Example: HMMs for CpG island finding

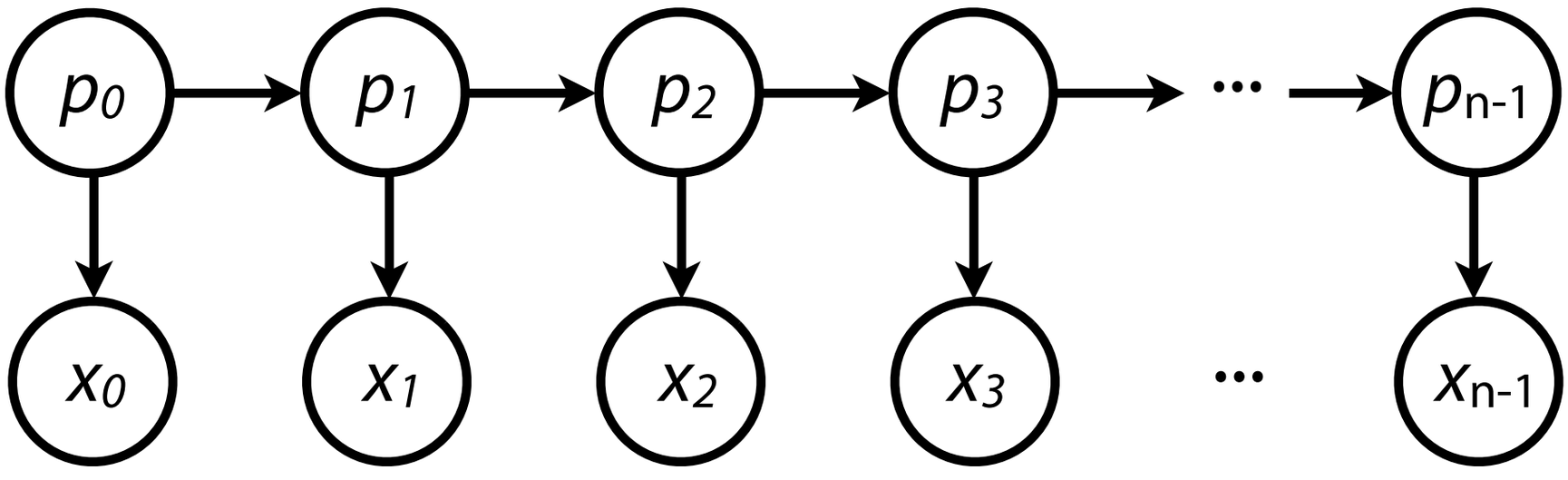
Ben Langmead

For original Keynote files, email me (ben.langmead@gmail.com)

Hidden Markov Model

We know what it is, how to calculate a joint probability, and how to find the most likely path given an emission sequence (Viterbi)

Can we design an HMM for finding **CpG islands**?

