# Genome Annotation

## Introduction to hidden Markov models

# Outline

- Motivation for hidden Markov models
- hidden Markov model definition

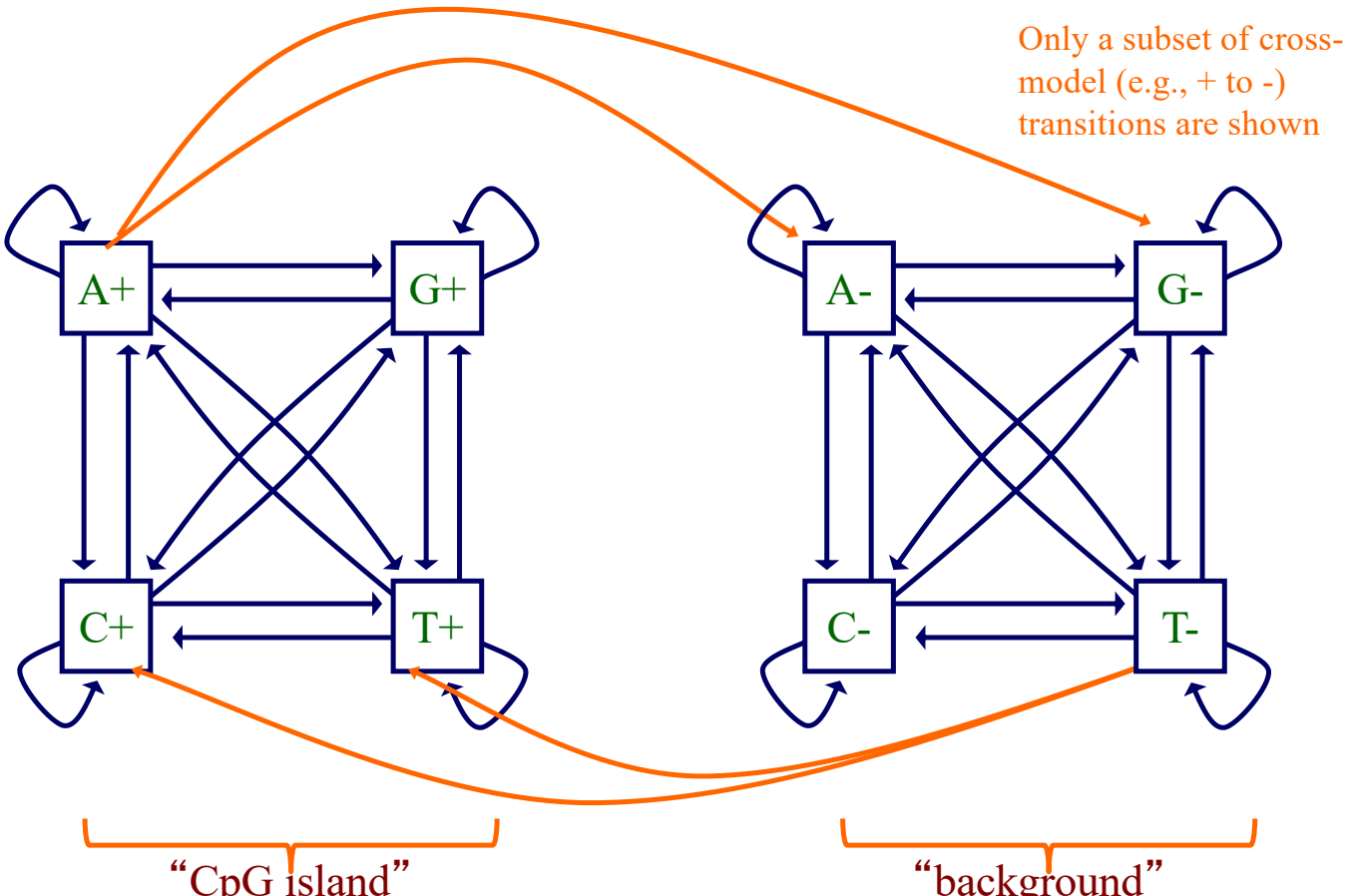
# Classifying sequences

- Markov chains
  - useful for modeling a single class of sequence
  - likelihood ratios of different models can be used to classify sequences
- What if a sequence contains multiple classes of elements?
  - Example: a whole genome sequence
  - How can we model such sequences?
  - How can we partition these sequences into their component elements?

