



# Ab initio methods

---

- Use information embedded in the genomic sequence *exclusively* to predict the gene structure.
- Find structure  $G$  representing gene boundaries + internal gene structure which maximizes the probability  $P(G|\text{genomic sequence})$ .
- Hidden Markov models are the predominant generative method for modeling the problem.

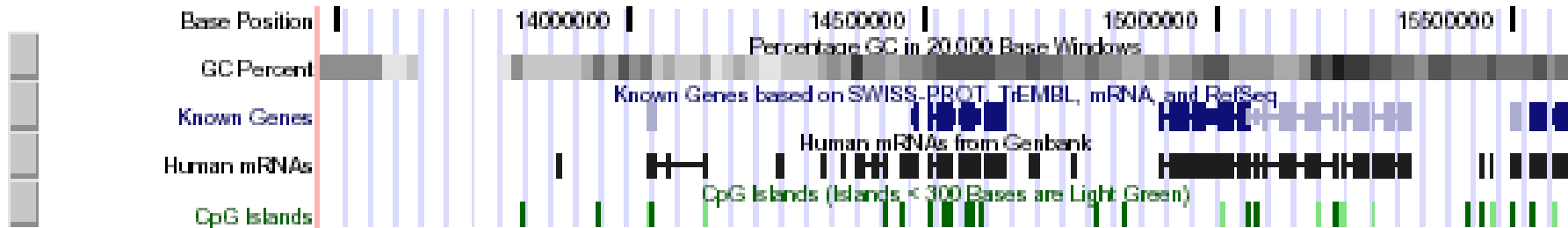


# Ab-initio methods

---

- Advantages
  - Intuitive, natural modeling
  - Prediction of 'novel' genes, *i.e.*, with no a priori known cDNA or protein evidence
- Caveats
  - Not effective in detecting alternatively spliced forms, interleaved or overlapping genes
  - Difficulties with gene boundary identification
  - Potentially large number of false positives with over-fitting

# A simple example: CpG Islands

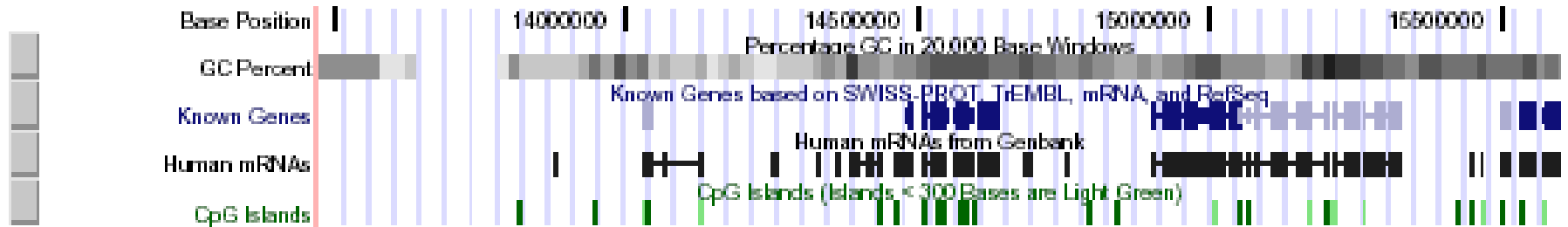


CpG nucleotides in the genome are frequently methylated. (Write CpG not to confuse with CG base pair)



