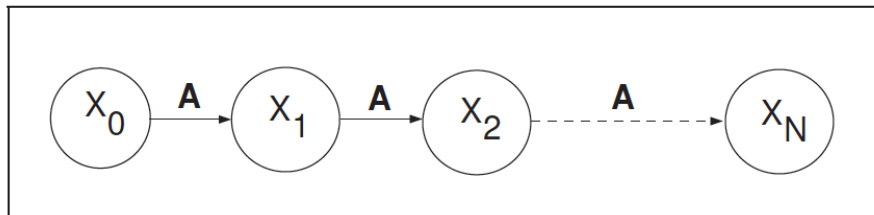


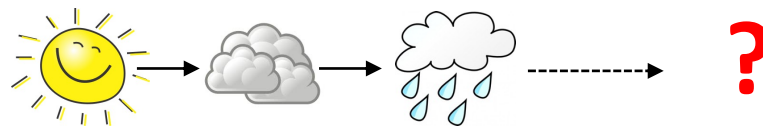
# Introducing Hidden Markov Models

## First – a Markov Model

**A Markov Model** is a chain-structured process where future states depend only on the present state, not on the sequence of events that preceded it.



The  $X$  at a given time is called the **state**.  
The value of  $X_n$  depends only on  $X_{n-1}$ .



**State** : sunny   cloudy   rainy   sunny ?

# The Markov Model



90 % sunny  
10% rainy

**State :** sunny sunny rainy sunny

A bracket is drawn under the first four states (sunny, sunny, rainy, sunny).

## State transition probability (table/graph)

(The probability of tomorrow's weather given today's weather)

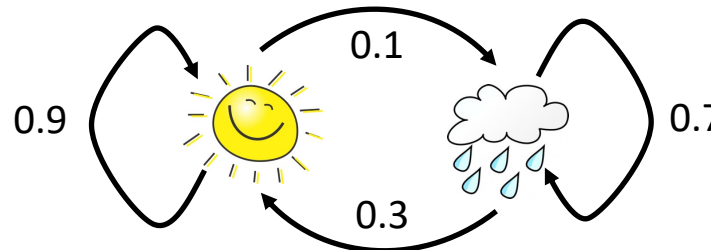
### Output format 1:

Today	Tomorrow	Probability
sunny	sunny	0.9
sunny	rainy	0.1
rainy	sunny	0.3
rainy	rainy	0.7

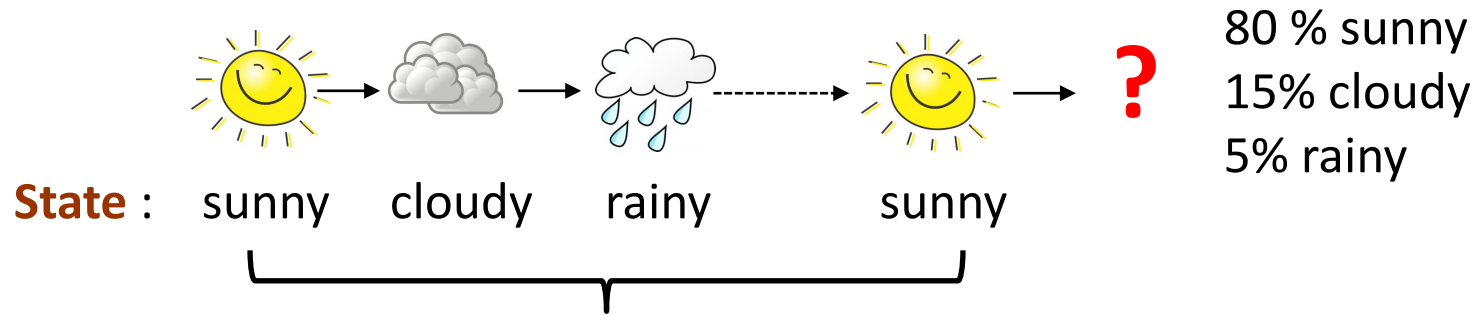
### Output format 2:

