# Identifying CpG islands using hidden Markov models

Matthew Macauley

Department of Mathematical Sciences
Clemson University
http://www.math.clemson.edu/~macaule/

Math 4500, Fall 2016

# CpG islands

On a DNA strand, a cytosine followed by guanine is a dinucleotide called *CpG*. The 'p' is for the *phosphate bond* between them.


Figure: *CpG* nucleotides on a DNA strand and its complement.

CpG's are often clustered in regions called CpG islands (CGIs).

CGIs are often associated with the promoter region of genes (where transcription begins).

Identifying CGIs can help identify new genes, some of which may be involved in cancer.

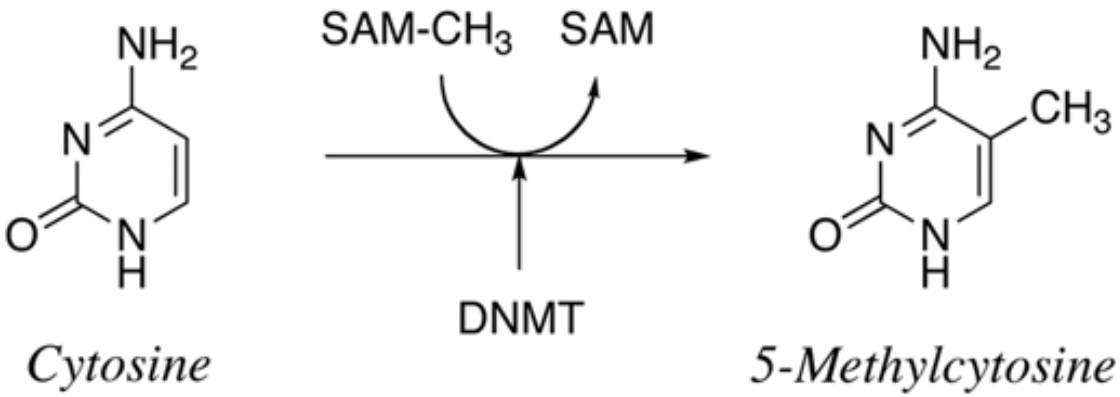
**Goal**

Given a genome of millions of base pairs, how can one identify the CpG islands?

# Cytosine methylation

Almost all cells in an organism have the same DNA sequence. The difference lies in the levels of *gene expression*.

One common way that genes are turned off is by a chemical change called methyalation at the promoter CGI.




Promoter regions of housekeeping genes are usually unmethylated.

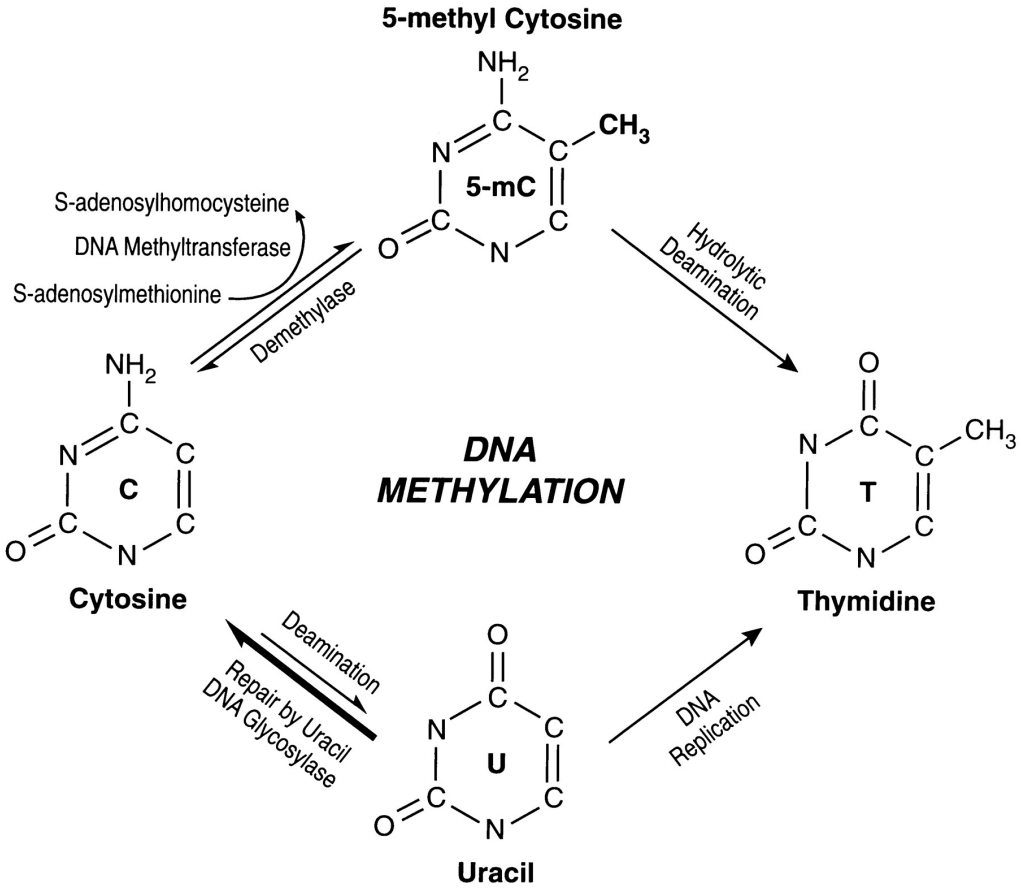
Appropriate methylation of CGIs is needed for normal development. If methylation occurs when it should not in tumor suppressor genes, then problems such as cancer can result.