Methylation often suppressed around genes, promoters  $\rightarrow$  CpG islands

# Example: CpG Islands



In CpG islands,  
CG is more frequent than in the rest of the  
genome

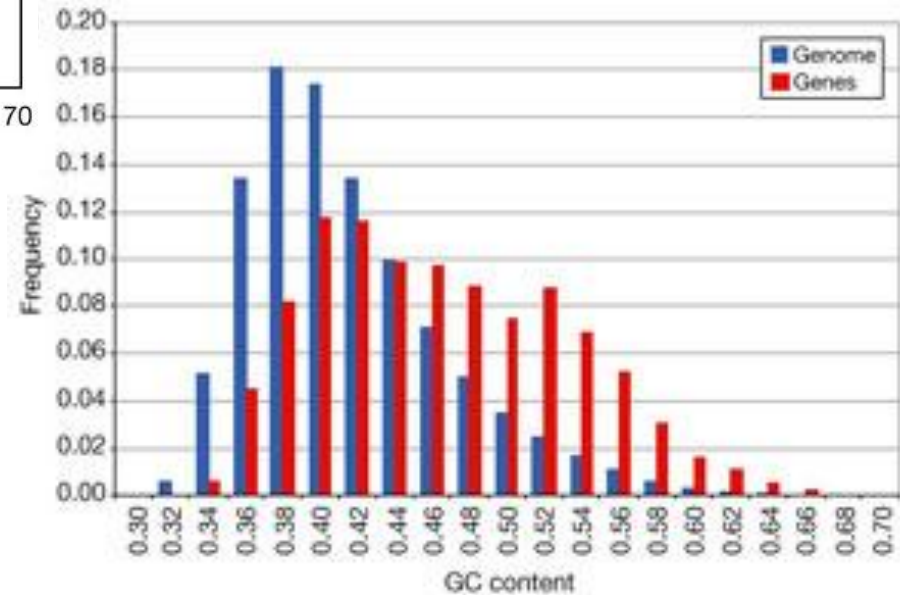
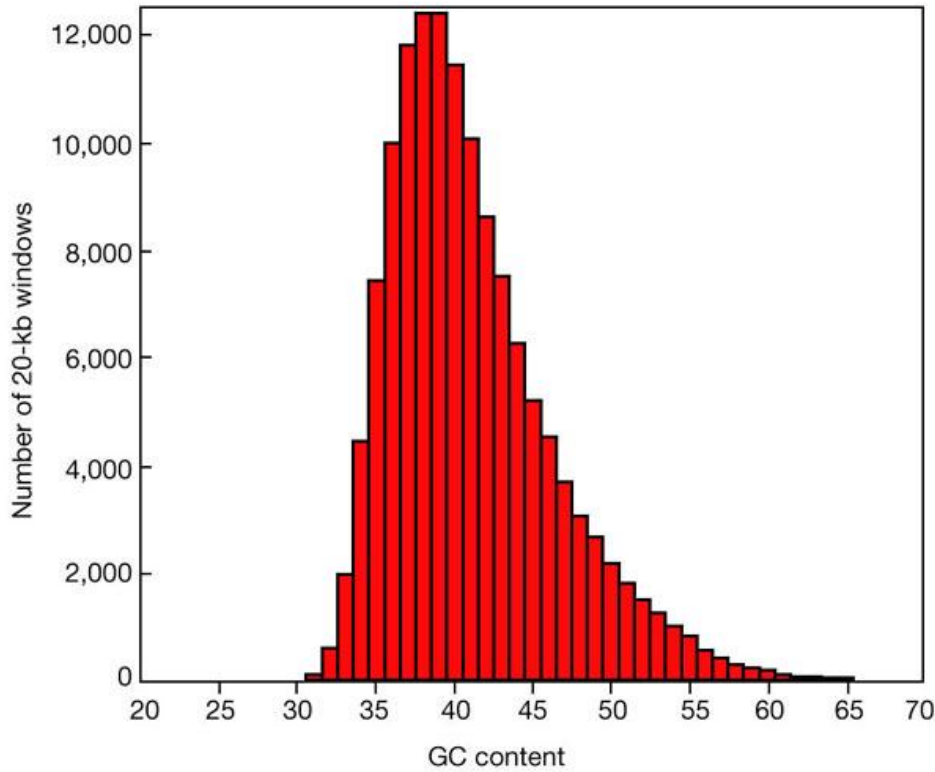