	sunny	rainy
sunny	0.9	0.1
rainy	0.3	0.7

### Output format 3:



# The Markov Model

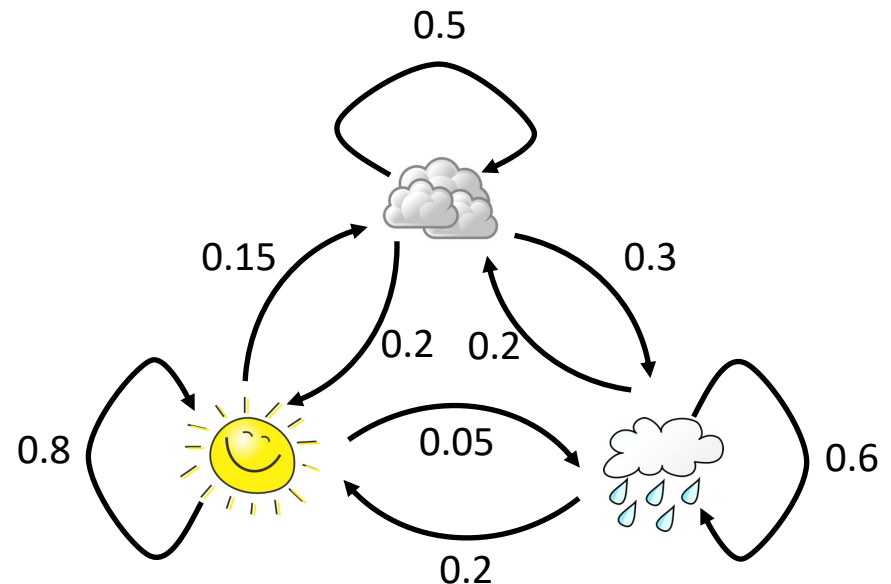


**State transition probability (table/graph)**

**Output format 1:**

Today	Tomorrow	Probability
sunny	sunny	0.8
sunny	rainy	0.05
sunny	cloudy	0.15
rainy	sunny	0.2
rainy	rainy	0.6
rainy	cloudy	0.2
cloudy	sunny	0.2
cloudy	rainy	0.3
cloudy	cloudy	0.5

