# Hidden Markov Models

# Markov Chains

- Simplest possible information graph structure: just a linear chain $X \to Y \to Z \to \cdots$.
- Obeys Markov property: joint probability factors into conditional probabilities that only depend on the *previous* variable

$$p(X, Y, Z, \cdots) = p(X)p(Y|X)p(Z|Y)\cdots$$

- Simple but useful for a remarkable range of problems in many fields.
- Widely used in bioinformatics: detecting features in sequences; sequence alignment; modeling evolution etc.

# Markov Model State Graphs

- Markov chains have a generic information graph structure: just a linear chain $X \to Y \to Z \to \cdots$.
- The more interesting aspect of how to build a Markov model is deciding *what states* it consists of, and what *state transitions* are allowed.
- This is represented by its state graph.
- Do not mix this up with an information graph!

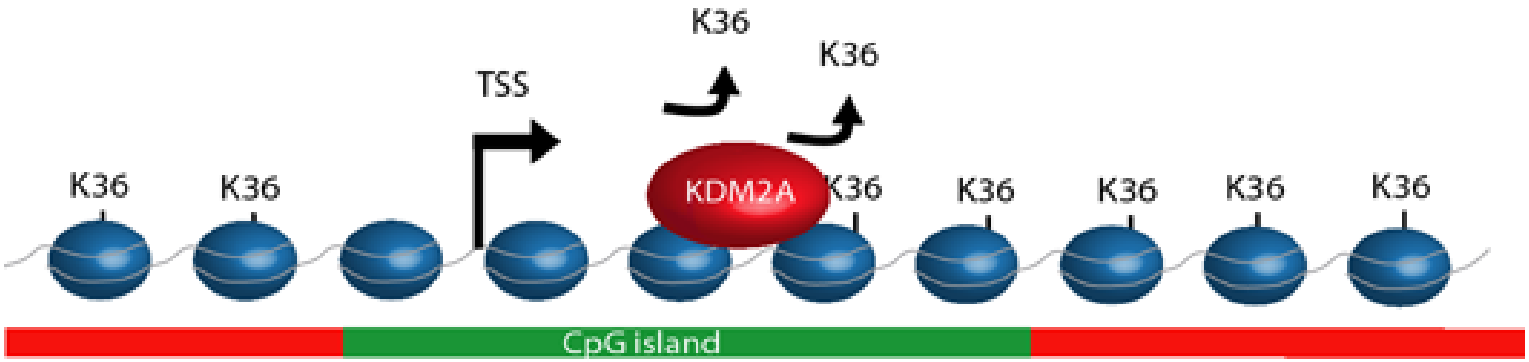
# Example: Occasionally Dishonest Casino

In a dice game, a casino uses a fair dice (all rolls have equal probability) and a loaded dice, but switches between them randomly:

- after each roll of the fair dice, they switch to the loaded dice with 5% probability.
- after each roll of the loaded dice, they switch to the fair dice with 10% probability.