# Revisiting the CpG question

- Given a sequence $x_1..x_L$ we can use two Markov chains to decide if $x_1..x_L$ is a CpG island or not.
- What do we do if we were asked to "find" the CpG islands in the genome?
- We have to search for these "islands" in a "sea" of non-islands

# One attempt: merge Markov chains

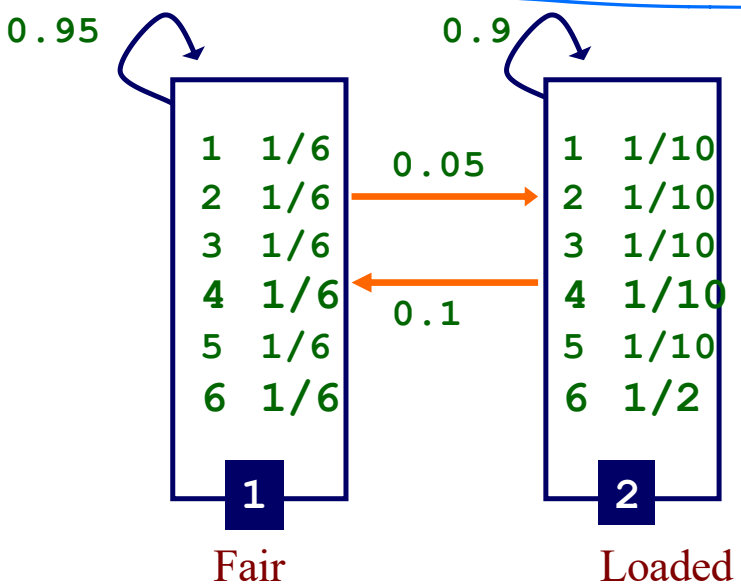



- Problem: when we observe a DNA sequence, we do not observe "A+" or "A-", we simply observe "A", for example.

# Hidden State

- We'll distinguish between the *observed* parts of a problem and the *hidden* parts
- in the Markov models we've considered previously, it is clear which state accounts for each part of the observed sequence
- in the model above, there are multiple states that could account for each part of the observed sequence
  - this is the hidden part of the problem

# An HMM for an occasionally dishonest casino




What is hidden? Which die is rolled

What is observed? Number (1-6) on the die

# Two HMM random variables

- Observed sequence

$$X = X_1, X_2, \ldots, X_L$$

- Hidden state sequence

$$\pi = \pi_1, \pi_2, \ldots, \pi_L$$

- HMM:
  - Markov chain over *hidden* sequence
  - Dependence between $\pi_i$ and $X_i$

# The Parameters of an HMM

- as in Markov chain models, we have transition probabilities

$$a_{kl} = \Pr(\pi_i = l \mid \pi_{i-1} = k)$$

probability of a transition from state $k$ to $l$

$\pi$ represents a path (sequence of states) through the model

- since we've decoupled states and characters, we also have emission probabilities