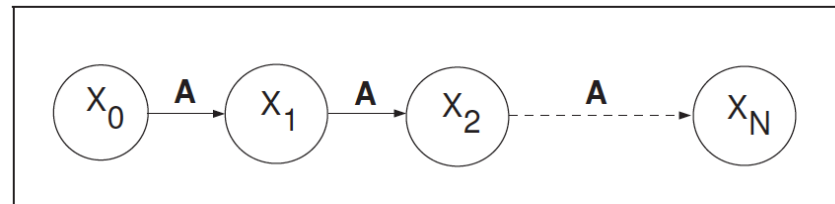
**Output format 3:**



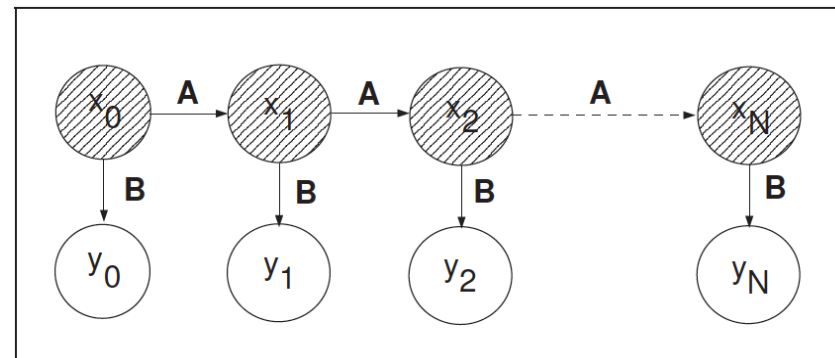
# The Hidden Markov Model

**A Hidden Markov Model** is a Markov chain for which the state is only partially observable.

**A Markov Model**



**A Hidden Markov Model**



**Hidden states** : the (TRUE) states of a system that can be described by a Markov process (e.g., the weather).

**Observed states** : the states of the process that are 'visible' (e.g., umbrella).

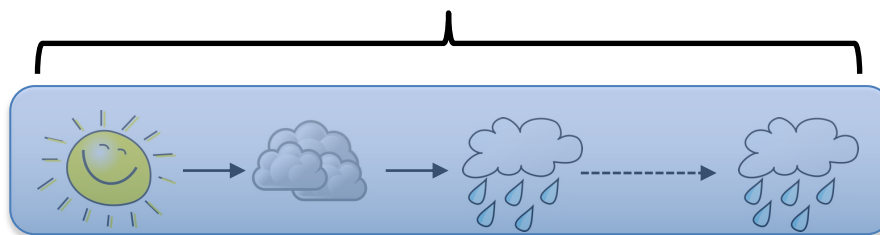
# The Hidden Markov Model

	sunny	rainy	cloudy
sunny	0.8	0.05	0.15
rainy	0.2	0.6	0.2
cloudy	0.2	0.3	0.5

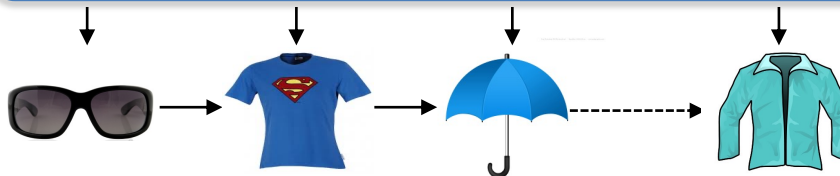
sum to 1

**State transition probability table**

**Hidden States**



**Observed States**



**State emission probability table**

	sunglasses	T-shirt	umbrella	Jacket
sunny	0.4	0.4	0.1	0.1
rainy	0.1	0.1	0.5	0.3
cloudy	0.2	0.3	0.1	0.4

sum to 1

The probability of observing a particular observable state given a particular hidden state

# The Hidden Markov Model

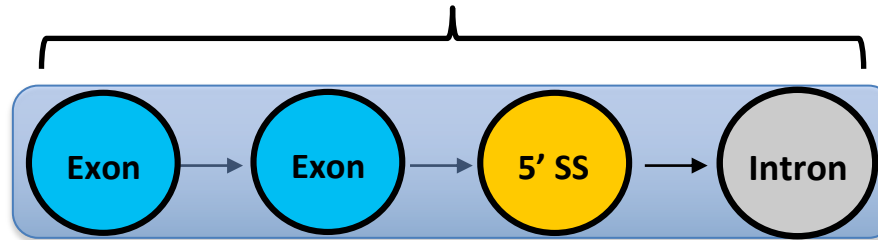
The probability of switching from one state type to another (ex. Exon - Intron).