In mammals, 70–80% of CpG cytosines are methylated, but it depends on the type of cell.

For example, hemoglobin genes should be methylated (and shut off) in skin cells but unmethylated (and expressed) in red blood cell precursors.

# Methylation and deamination

5-methyl cytosine can be deaminated to produce thymine (T), which is a mutation. As a result, there is a lack of CpG sites in methylated DNA.




## Rule of thumb

On an evolutionary timescale, unmethylated C's tend to persist and methylated C's tend to be eliminated.


**Left**: CpG sites at 1/10 nucleotides, constituting a CpG island. The sample is of a gene-promoter, the highlighted ATG consitutes the start codon.

**Right**: CpG sites present at every 1/100 nucleotides, consituting a more normal example of the genome, or a region of the genome that is commonly methylated.

# How to define a CpG island

The human genome has a 42% GC content. Thus, the expected frequency of a CpG $0.21 \cdot 0.21 = 4.41\%$. However, the actual frequency is 1%.

The percent combined $C + G$ content ($\%C + G$) is defined "exactly how you would expect."

If dinucleotides were formed by randomly choosing two nucleotides, then the expected number of *CpG*'s would be

$$\frac{(\#\text{ C's}) \cdot (\#\text{ G's})}{\text{length of sequence}}$$

The observed over expected CpG ratio (O/E CpG) is: $\dfrac{\text{observed } \#\text{ CpG's}}{\text{expected } \#\text{ CpG's}}$.