# GC content

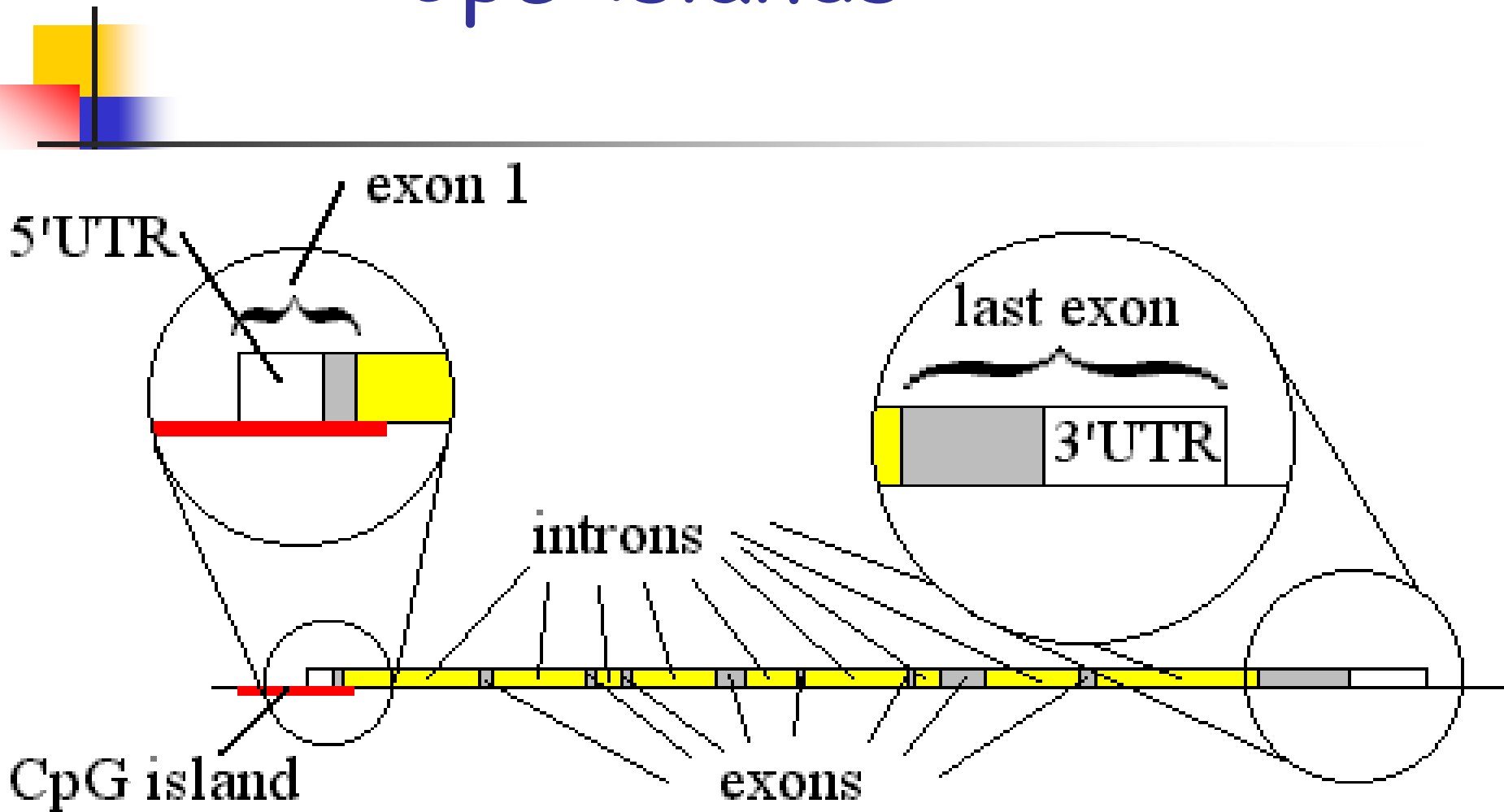
---

- Genes tend to be rich in GC content
- GC-rich genomes are more thermodynamically stable