	exon	5'SS	intron
exon	0.9	0.1	0
5'SS	0	0	1
intron	0	0	0.9

sum to 1

State transition probability table

Hidden States



Observed States



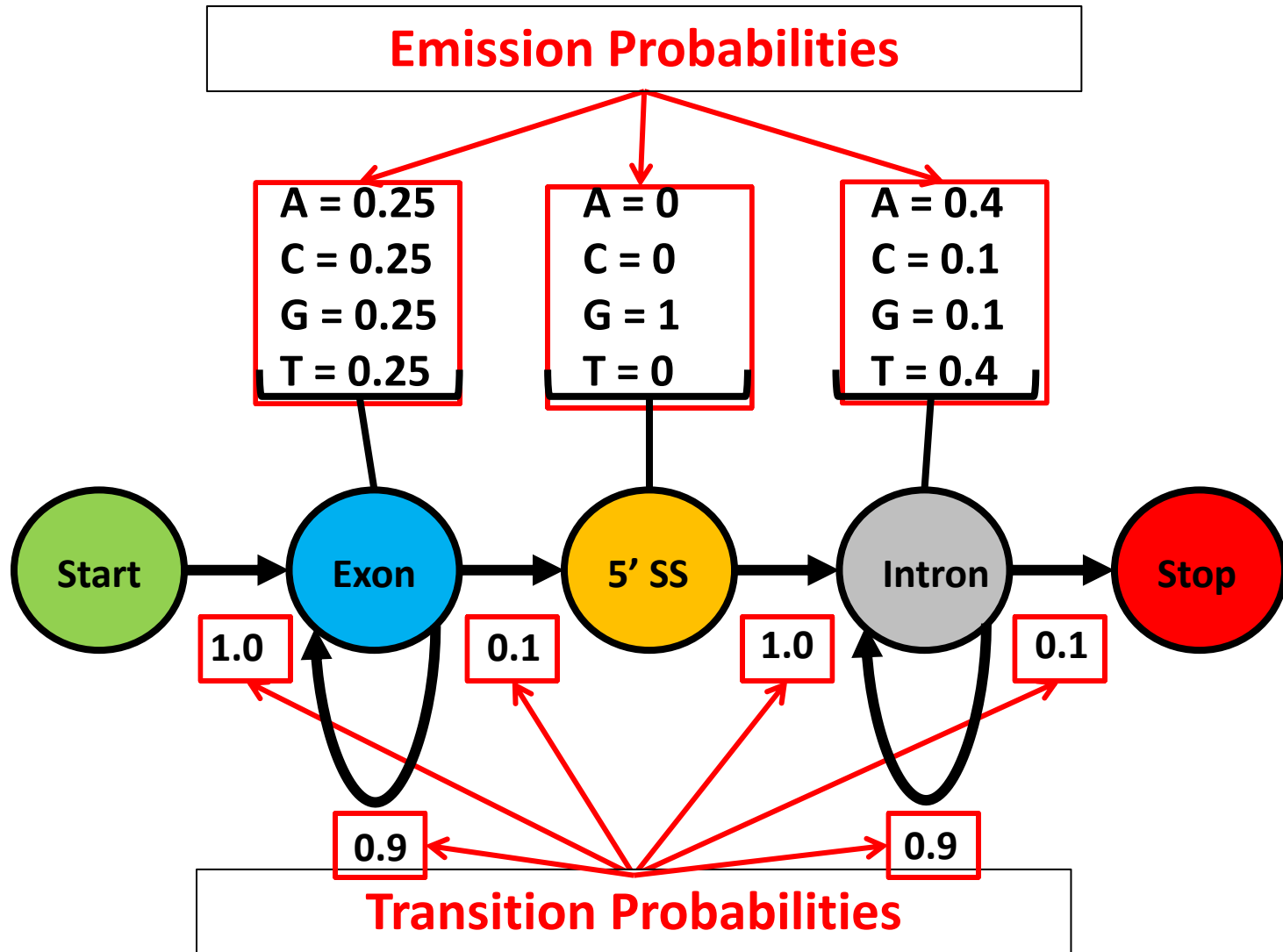
State emission probability table

	A	C	G	T
exon	0.25	0.25	0.25	0.25
5'SS	0	0	1	0
intron	0.4	0.1	0.1	0.4

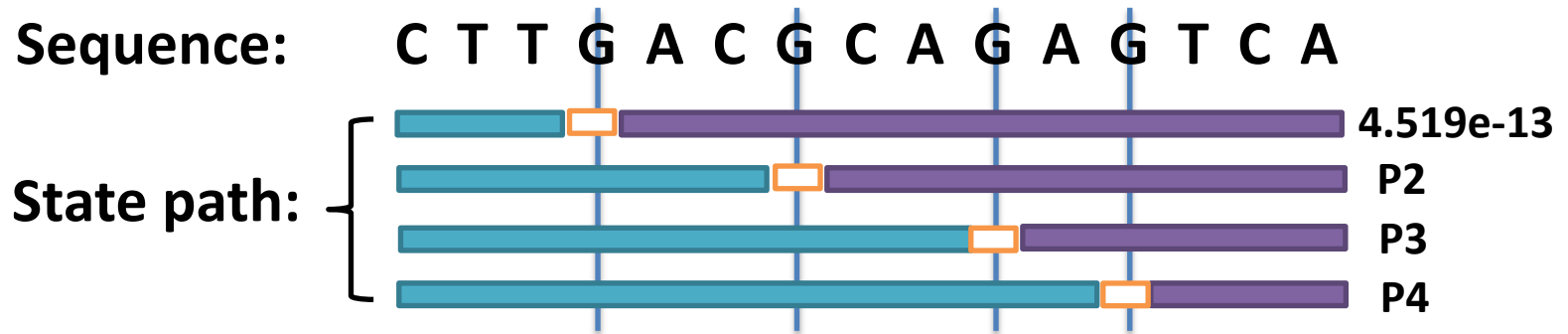
sum to 1

The probability of observing a nucleotide (A, T, C, G) that is of a certain state (exon, intron, splice site)

