 259
```

86%

All positions: 1-9

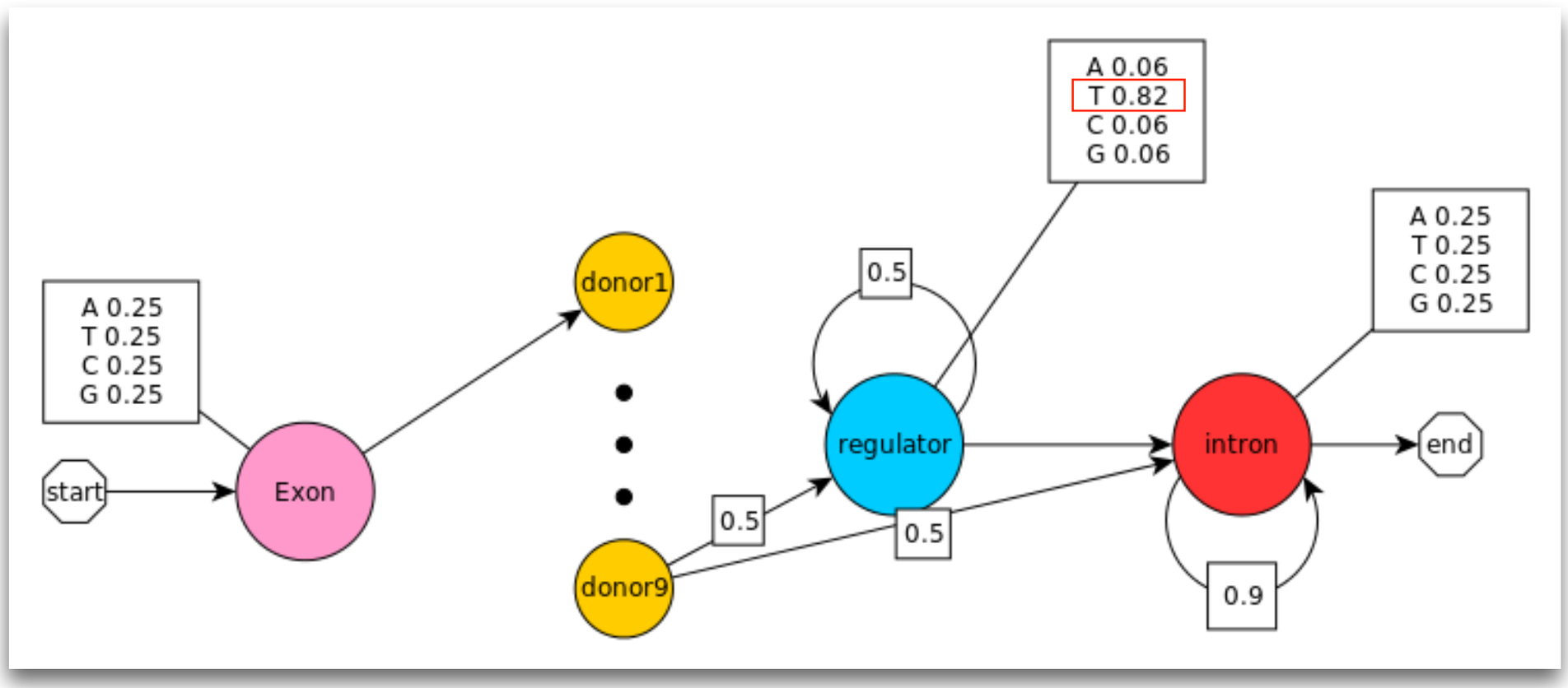
Big test set

```
Sensitivity: 0.940932642487047
Specificity: 0.940932642487047
True positive: 1816
False Positive/False negative: 114
```

94%