# GC content of the human genome: mean 41%



# CpG islands





# Two problems

---

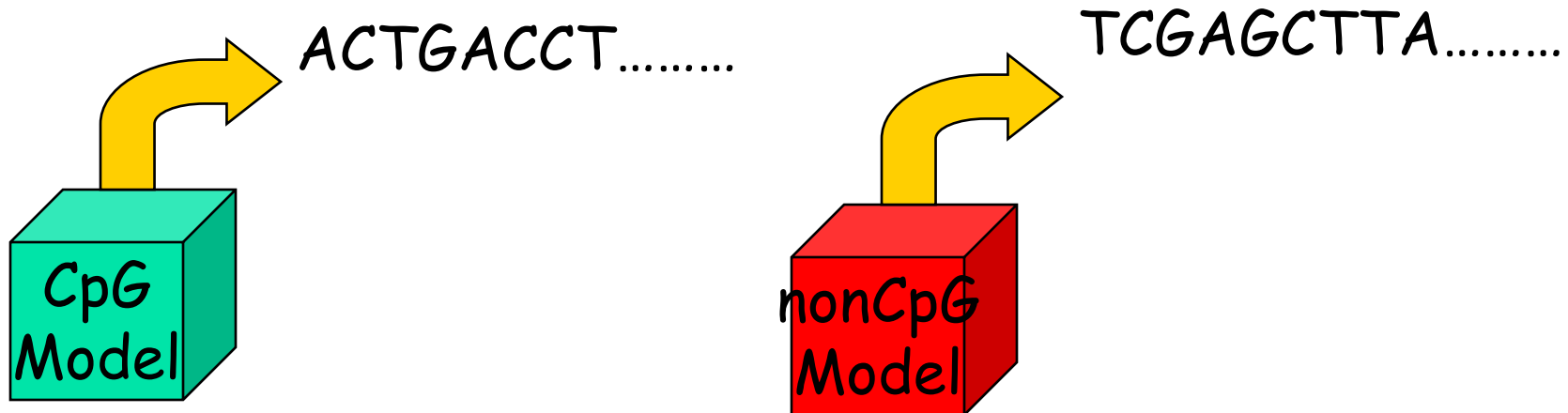
- Given a short DNA sequence, does it come from a CpG island or not?
  - Is this part of a CpG island or not?
- How to find the CpG islands in a long sequence?





# Generative models

---



Models generate sequences of strings in the  $A, T, C, G$  alphabet. Model parameters are tuned to reflect characteristics of CpG and non CpG islands.



# Markov processes: a quick intro

---

- We are interested in predicting weather ( $w$ ), which can be either be sunny ( $s$ ) or rainy ( $r$ ).  $s$  and  $r$  are values of the random variable  $w$ .
- The weather on a given day depends only on the weather on the previous day.

$$P(w_t | w_{t-1}, \dots, w_1) = P(w_t | w_{t-1})$$

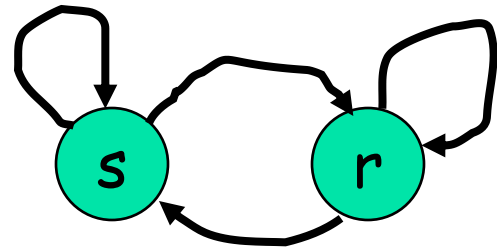
This is the **Markov** property.

# Markov process example

- We have knowledge of the **transition probabilities** between sunny and rainy days.

Rows of the transition matrix sum to 1.

$$\begin{array}{c} \text{s} \\ \text{r} \end{array} \begin{array}{cc} \text{s} & \text{r} \\ \left[ \begin{array}{cc} 0.9 & 0.1 \\ 0.5 & 0.5 \end{array} \right] \end{array}$$



- We know the **initial probabilities** of s and r.



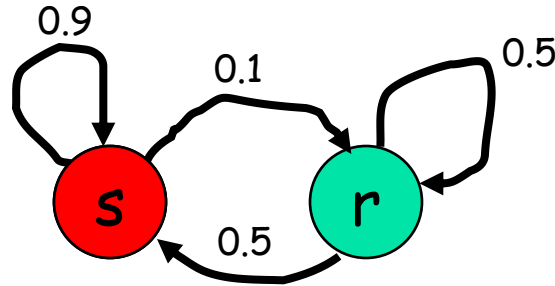
# Generating weather sequences

---

- Lets say we start with a sunny day.
- Now we consult our transition matrix and find that  $P(w_t | w_{t-1}=s) = [0.9 \ 0.1]$ . It is more likely that the next day will be sunny too.
- We repeat this process, flipping coins biased by the probability  $P(w_t | w_{t-1})$  to get a sequence representing weather for a consecutive set of days.

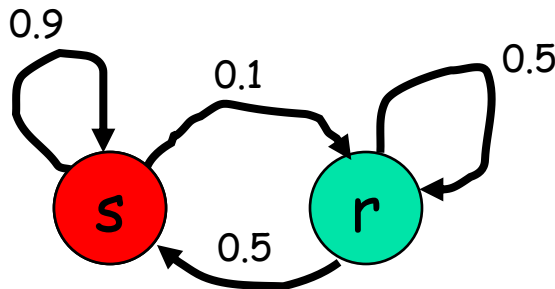
# Generating sequences

$$\begin{matrix} & \begin{matrix} s & r \end{matrix} \\ \begin{matrix} s \\ r \end{matrix} & \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} \end{matrix}$$