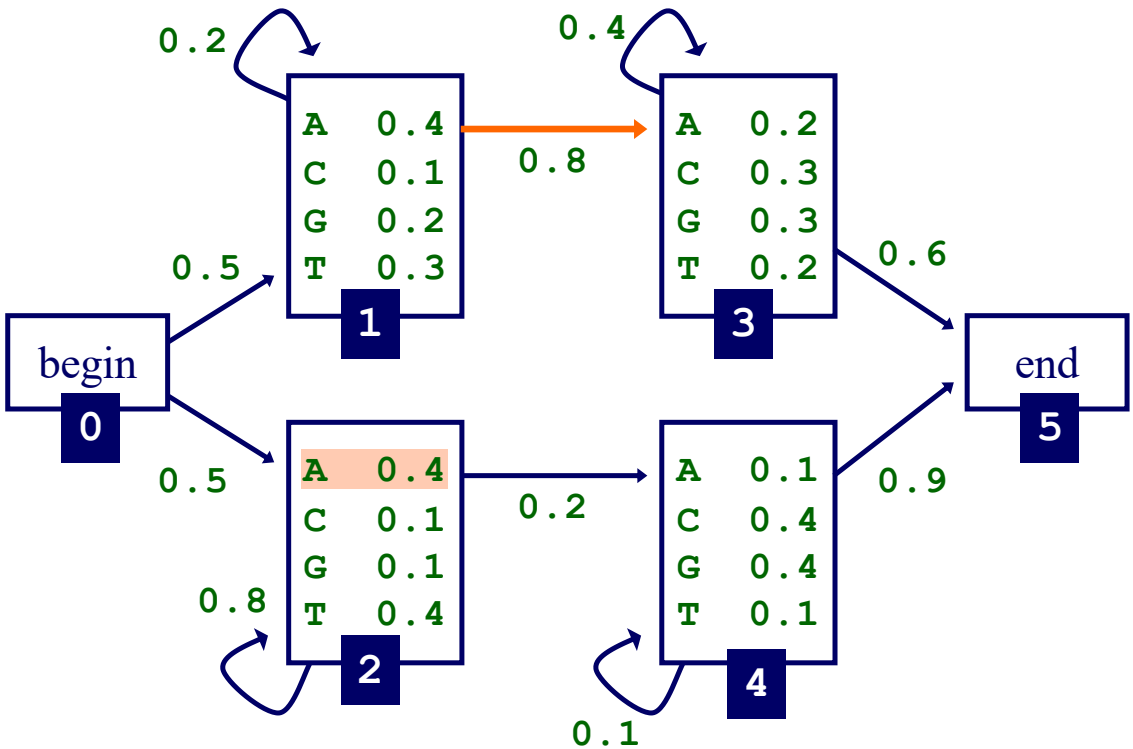
$$e_k(b) = \Pr(X_i = b \mid \pi_i = k)$$

probability of emitting character b in state $k$

# A Simple HMM with Emission Parameters

$a_{13}$ probability of a transition from state 1 to state 3

$e_2(\mathrm{A})$ probability of emitting character *A* in state 2



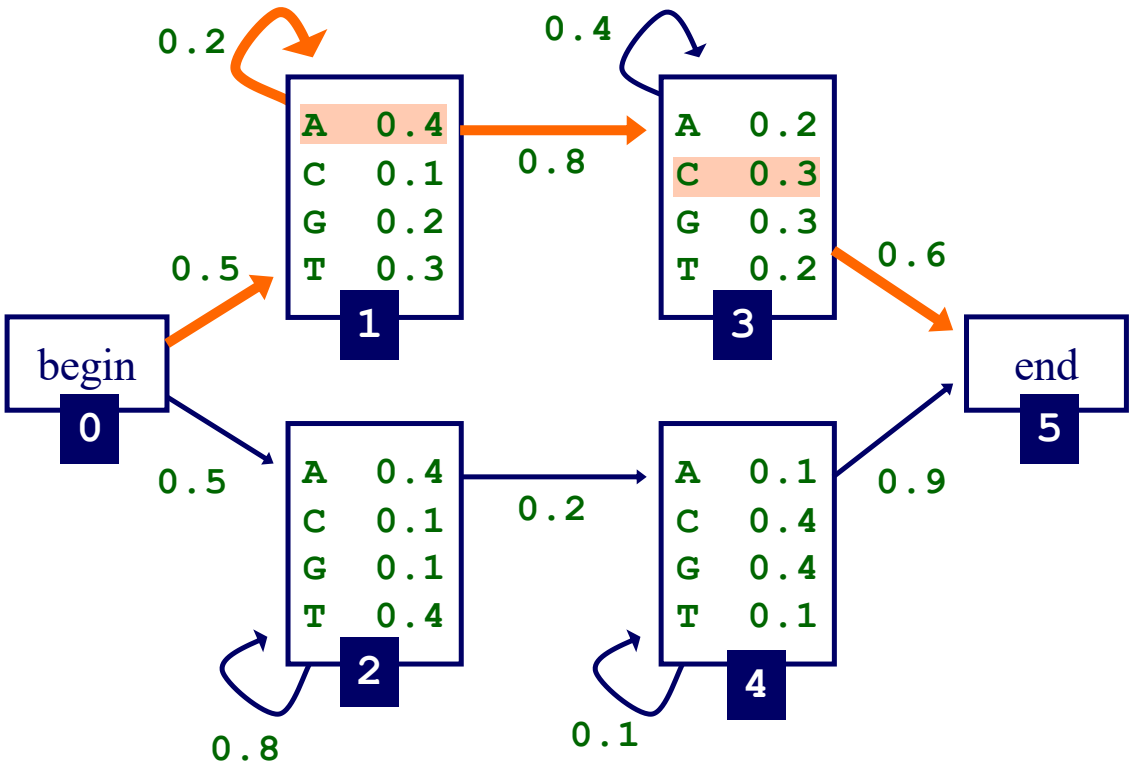
# How Likely is a Given Path and Sequence?