# The Hidden Markov Model



# Splicing Site Prediction Using HMMs



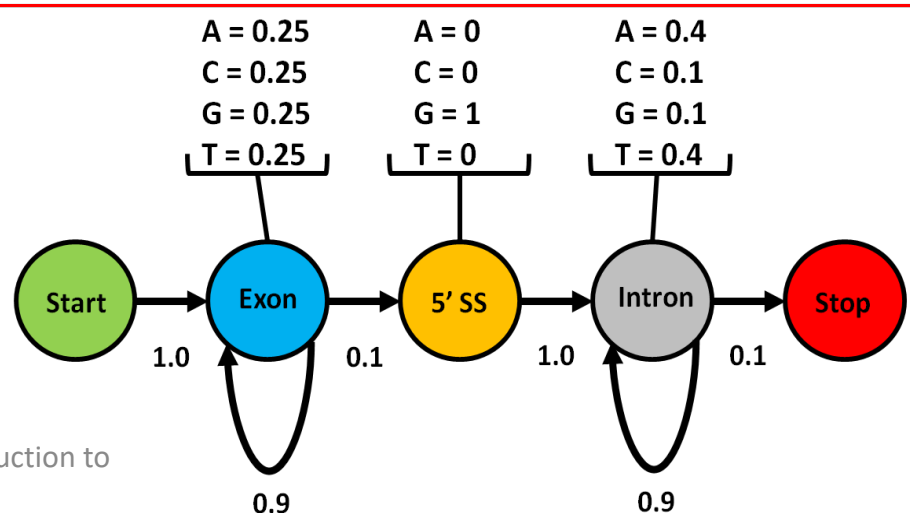
To calculate the *probability* of each state path, multiply all transition and emission probabilities in the state path.

$$\text{Emission} = (0.25^3) \times 1 \times (0.4 \times 0.1 \times 0.1 \times 0.1 \times 0.4 \times 0.1 \times 0.4 \times 0.1 \times 0.4 \times 0.1 \times 0.4)$$

$$\text{Transition} = 1.0 \times (0.9^2) \times 0.1 \times 1 \times (0.9^{10}) \times 0.1$$

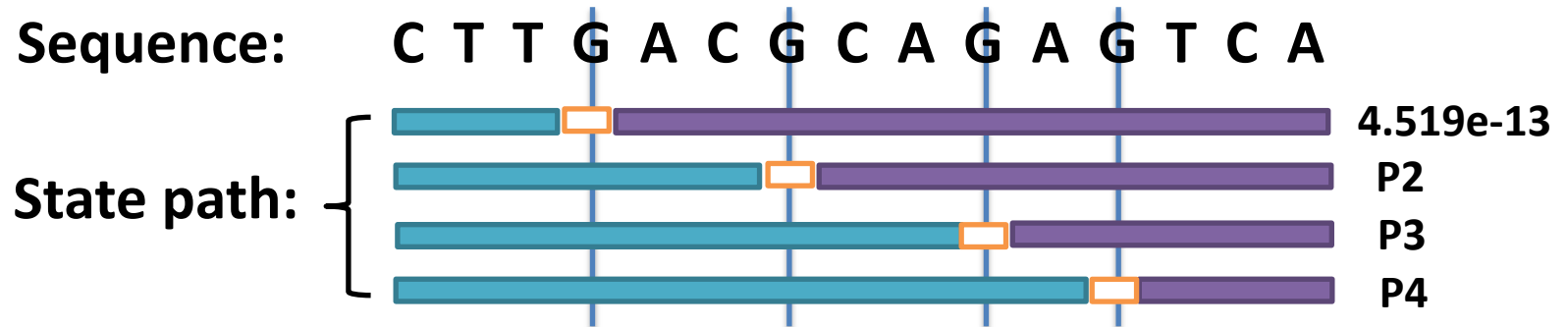
$$\begin{aligned} \text{State path} &= \text{Emission} \times \text{Transition} \\ &= 1.6e-10 \times 0.00282 \\ &= 4.519e-13 \end{aligned}$$

The state path with the highest probability is most likely the correct state path.