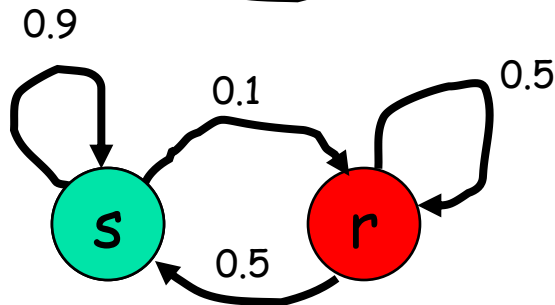


sequence

s



s



r



# Prediction

---

- Suppose day 1 is rainy . We will represent this as a vector of probabilities over the two values.

$$p(1) = [0 \ 1];$$

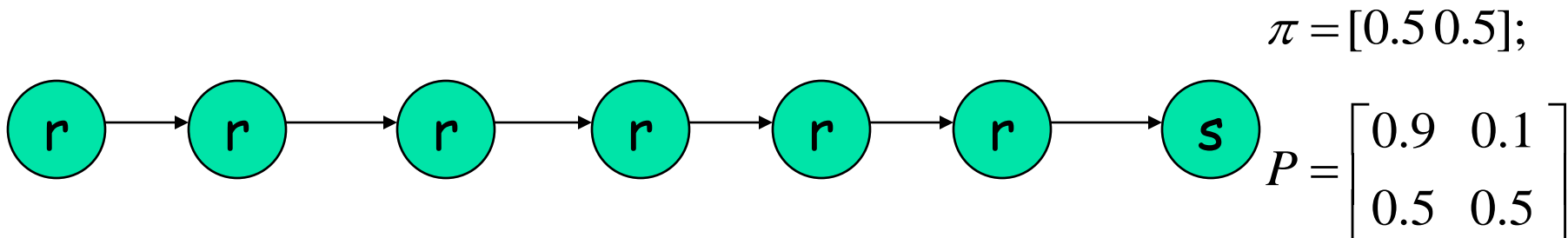
- How do we predict weather on day 2 given  $p(1)$  and the transition probabilities  $P$ ?
- From  $P$ , we can see that the probability of day 2 being sunny is .5, and for being rainy is 0.5

$$p(1) * P = [0.5 \ 0.5];$$

# Probability of a sequence

- Given a Markov model specified by an initial state probability vector  $\pi$ , and a transition probability matrix  $P$ , what is the probability of observing the sequence "rrrrrrs"?

$$\begin{aligned} P(X = rrrrrrs) &= \pi(r)P(r|r)P(r|r)P(r|r)P(r|r)P(r|r)P(s|r) \\ &= \pi(r) \prod_{t=2..7} P(x_t | x_{t-1}) = (0.5)^7 \end{aligned}$$



# Which weather pattern is more likely?

- Given a transition model

$$\begin{array}{c} s \\ r \end{array} \begin{array}{cc} s & r \\ \left[ \begin{array}{cc} 0.9 & 0.1 \\ 0.5 & 0.5 \end{array} \right] \end{array}$$

- And an initial state distribution: [0.5 0.5]
- And two sequences: rrrrrrs and ssssssr  
Which is more likely, given the model?



# Comparing likelihoods

---

$$P(X = rrrrrrs | Model) = \pi(r)[P(r | r)]^5 P(s | r) = (0.5)^7$$

$$P(X = ssssssr | Model) = \pi(s)[P(s | s)]^5 P(r | s) = 0.5 * (0.9)^5 * 0.1$$



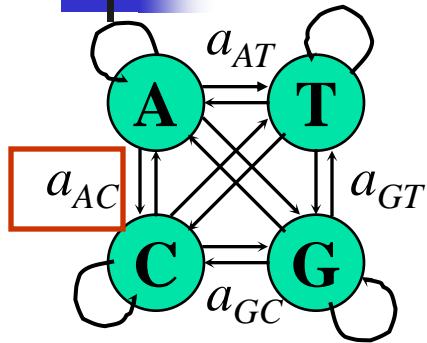
# Markov models (summary)

---

- States:  $S = \{s_1, \dots, s_N\}$ ,  $N$  states
- Transition probability:
  - $a_{ij} = P(X_{t+1}=s_j | X_t=s_i)$ ,  $i, j$  in  $[1..N]$
- Initial state probability
  - $pi_i = P(X_1=s_i)$ ,  $i$  in  $[1..N]$

Model generates sequences of states from  $S$ , and we can compute how likely a sequence is given the model.