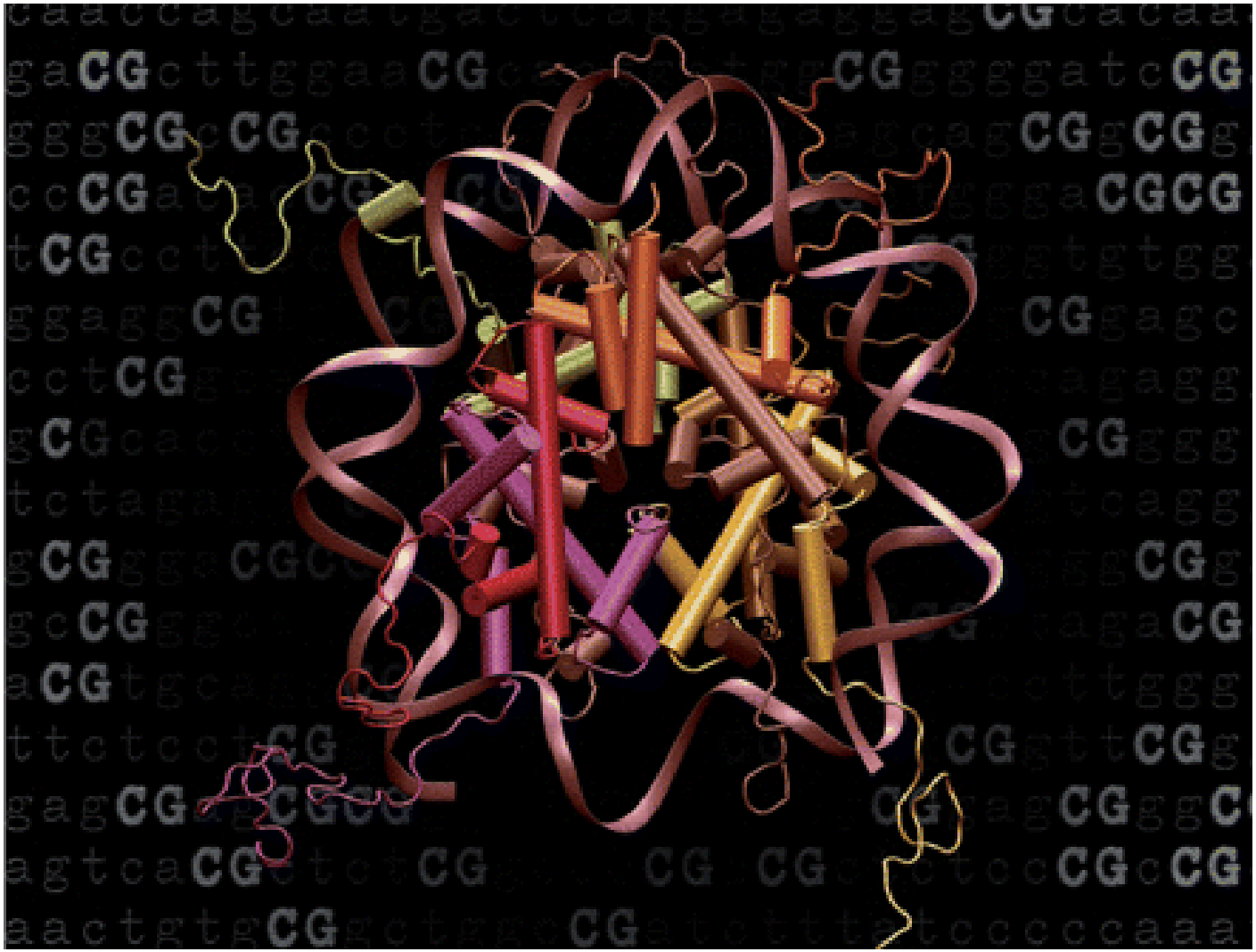
What's the state graph?

# Example: CpG Islands




- CpG islands are segments of DNA with unusually high frequencies of CG dinucleotide (i.e. a C followed by a G; the “p” refers to the phosphate linker on the DNA backbone).
- CpG islands are often binding sites for proteins that regulate the expression of a nearby gene, and *methylation* (a chemical modification of DNA that occurs in cells) of CpG islands helps control these regulatory interactions.
- CpG islands play an important role in tumors, which use them to turn off genes that block tumor growth.

# Nucleosome Binding To CpG Islands



# Modeling CpG Islands as a Markov Chain

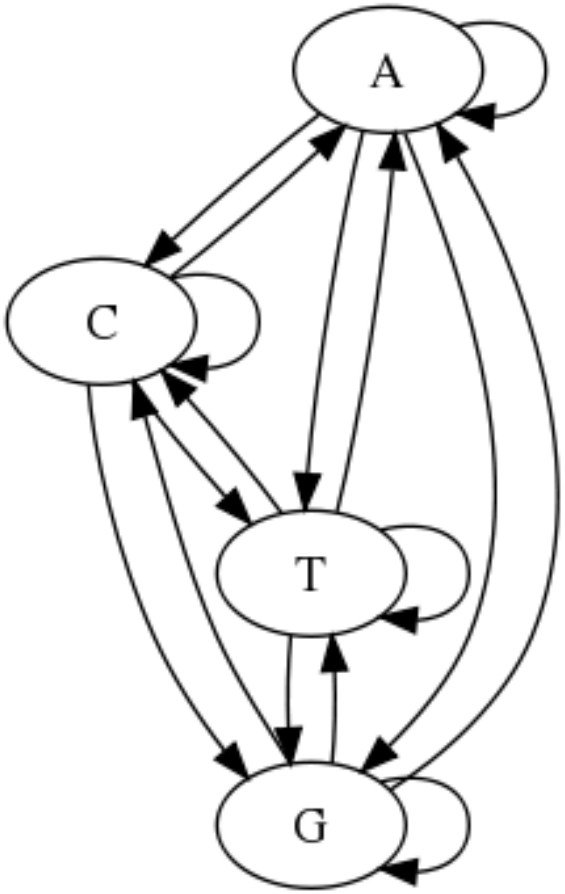
CpG islands show very different conditional probabilities $p(X_{t+1}|X_t)$ than non-CpG island sequence. This suggests we can use a Markov chain model to detect them in any sequence.