# Identification of the Most Likely Splice Site

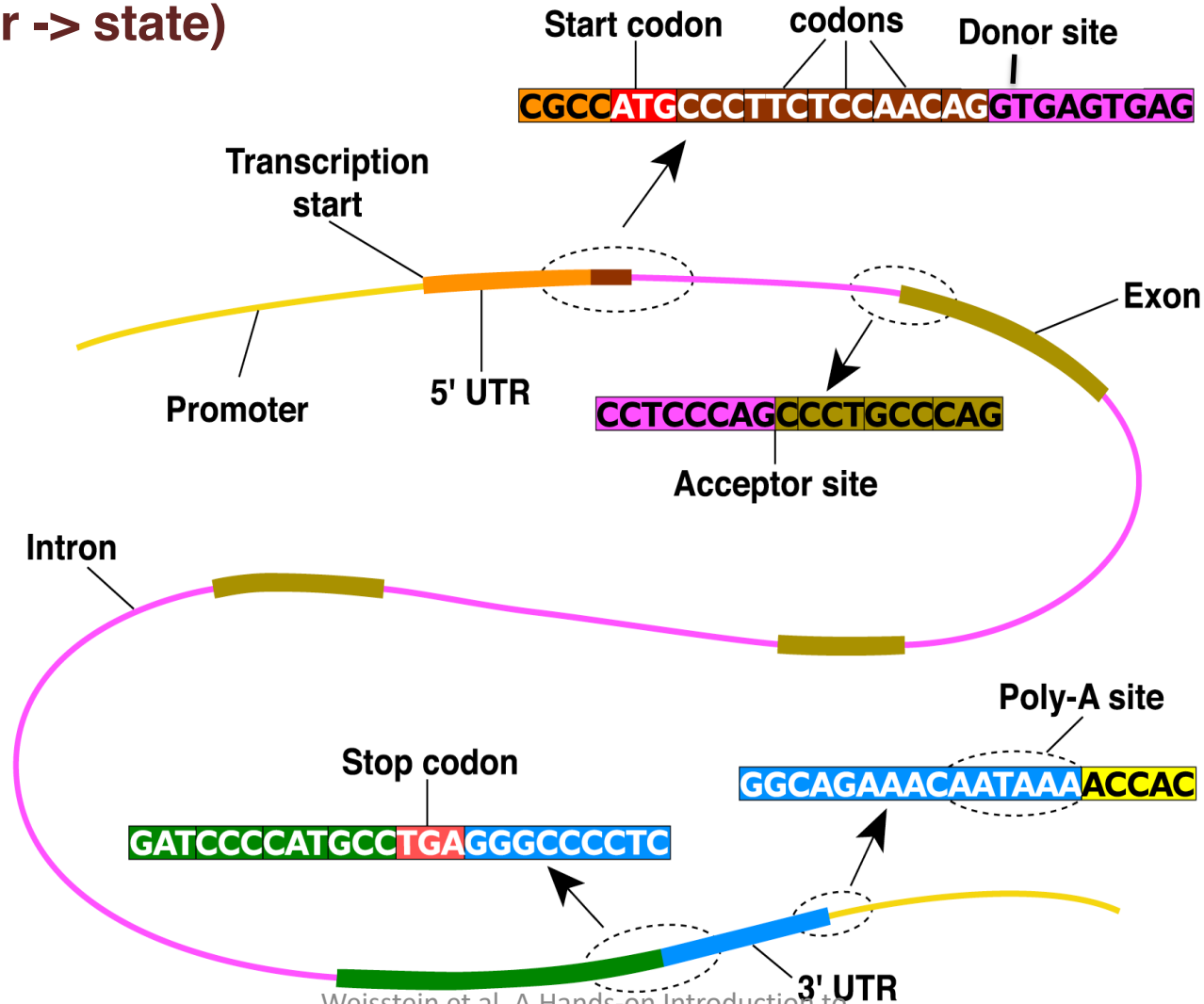


The *likelihood* of a splice site at a particular position can be calculated by taking the probability of a state path and dividing it by the sum of the probabilities of all state paths.