# Markov models for CpG islands



A state for each of the four letters A, C, G, and T in the DNA alphabet

$$a_{st}^+ = \frac{c_{st}^+}{\sum_{t'} c_{st'}^+}$$

From a set of known CpG islands, and non CpG islands, estimate the transition probabilities

+	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	.274	.188
G	.161	.339	.375	.125
T	.079	.355	.384	.182

-	A	C	G	T
A	.300	.205	.285	.210
C	.322	.298	.078	.302
G	.248	.246	.298	.208
T	.177	.239	.292	.292



# Using the model

---

- To use the model for classification of a given sequence, calculate the log-odds ratio.
- Is the sequence more likely to come from a CpG island or a non-CpG region?

$$P(x | CpG) > P(x | nonCpG)$$

$$\frac{P(x | CpG)}{P(x | nonCpG)} > 1$$

$$\log \frac{P(x | CpG)}{P(x | nonCpG)} > 0$$



Log-odds ratio

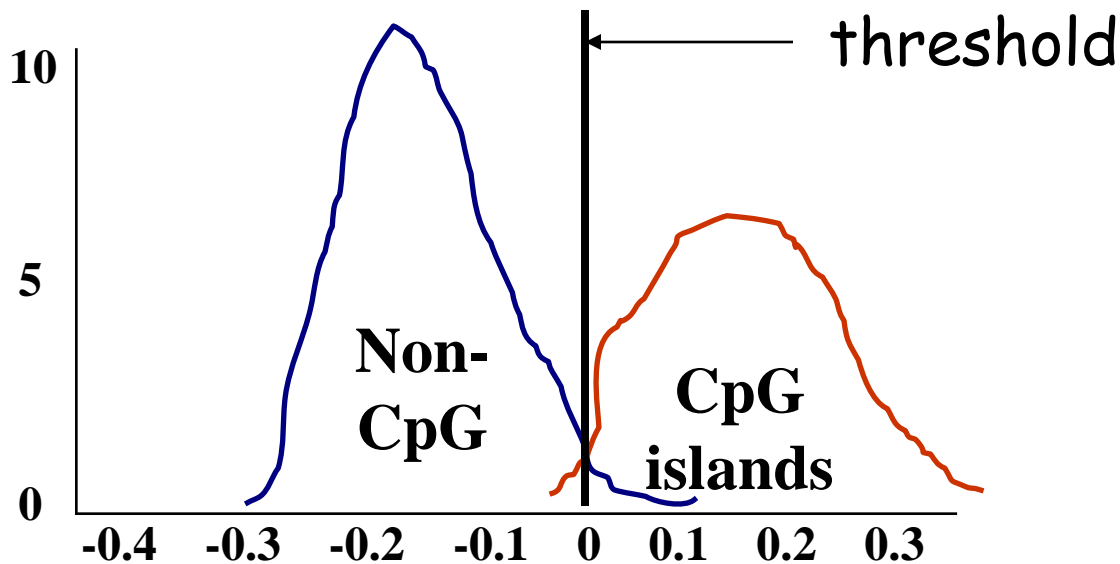


# The log-odds ratio

---

$$S(x) = \log \frac{P(x/CpG)}{P(x/nonCpG)} = \sum_{i=1}^L \log \frac{a_{x_{i-1}x_i}^+}{a_{x_{i-1}x_i}^-}$$

# Histogram of log-odds scores



Given a short sequence  $x$ , does it come from CpG island (Yes-No question)?

**Decision rule: if  $S(x) > 0$  then CpG else non-CpG**



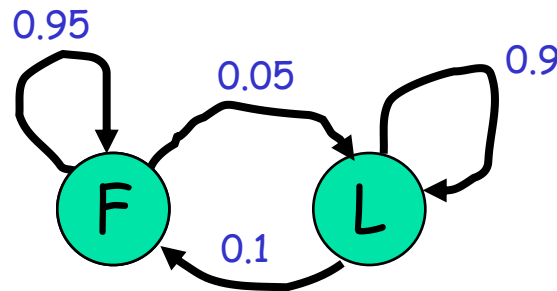
# How to locate CpG islands?