- Given a 4 x 4 table of the conditional probabilities $p(X_{t+1}|X_t)$ measured in CpG islands, describe how you would construct a Markov chain model of CpG island sequences. Specifically, describe the state graph structure you would use:
  - what are its nodes?
  - what edges exist?
  - what will you use as the transition probabilities associated with the edges?

# Modeling CpG Islands as a Markov Chain Answer

- The nodes (states) of our Markov chain state graph are just the four nucleotides A, C, G, T.
- Every node has four outgoing edges, to itself and the other three nucleotides.
- The transition probabilities are just the conditional probabilities $p(X_{t+1}|X_t)$ given by the table.

# CpG Island State Graph




- Note the *self-edges*; this is common in MC state graphs.
- (Yes, *dot* chose a really silly layout.)

# Deriving the Transition Matrix

Given a table of the 16 dinucleotide frequencies $p(X_t, X_{t+1})$ measured for CpG islands, indicate how you would derive the transition matrix for your Markov chain model.

# Deriving the Transition Matrix Answer

- We can easily turn dinucleotide frequencies into the conditional probabilities we need for a Markov chain:

$$p(X_{t+1} = G|X_t = C) = \frac{p(X_{t+1} = G, X_t = C)}{\sum_{X_{t+1}} p(X_{t+1}, X_t = C)}$$

- These become the *transition probabilities* $\tau_{ij}$ of the Markov chain.

# Uniform Probabilities?

Say we wish our model of the Occasionally Dishonest Casino to have the same hidden state probability $p(X_t = F)$ at all times $t$. Assuming the transition probabilities are $\tau_{FL} = p(L|F) = 0.05$ and $\tau_{LF} = p(F|L) = 0.1$, how can we achieve this goal? (If this is not possible with the given data, just say "Insufficient data").

# Uniform Probabilities? Answer

- If we set $p(X_1 = s_i) = \pi_i$ to the *stationary distribution*, then the distribution at all times will just be $\vec{\pi}$.
- For this two-state case it's easy to get $\vec{\pi}$ straight from the transition probabilities:

$$\pi_F = \frac{\tau_{LF}}{\tau_{LF} + \tau_{FL}} = \frac{2}{3}$$

# Viterbi Algorithm

- Finds an optimal path (a sequence of hidden states) that maximizes the joint probability of the observable and hidden variables $p(\vec{X}, \vec{\theta})$.

$$V_{ti} = \text{Max}\left[p(X_t = s_i | X_{t-1} = s_j) V_{t-1,j}\right]$$

- Recursive algorithm, sometimes called “dynamic programming”
- Computationally efficient: finds the best path out of $O(m^n)$ possible paths but only takes $O(nm^2)$ time.

# Viterbi subpaths?

Say $\vec{V}^n$ is the Viterbi optimal path of states to a destination state $\Theta_n = s_i$ in a homogeneous HMM. In other words, there is no other path to $\Theta_n = s_i$ with greater probability. Now choose any point $\Theta_t = s_j$ on path $\vec{V}^n$ where $t < n$. What is the probability that the subsegment of path $\vec{V}^n$ ending at $\Theta_t = s_j$ is itself a Viterbi maximum? (In other words, there is no other path to $\Theta_t = s_j$ with greater probability.)

1. 1.
2. approximately $|S|^{-1}$, where $|S|$ is the number of distinct states for each variable in the homogeneous HMM.
3. 0.
4. The probability depends on the details of the HMM's probability model.