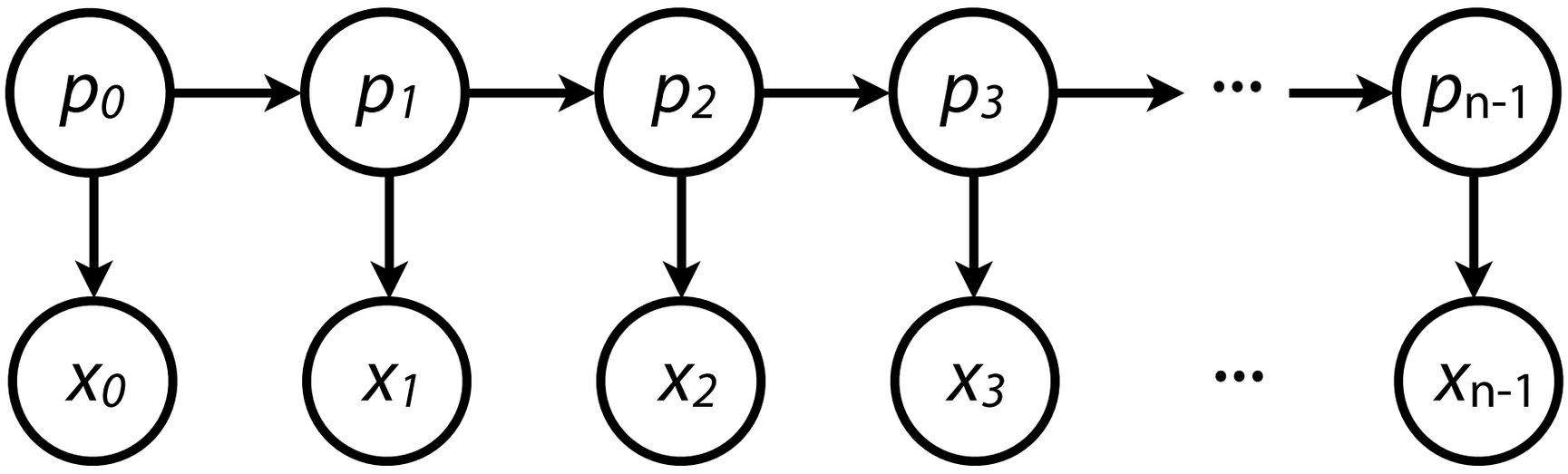


Hidden Markov Model

Idea 1:

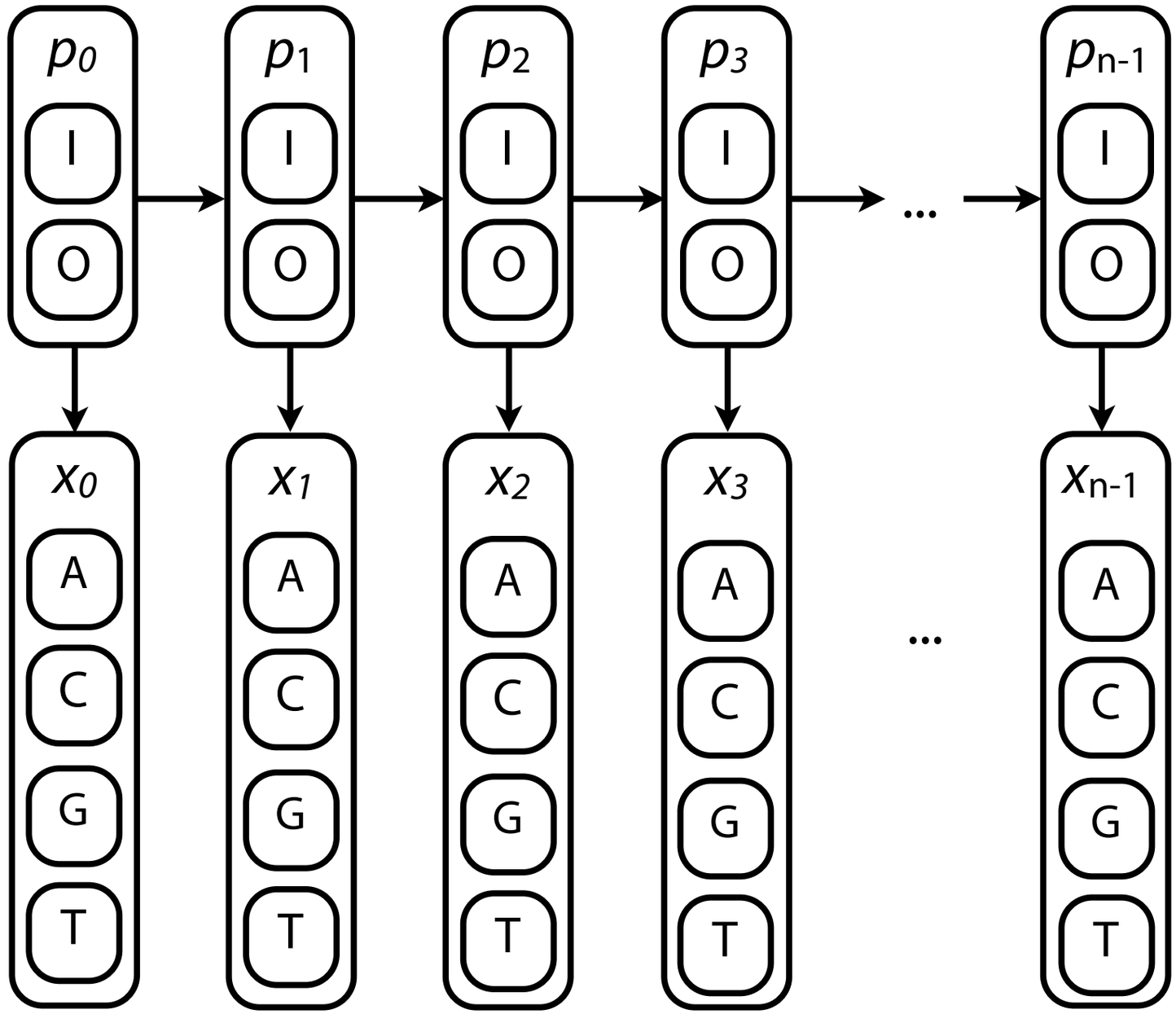
States (Q) = { }

Emissions (∑) = { }



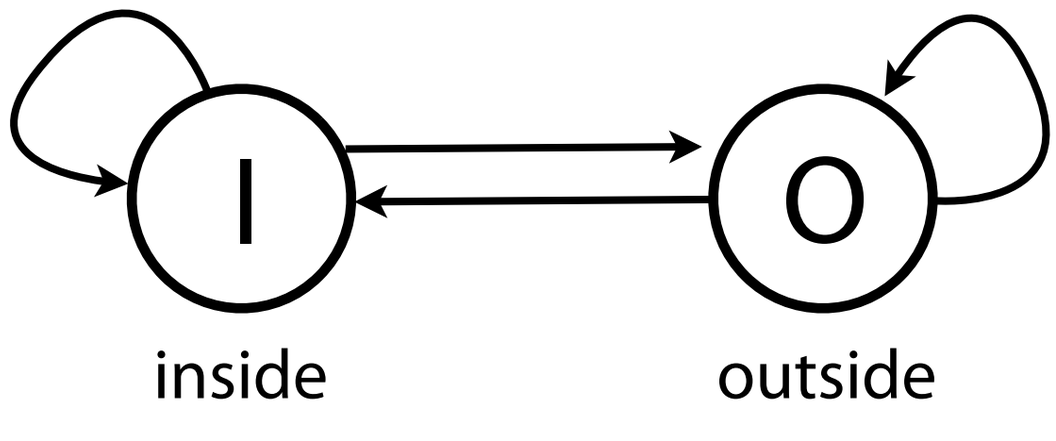
Hidden Markov Model

Idea 1: Q = { inside, outside }, ∑ = { A, C, G, T }



Hidden Markov Model

Idea 1: $Q = \{$ inside, outside $\}$, $\Sigma = \{$ A, C, G, T $\}$



A	I	O
I		
O		

Transition matrix

E	A	C	G	T
I				
O				

Emission matrix

Hidden Markov Model

Idea 1: Q = { inside, outside }, ∑ = { A, C, G, T }

Training data consists of bases and associated I/O annotations:

```
ATCGTCTCGATATAGCGCGGTACGCGATTAACGATATATATCGGCGTGCCGTACG···
OOOOOOOOOOOOOOIIIIIIIIIIIOOOOOOOOOOOOOOIIIIIIIIIIIIII···
```