---

- Given a DNA sequence, find the CpG islands in it, if any.
- Approach: Calculate the log-odds score for a window of  $w$  nucleotides around every base in the sequence. Predict as CpG islands, those with a positive log-odds score.
- Problem: What should the size of the window  $w$  be? Predictions are sensitive to choice of  $w$ .

# The occasionally dishonest casino

- A casino uses a fair coin most of the time, but occasionally they switch to a loaded coin. You can't see which coin they are using, just the results of the flips (heads and tails) are visible.



h:0.5  
t:0.5

h:0.1  
t:0.9

Hidden state

Observables

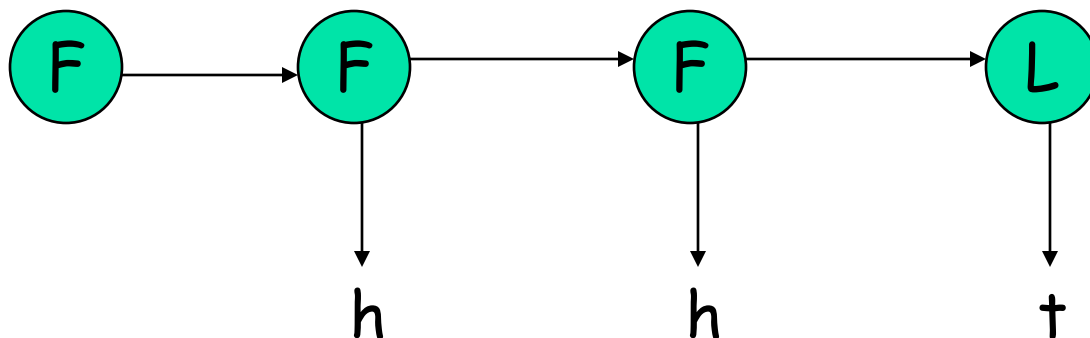
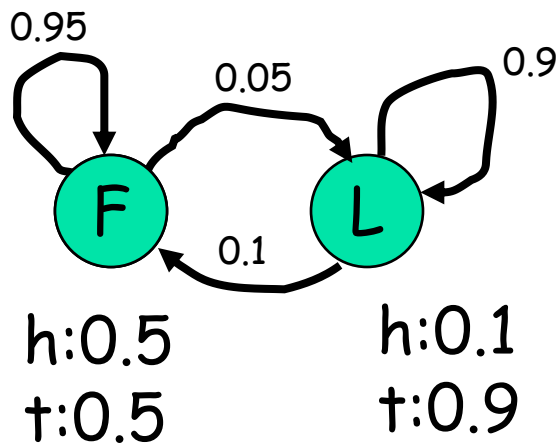


# Generating coin flips

---

- Start in one of the states, F or L (i.e., pick a fair or loaded coin to start with) (initial probabilities).
- Move to the next state (F or L), based on the transition probabilities. Generate an h or t based on the emission probabilities of that state.
- Repeat above step.

# Generating flips (take 2)



State sequence: FFFL (unobserved)  
Obs sequence : hht (observed)



# Hidden Markov Models

---

- $S = \{s_1, \dots, s_N\}$ ,  $N$  states
- $O = \{o_1, \dots, o_M\}$ ,  $M$  observation symbols
- $a_{ij} = P(S_{t+1}=s_j | S_t=s_i)$ ,  $i, j$  in  $[1..N]$ ; **transition probabilities**
- $b_i(k) = P(E_t=o_k | S_t=s_i)$ ,  $k$  in  $[1..M]$ ,  $i$  in  $[1..N]$ ; **emission probabilities**
- $\pi_i = P(S_1=s_i)$ ,  $i$  in  $[1..N]$ ; **initial state probabilities**

$\lambda = (A, B, \pi)$  specifies the HMM model



# Dishonest casino as an HMM

---

- $N = 2, S = \{F, L\}$

- $M = 2, O = \{h, t\}$

- $A =$ 

	F	L
--	---	---

F	$\begin{bmatrix} 0.95 & 0.05 \end{bmatrix}$
L	$\begin{bmatrix} 0.10 & 0.90 \end{bmatrix}$

- $B =$

	h	t
F	$\begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$	
L	$\begin{bmatrix} 0.1 & 0.9 \end{bmatrix}$	

- $\pi = [1 \ 0]$