# Viterbi subpaths? Answer

The correct answer is 1. The Viterbi algorithm is based on a proof by induction: namely that we can find an optimal path to $\Theta_t = s_j$ assuming that already know optimal paths to $\Theta_{t-1} = s_i$. Thus by definition *every* subpath of a Viterbi optimal path must itself be a Viterbi optimal path.

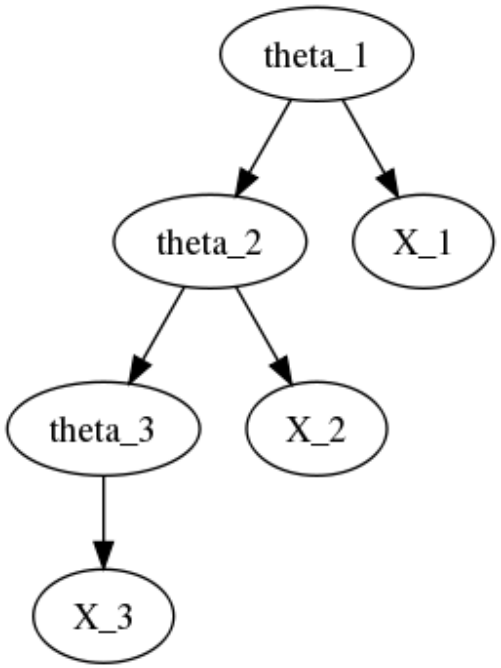
# Predicting CpG Islands Using an HMM

You wish to use your Markov chain model of CpG islands to predict which sequence regions in the human genome are actually CpG islands. You are given dinucleotide frequencies for non-CpG island sequences, and the probabilities of transitioning from CpG island to non-Cpg island at any nucleotide, and vice versa.

- describe the state graph of your HMM (you do not need to *draw* all its details; just state its structure unambiguously).
- draw an information graph of your HMM, including both the hidden and observable variables associated with three consecutive nucleotides in the genome sequence.

# Predicting CpG Islands Using an HMM Answer

- The state graph consists of four nodes for the CpG model (call them A+, etc.), plus four nodes for the non-CpG model (A-, etc.).
- Any node can transition to any node, including itself.
- The information graph includes a hidden state $\theta_t$ for each nucleotide, which emits a single observation $X_t$ (the genomic sequence).



# Emission Probabilities

- As the information graph highlights, there is a conditional probability term $p(X_t|\theta_t)$ for each nucleotide in the genome sequence. These are called the *emission probabilities*.
- Naturally, each hidden state can only emit one possible observed letter (i.e. A+ $\rightarrow$ A).
- Does this mean $\theta_t$ is actually observable?
- No. For each observed letter (e.g. A), there are *two* hidden states (e.g. A+, A-) that could have emitted it.

# CpG Island HMM Transition Matrix

Indicate what if anything you need to add / alter in the transition matrix from your original CpG island Markov chain.

# CpG Island HMM Transition Matrix Answer

- For transitions within the CpG model, we simply multiply the original $\tau_{ij}$ by $p(+|+)$, the probability of remaining in a CpG island at any given nucleotide position.
- For transitions within the non-CpG model, we multiply by $p(-|-)$, the probability of remaining in the non-CpG model.
- For transitions from CpG model to non-CpG model, we can just multiply $p(-|+)$ by $p(j|-)$, the probability of observing nucleotide $j$ at any given position in the non-CpG model. This is just given by the stationary distribution $\pi_j$ for the non-CpG model.
- We treat transitions from non-CpG model to CpG model in the same way.

# CpG Island Coupling

What ties the hidden state to the observable in your CpG island HMM? In other words, what information enables the HMM to predict the location of CpG islands in the genomic sequence?