## Definition (Gardiner-Garden, Frommer, 1987)

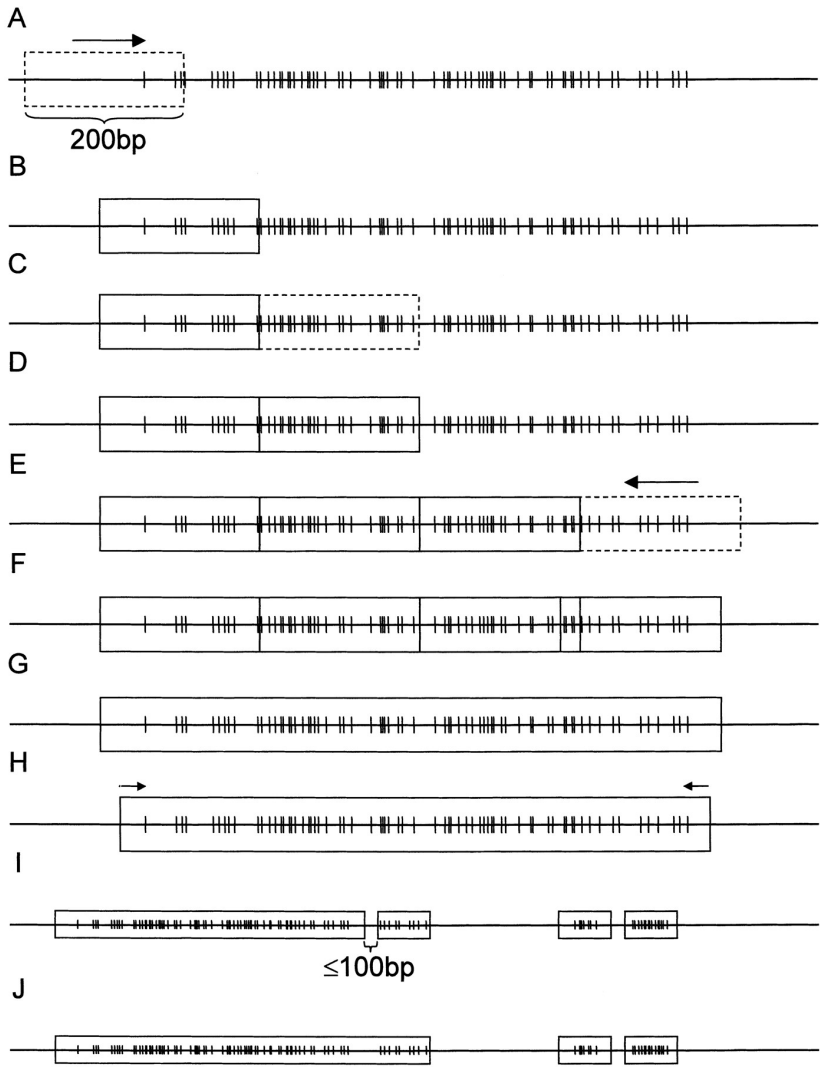
A subsequence in a vertebrate genome is CpG island if:

1. it has length at least 200 bp;
2. $\%C + G \geqslant 50\%$;
3. O/E CpG $\geqslant 0.6$;

There is no universal standard for these values. Another paper (Takai & Jones) used 500 bp, $\%C + G \geqslant 55\%$, and O/E CpG $\geqslant 0.65$.

# Finding CpG islands

One method for inferring CpG islands is purely algorithmic: using a sliding window.



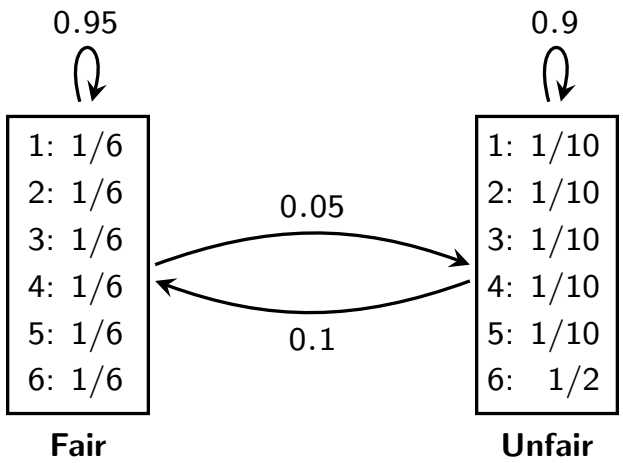

The remainder of this lecture will focus on an alternative approach: hidden Markov models.

# The occasionally dishonest casino

Suppose a casino hosts a simple game with two dice: one fair and one unfair.

- FAIR: $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = 1/6$.
- UNFAIR: $p(1) = p(2) = p(3) = p(4) = p(5) = 1/10$, $p(6) = 1/2$.

The casino switches between fair and unfair die according to the following probabilities:




*You cannot tell which die the casino is using.* This is a hidden Markov model (HMM).

Suppose that the outcome of the game is the following:

- WIN: roll 1, 2, 3, or 4.
- LOSE: roll 5 or 6.