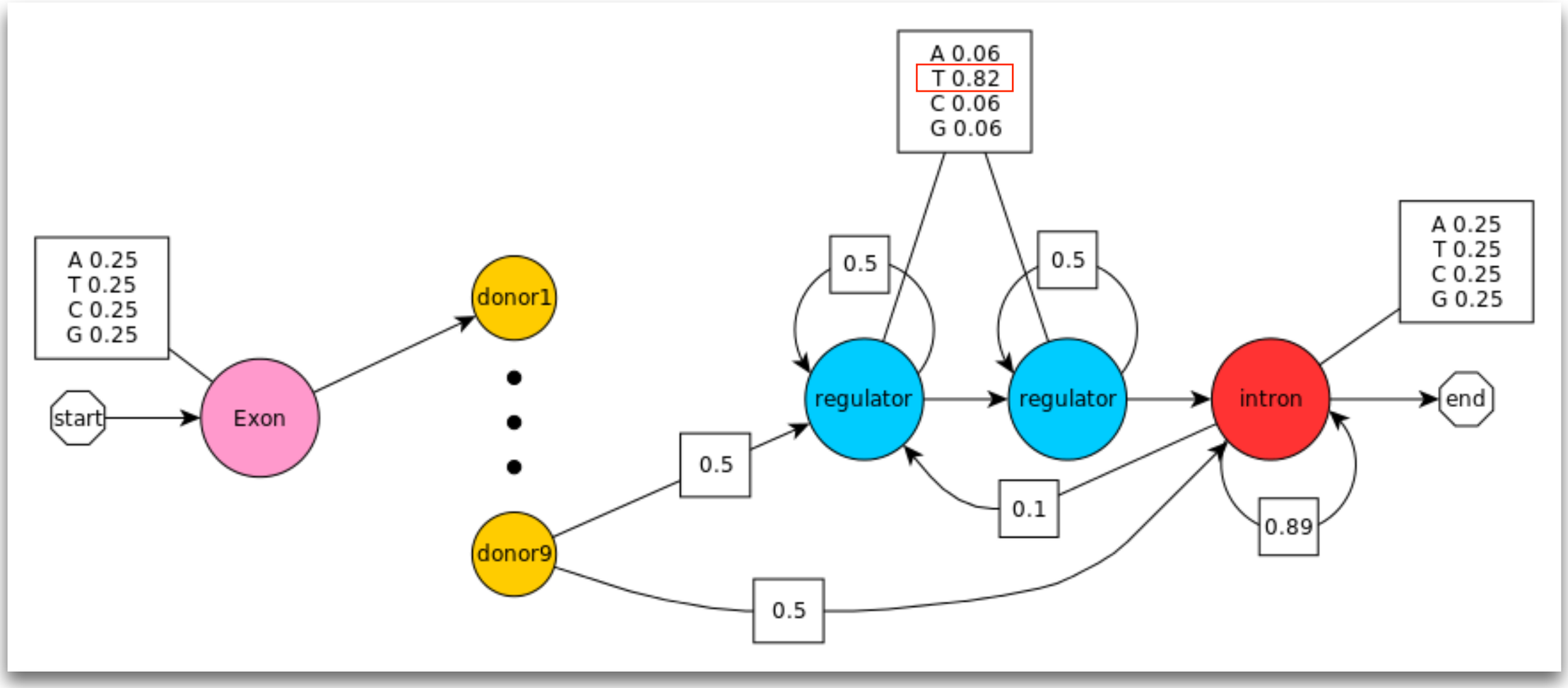
Improving the toy model: binding of the TIA

- TIA: regulation of pre-mRNA splicing and mRNA translation
    - we add a new state —> Splicing regulator (S)



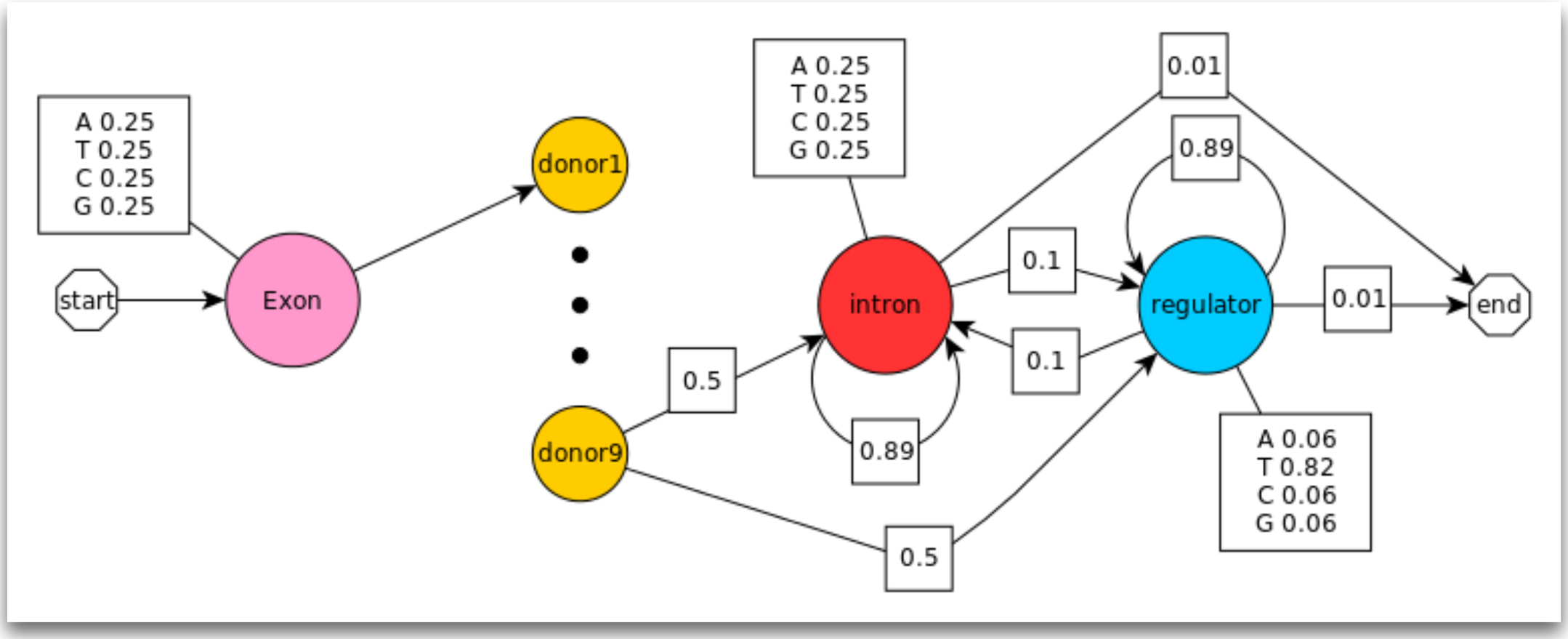
Improving the toy model: binding of the TIA

- Improve TIA model adding another S state and allowing to go from intron to regulator.



Improving the toy model: binding of the TIA

- TIA 3. A little bit different.