# CpG Island Coupling Answer

The coupling of observable to hidden state is captured by two things:

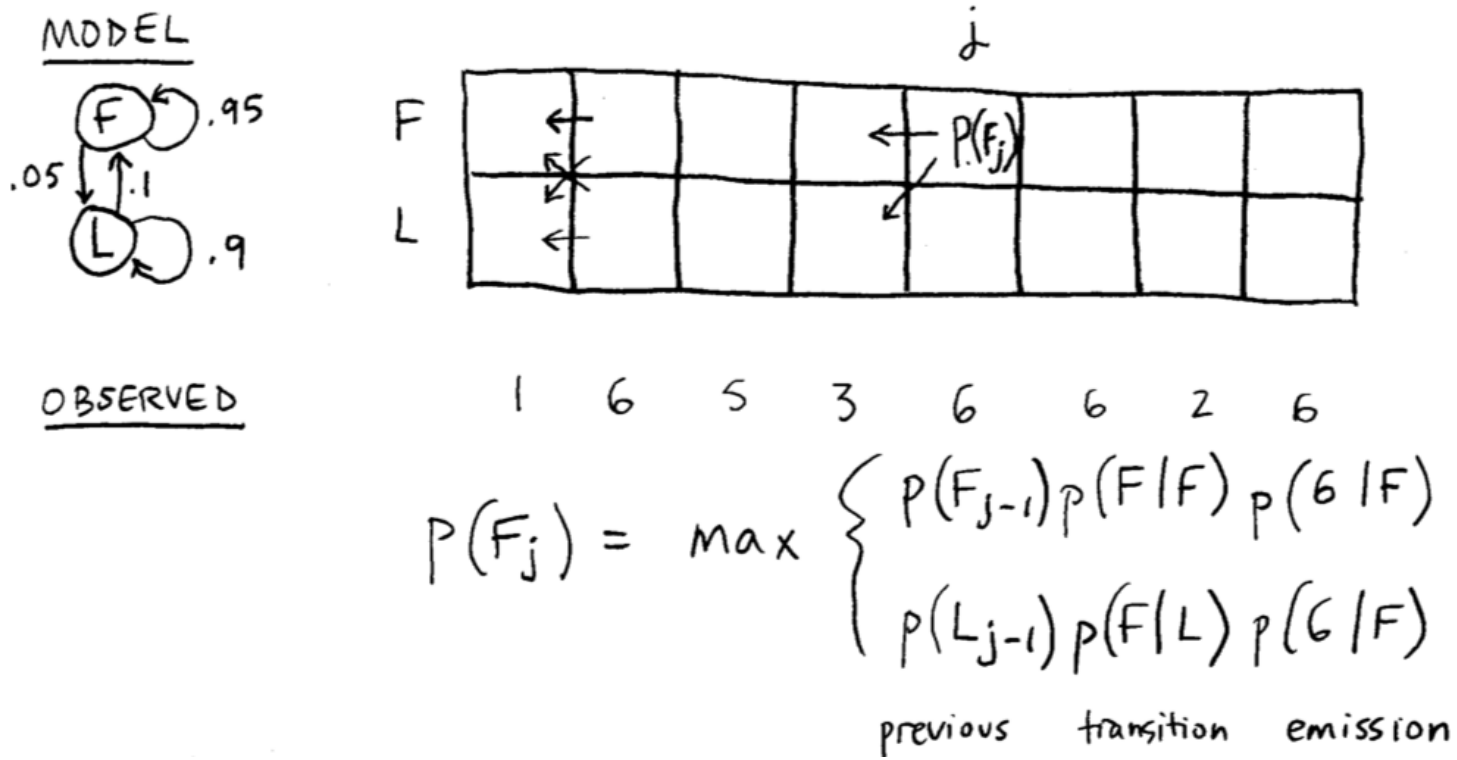
- by the emission probabilities (in this case, trivially);
- by the fact that the dinucleotide frequency tables for CpG vs. non-CpG models are so different.

# Occasionally Dishonest Casino Viterbi

To make the Viterbi algorithm concrete, let's draw the calculation:

- draw a rectangular matrix whose rows are the possible hidden states $s_i$, and whose columns are the successive time steps $t$.
- For one time step $t$, draw the possible transitions at that time step as arrows on the matrix.
- Write the Viterbi condition for choosing the optimal path probabiity to state $\theta_t = F$

# Occasionally Dishonest Casino Viterbi Answer




- Shorthand notation: Interpret $j$ as our time index.
- Interpret $p(F_j) \equiv p(\theta_1^*, \theta_2^*, ..., \theta_j^* = F, \vec{O}^j)$