# 02-712 Week 13

Biological Modeling and Simulation

Aidan Jan

November 18, 2025

## Hidden Markov Models

Suppose we are given a 5' splice site recognition problem. We have a strand of DNA

```
--------------------------[EXON]---------|SPLICE SITE|--[INTRON]----------------
```

Basically, we have a strand of DNA composed of one exon followed by one intron. How do we decide where the splice site is, or where it switches from the exon to the intron? We have some statistical properties about the strand:

- An exon typically contains all four nucleotides an about-equal amount
- Splice sites are `G` base pairs 95% of the time, and the other 5% are `A`'s.
- Introns are `A-T` rich. Around 40% of introns are `A`'s and `T`'s (and around 10% `C`'s and `G`'s, each).

We can define this as a Markov model. We have 5 states:

- Start state (source node)
- Exon state ($E$)
- 5' splice state ($5'$)
- Intron state ($I$)
- End state (sink node)

We have the emission probabilities (model the base composition), and transmission probabilities between states $E \to 5' \to I$. For our example, assume that we start with

- State E:
  - Emission: {A = 0.25, C = 0.25, T = 0.25, G = 0.25}
  - Transition: {Self loop: 0.9, 5': 0.1}
- State 5':
  - Emission: {A = 0.05, C = 0, G = 0, T = 0.95}
  - Transition: {Self loop: 0, I: 1.0}
- State I:
  - Emission: {A = 0.4, C = 0.1, T = 0.4, G = 0.1}
  - Transition: {Self loop: 0.9, End state: 0.1}

Now, consider the sequence:

```
CTTCATGTGAAAGCAGACGTAAG
```

We can have a **state path**, which represents the states the Markov model passes through as it goes through the sequence one letter at a time (i.e., Markov Chain.)

```
Sequence:  CTTCATGTGAAAGCAGACGTAAG
Statepath: EEEEEEEE5IIIIIIIIIIIIII
```

In a Hidden Markov Model (HMM), the state path is hidden to us. It happens in the background based on the most likely probability of occurring. Our goal is given the model, we want to *infer* the statepath. If we are able to infer the statepath, then we can find out where the 5' splice site is.

## Calculating HMM Probabilities

The probability $p(S, \pi | \text{HMM}, \theta)$ that the HMM with parameter $\theta$ generates a state path $\pi$ and observed segment $S$ is the product of all emission and transition probabilities.

For our example here, we can understand the model's choice by calculating every possible statepath. The E state is only present before the 5' site and the I state is only present after the 5' site. Additionally, the 5' site can only be on an `A` or a `G`. Therefore, since there are only 14 different `A`'s or `G`'s in the sequence. We will use log probabilities since multiplying many probabilities less than 1 together usually leads to rounding errors.

Even then, many of the statepaths would have very similar probability values. So how confident are we that the most likely one is *actually* the statepath that occurred? For this, we use a process called **posterior decoding**.

$$p(\text{residue} + \text{emitted by state} k) = \frac{\sum p(\text{state paths that use } k \text{ to generate})}{\sum p(\text{all state paths})}$$