Would you play this game?

# The occasionally dishonest casino

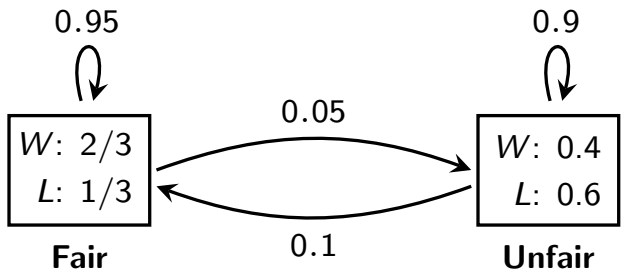
## 3 canonical questions

Given a sequence of roles by the casino:

123626362511512666122161452152616666161166126162664366626223451612426

one may ask:

1. Decoding: When was the casino rolling the fair vs. the unfair die?
2. Evaluation: How likely is this sequence given our model?
3. Learning: Can we deduce the probability parameters if we didn't know them? (e.g., "*how loaded are the die?*", and "*how often does the casino switch?*")




We'll analyze these questions but for simplicity, only record wins vs. losses:

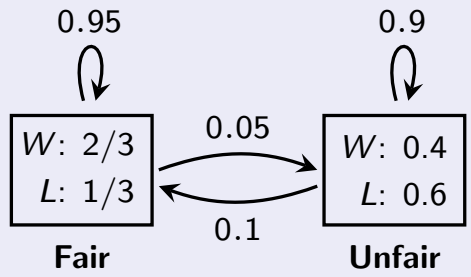
WWWLWLWLWLWWLWWLLLWWWWLWWLWWLWLWLLLWLWWLLWWLWLWLLWWLLLWLWWWWLWLWWWWL

# Two examples of Hidden Markov models

The parameters of an HMM can be encoded in a table.

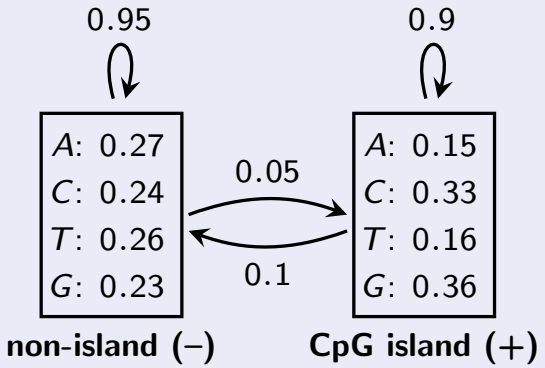
## HMM for the occasionally dishonest casino

| State | Transitions | | Emissions | | Initial distribution |
|---|---|---|---|---|---|
| | **F** | **U** | **W** | **L** | |
| **F** | .95 | .05 | 2/3 | 1/3 | .5 |
| **U** | .1 | .9 | .4 | .6 | .5 |