Improving the toy model: binding of the TIA

- Results:
  - TIA 1 :

```
Sensitivity: 0.926424870466321
Specificity: 0.926424870466321
True positive: 1788
False Positive/False negative: 142
```

SN: 92.64%

  - TIA2

```
Sensitivity: 0.938341968911917
Specificity: 0.938341968911917
True positive: 1811
False Positive/False negative: 119
```

SN: 93.83%

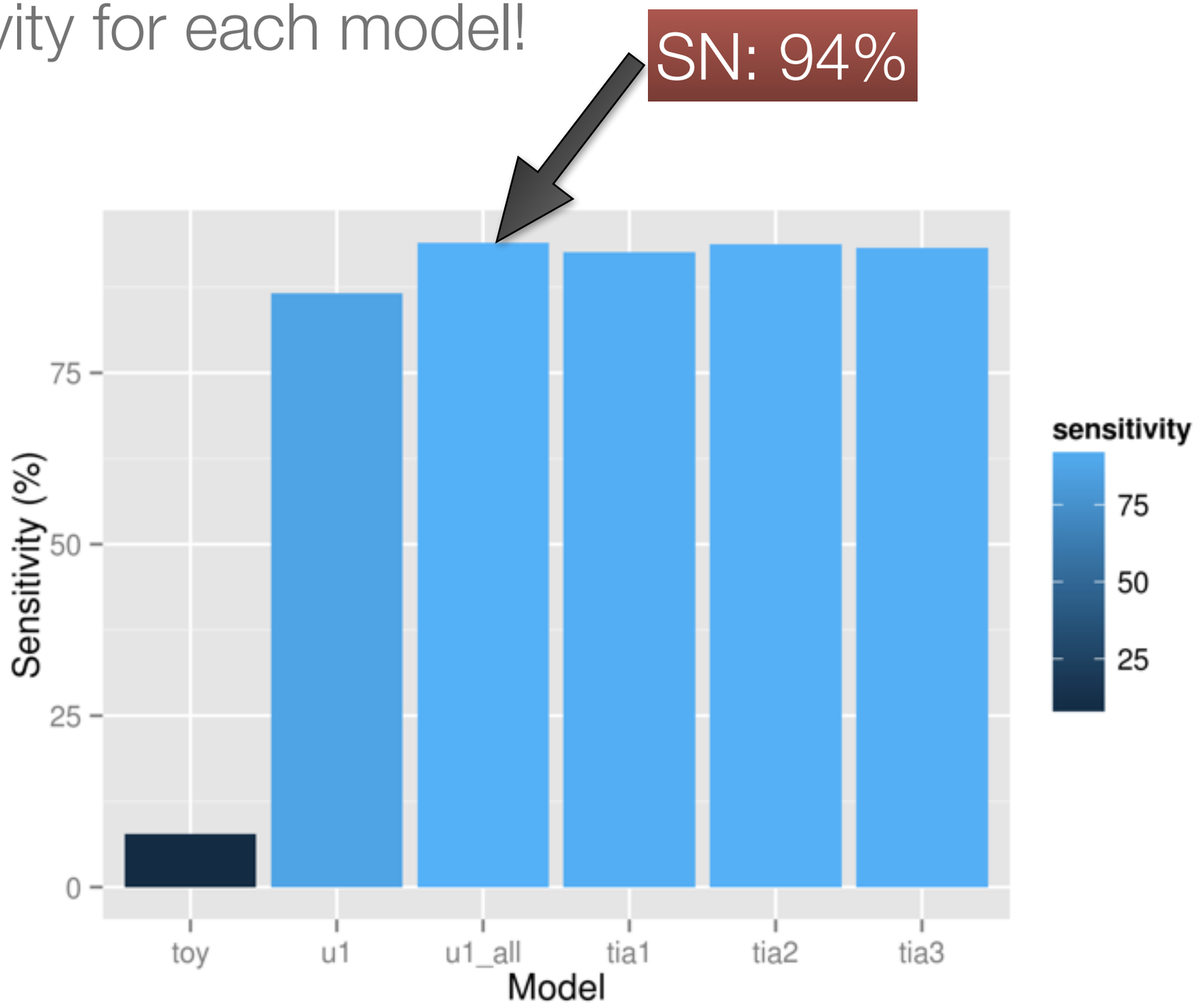
  - TIA3

```
Sensitivity: 0.931606217616580
Specificity: 0.931606217616580
True positive: 1798
False Positive/False negative: 132
```

SN: 93.16%

Discussion

- Sensitivity for each model!



Discussion

- The best model we obtained is U1_all with sensitivity/ specificity of 94%!

- Why?

  “The contribution of the TIA-1 binding becomes negligible upon improving base-pairing complementarity between U1 snRNA and 5’ splice site.” (Izquierdo et al. 2005)

Thank you very much for your attention!