- the probability that the path $\pi_1 .. \pi_L$ is taken and the sequence $X_1 ... X_L$ is generated:

$$\Pr(X_1 ... X_L, \pi_1 .. \pi_L) = a_{0\pi_1} a_{\pi_L N} \prod_{i=1}^{L-1} a_{\pi_i \pi_{i+1}} \prod_{i=1}^{L} e_{\pi_i}(X_i)$$

(assuming begin/end are the only silent states on path)

# How Likely Is A Given Path and Sequence?




$$\Pr(\text{AAC},113) = a_{01} \times e_1(\text{A}) \times a_{11} \times e_1(\text{A}) \times a_{13} \times e_3(\text{C}) \times a_{35}$$

$$= 0.5 \times 0.4 \times 0.2 \times 0.4 \times 0.8 \times 0.3 \times 0.6$$

# Three Important Questions

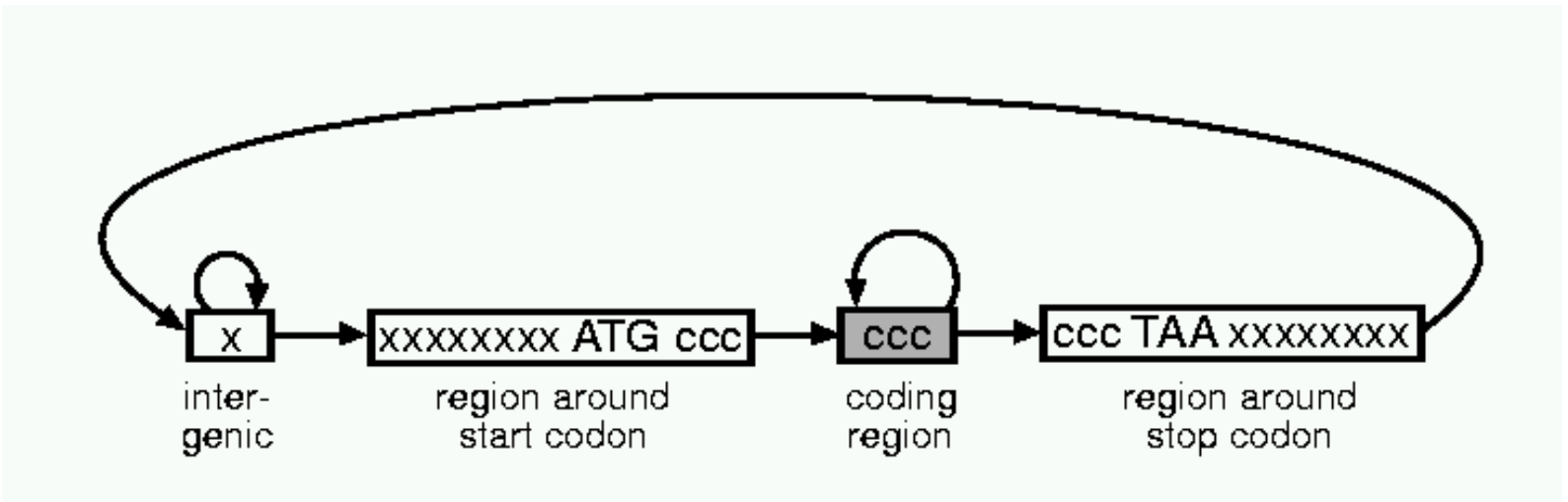
- How likely is a given sequence?
  the Forward algorithm

- What is the most probable "path" (sequence of hidden states) for generating a given sequence?
  the Viterbi algorithm

- How can we learn the HMM parameters given a set of sequences?
  the Forward-Backward (Baum-Welch) algorithm

# Simple HMM for Gene Finding

Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences



# Summary

- Hidden Markov models are extensions to Markov chains enabling us to model and segment sequence data
- HMMs are defined by a set of states and emission characters, transition probabilities and emission probabilities