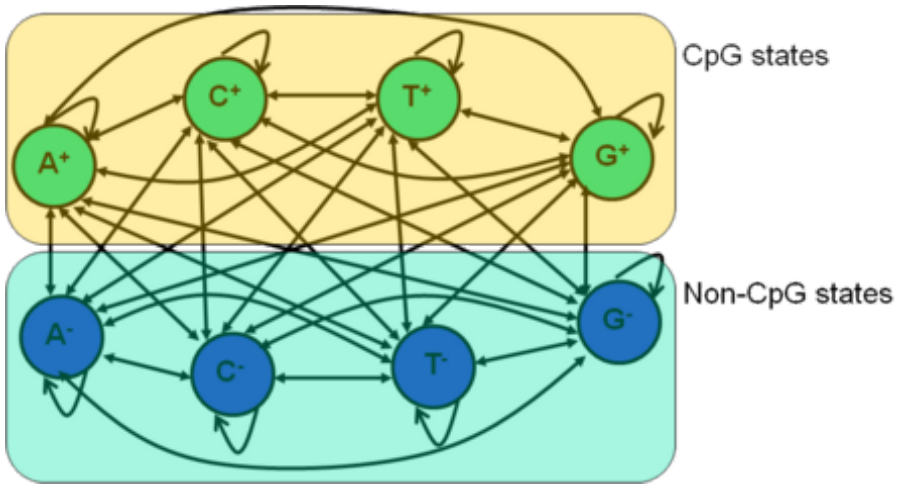

## HMM for CpG islands (simple)

| State | Transitions | | Emissions | | | | Init. dist. |
|---|---|---|---|---|---|---|---|
| | − | + | **A** | **C** | **T** | **G** | |
| − | .95 | .05 | .27 | .24 | .26 | .23 | .5 |
| + | .1 | .9 | .15 | .33 | .16 | .36 | .5 |



# A better hidden Markov model for CpG islands

A "better" HMM model should incorporate the fact that transmission probabilities within CpG islands are much different than the rest of the genome.




The following is from a sequence of annotated human DNA of length $\approx$60,000.

| | **Transitions** | | | | | | | | **Emissions** | | | | **Init.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $A_-$ | $C_-$ | $T_-$ | $G_-$ | $A_+$ | $C_+$ | $T_+$ | $G_+$ | A | C | T | G | |
| $A_-$ | .300 | .205 | .210 | .285 | $(1-q)/4$ | $(1-q)/4$ | $(1-q)/4$ | $(1-q)/4$ | 1 | 0 | 0 | 0 | .125 |
| $C_-$ | .322 | .298 | .302 | .078 | $(1-q)/4$ | $(1-q)/4$ | $(1-q)/4$ | $(1-q)/4$ | 0 | 1 | 0 | 0 | .125 |
| $T_-$ | .248 | .246 | .208 | .298 | $(1-q)/4$ | $(1-q)/4$ | $(1-q)/4$ | $(1-q)/4$ | 0 | 0 | 1 | 0 | .125 |
| $G_-$ | .177 | .239 | .292 | .292 | $(1-q)/4$ | $(1-q)/4$ | $(1-q)/4$ | $(1-q)/4$ | 0 | 0 | 0 | 1 | .125 |
| $A_+$ | $(1-p)/4$ | $(1-p)/4$ | $(1-p)/4$ | $(1-p)/4$ | .180 | .274 | .120 | .426 | 1 | 0 | 0 | 0 | .125 |
| $C_+$ | $(1-p)/4$ | $(1-p)/4$ | $(1-p)/4$ | $(1-p)/4$ | .171 | .368 | .188 | .274 | 0 | 1 | 0 | 0 | .125 |
| $T_+$ | $(1-p)/4$ | $(1-p)/4$ | $(1-p)/4$ | $(1-p)/4$ | .161 | .339 | .125 | .375 | 0 | 0 | 1 | 0 | .125 |
| $G_+$ | $(1-p)/4$ | $(1-p)/4$ | $(1-p)/4$ | $(1-p)/4$ | .079 | .355 | .182 | .384 | 0 | 0 | 0 | 1 | .125 |

# Three canonical HMM problems, formalized

## Problem #1: Decoding

Given an observed path $x = x_1 x_2 x_3 \cdots x_\ell$, what is the most likely hidden path $\pi = \pi_1 \pi_2 \pi_3 \cdots \pi_\ell$ to emit $x$? That is, compute

$$\pi_{\max} = \arg\max_{\pi} P(\pi|x) = \arg\max_{\pi} P(x, \pi)$$

## Problem #2: Evaluation

Given an observed path $x = x_1 x_2 x_3 \cdots x_\ell$, what is its probability $P(x)$? That is, compute

$$P(x) = \sum_{\pi} P(x, \pi), \qquad \text{where} \quad P(x, \pi) = a_{0\pi_1} \prod_{i=1}^{\ell} e_{\pi_i}(x_i) a_{\pi_i, \pi_{i+1}}$$

and the sum is over all hidden sequences $\pi = \pi_1 \pi_2 \cdots \pi_\ell$.

## Problem #3: Learning

Given an observed sequence $x$ (or set of sequences), what are the HMM parameters that make $x$ mostly likely to occur?