$$\text{likelihood of a splice site in state path \#1} = \frac{4.519\text{e-}13}{4.519\text{e-}13 + P2 + P3 + P4}$$

# HMMs and Gene Prediction

(color -> state)

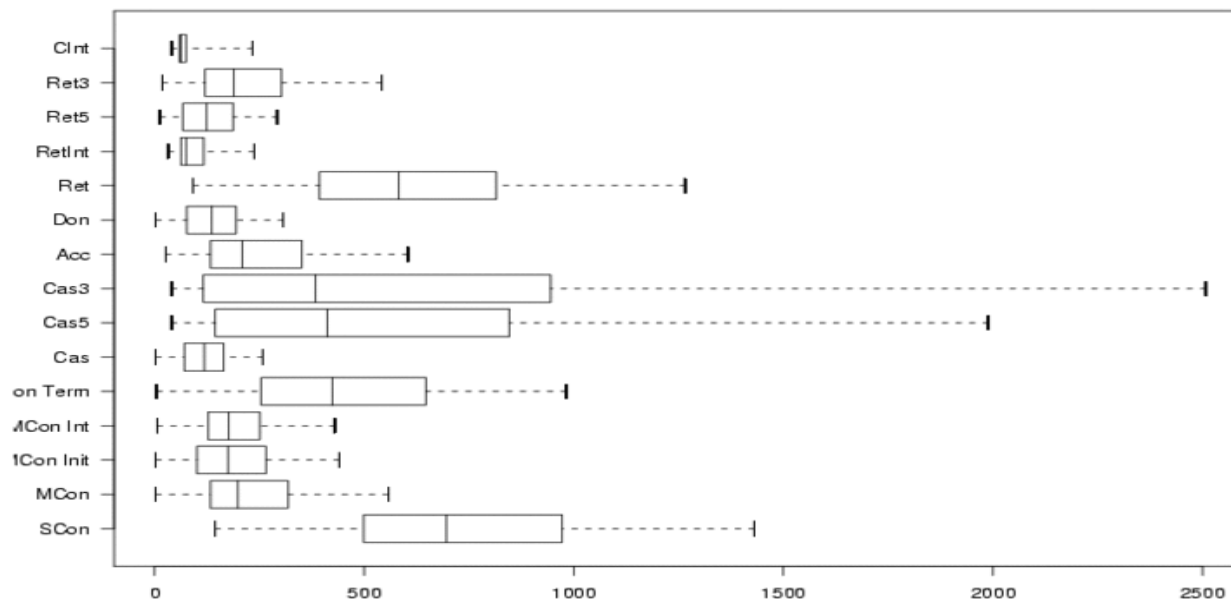


# HMMs and Gene Prediction

The accuracy of HMM gene prediction depends on **emission probabilities** and **transition probabilities**.

**Emission probabilities** are calculated based on the base composition in that particular state in the training data.

**Transition probabilities** are calculated based on the average lengths of that particular state in the training data.



**Exon length boxplots**  
(DEDB, *Drosophila melanogaster* Exon Database)

Homework Question: How do transition probabilities affect the length of predicted ORFs?

# Conclusions

- Hidden Markov Models have proven to be useful for finding genes in unlabeled genomic sequence. HMMs are the core of a number of gene prediction algorithms (such as Genscan, Genemark, Twinscan).
- Hidden Markov Models are machine learning algorithms that use ***transition probabilities*** and ***emission probabilities***.
- Hidden Markov Models label a series of observations with a ***state path***, and they can create multiple state paths.
- It is mathematically possible to determine which state path is most likely to be correct.