A	I	O
I		
O		

Transition matrix

E	A	C	G	T
I				
O				

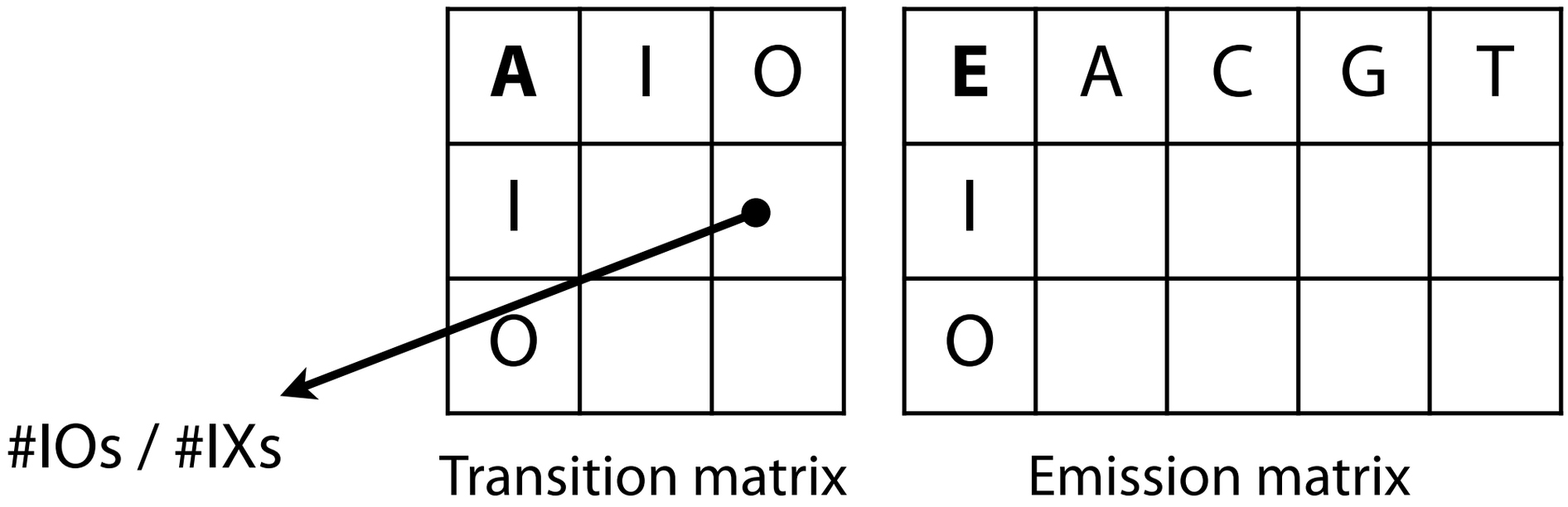
Emission matrix

Hidden Markov Model

Idea 1: Q = { inside, outside }, ∑ = { A, C, G, T }

Training data consists of bases and associated I/O annotations:

```
ATCGTCTCGATATAGCGCGGTACGCGATTAACGATATATATCGGCGTGCCGTACG···
OOOOOOOOOOOOOOIIIIIIIIIIIIOOOOOOOOOOOOOOOIIIIIIIIIIIIII···
```

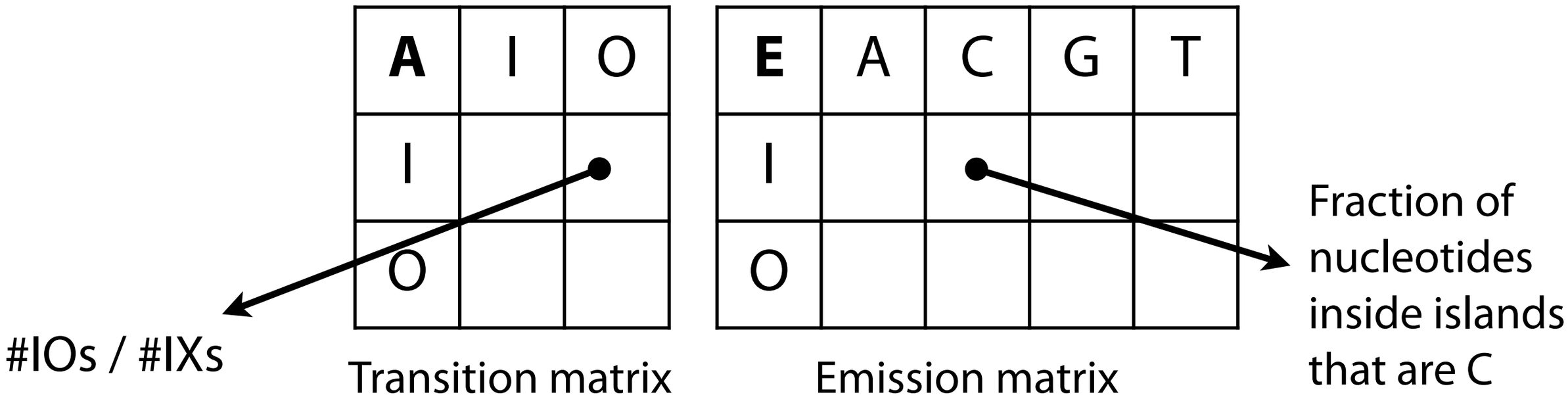


Hidden Markov Model

Idea 1: Q = { inside, outside }, ∑ = { A, C, G, T }

Training data consists of bases and associated I/O annotations:

```
ATCGTCTCGATATAGCGCGGTACGCGATTAACGATATATATCGGCGTGCCGTACG···
OOOOOOOOOOOOOOIIIIIIIIIIIOOOOOOOOOOOOOOOIIIIIIIIIIIIII···
```



Hidden Markov Model

Example 1 using HMM idea 1:

A	I	O
I	0.8	0.2
O	0.2	0.8

E	A	C	G	T
I	0.1	0.4	0.4	0.1
O	0.25	0.25	0.25	0.25

I
0.5
0.5

x: ATATATACGCGCGCGCGCGATATATATATATATA

p: OOOOOOOIIIIIIIIIIIIIOOOOOOOOOOOOO

(from Viterbi)

Hidden Markov Model

Example 1 using HMM idea 1:

A	I	O
I	0.8	0.2
O	0.2	0.8

E	A	C	G	T
I	0.1	0.4	0.4	0.1
O	0.25	0.25	0.25	0.25

I
0.5
0.5

x: ATATCGCGCGCGATATATATCGCGCGCGATATATATAT

Hidden Markov Model

Example 1 using HMM idea 1:

A	I	O
I	0.8	0.2
O	0.2	0.8

E	A	C	G	T
I	0.1	0.4	0.4	0.1
O	0.25	0.25	0.25	0.25

I
0.5
0.5

x: ATATCGCGCGCGATATATCGCGCGCGATATATAT

p: OOOOIIIIIIIIOOOOOOIIIIIIIIOOOOOOOO

Hidden Markov Model

Example 1 using HMM idea 1:

A	I	O
I	0.8	0.2
O	0.2	0.8

E	A	C	G	T
I	0.1	0.4	0.4	0.1
O	0.25	0.25	0.25	0.25

I
0.5
0.5

x: ATATATACCCCCCCCCCCCCCCATATATATATATATA

Hidden Markov Model

Example 1 using HMM idea 1:

A	I	O
I	0.8	0.2
O	0.2	0.8

E	A	C	G	T
I	0.1	0.4	0.4	0.1
O	0.25	0.25	0.25	0.25

I
0.5
0.5

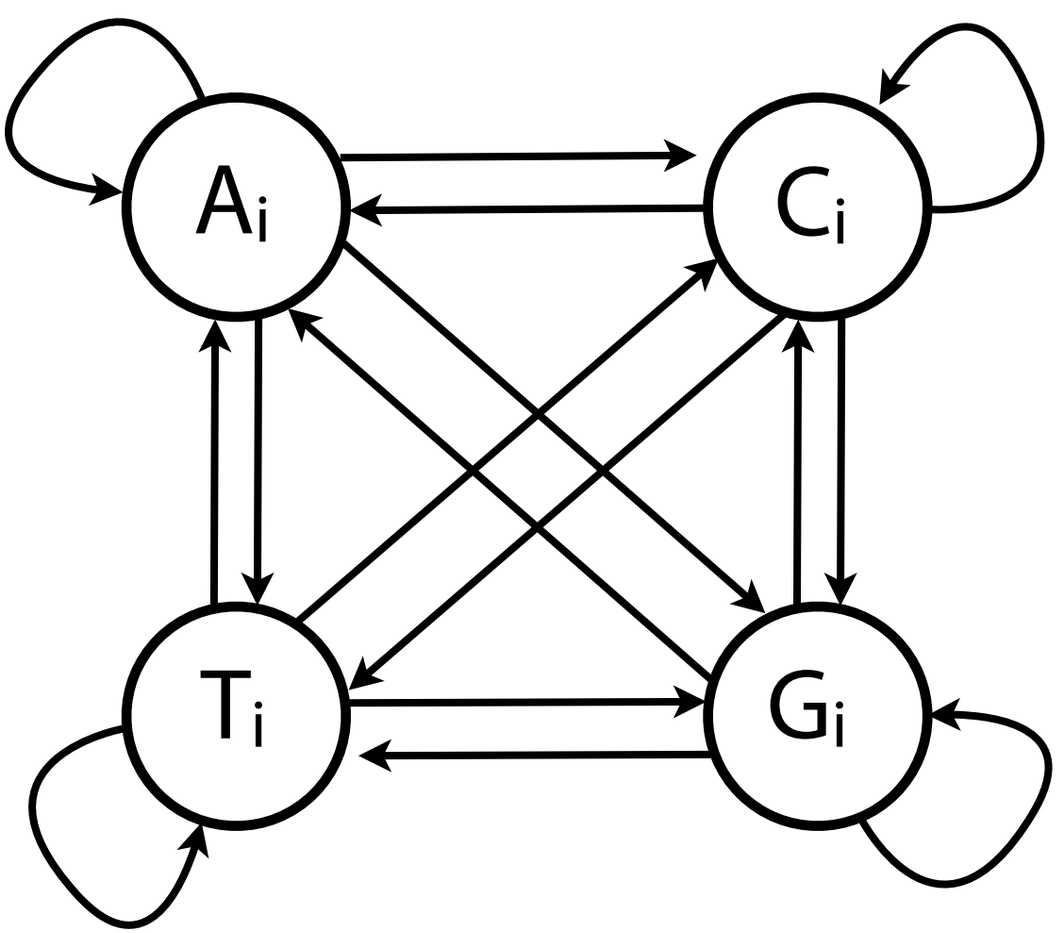
x: ATATATACCCCCCCCCCCCCCATATATATATATATA

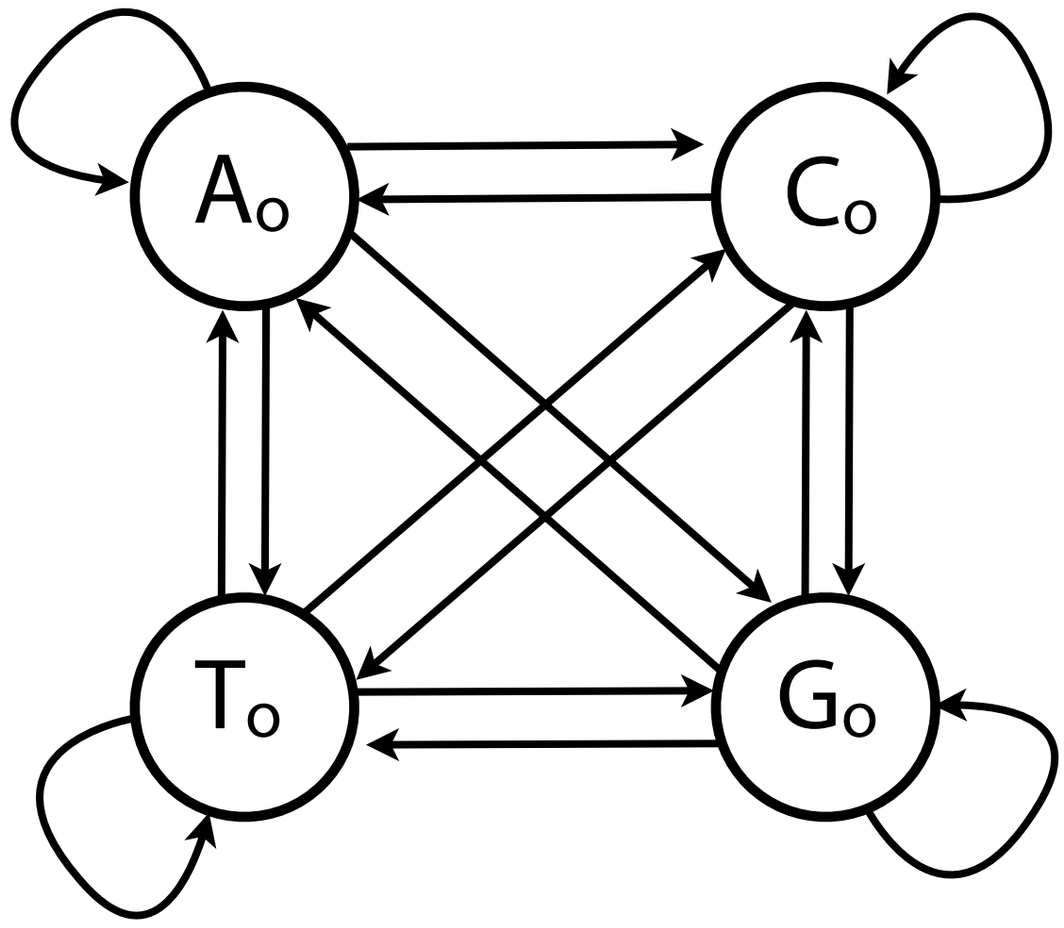
p: OOOOOOOIIIIIIIIIIIIIOOOOOOOOOOOOO

😬 - clearly not a CpG island

Hidden Markov Model

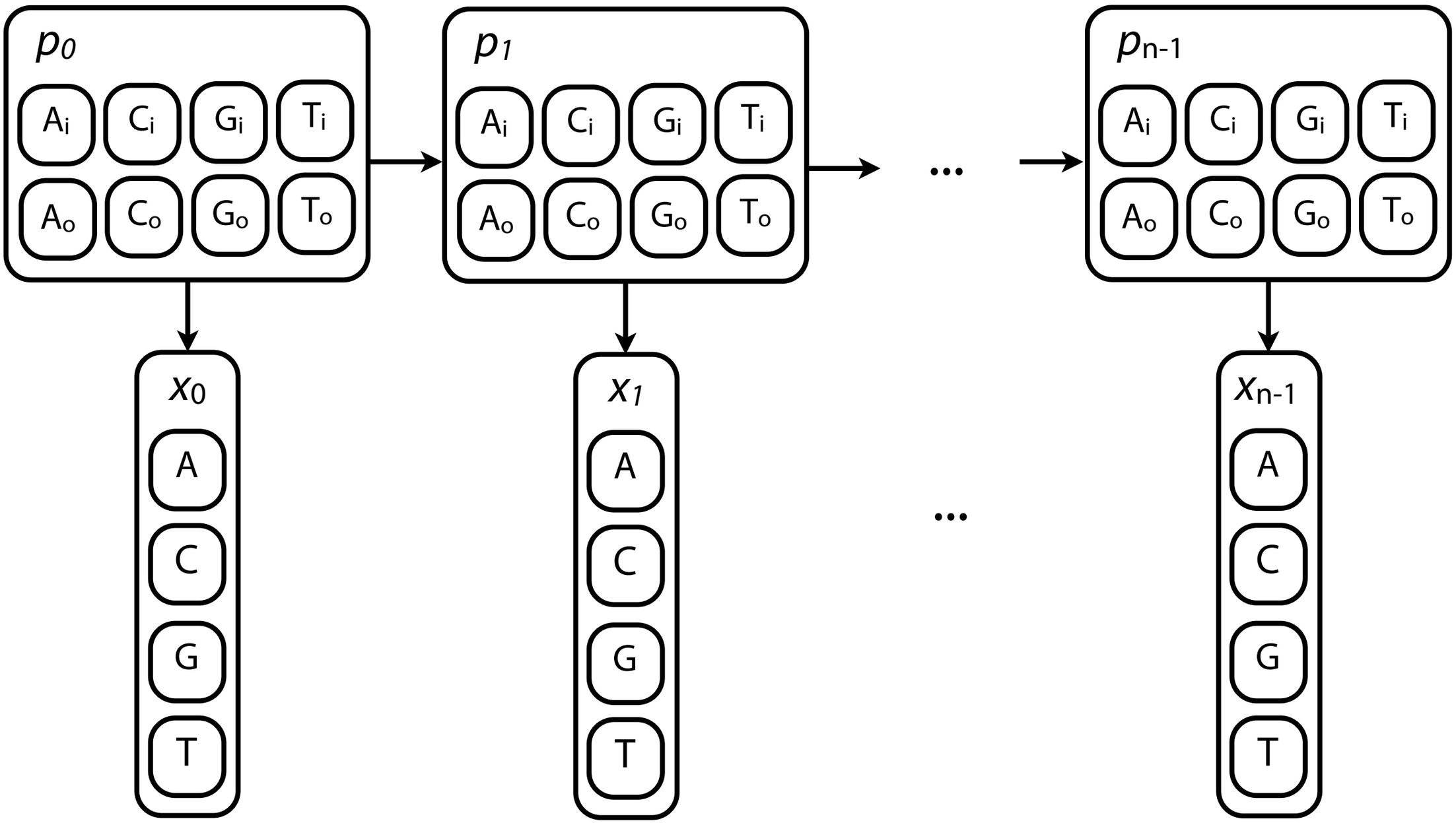
Recall our 2-chain model from before





Hidden Markov Model

Idea 2: $Q = \{ A_i, C_i, G_i, T_i, A_o, C_o, G_o, T_o \}$, $\Sigma = \{ A, C, G, T \}$



Hidden Markov Model

Idea 2: $Q = \{ A_i, C_i, G_i, T_i, A_o, C_o, G_o, T_o \}$, $\Sigma = \{ A, C, G, T \}$

A	A_i	C_i	G_i	T_i	A_o	C_o	G_o	T_o
A_i								
C_i								
G_i								
T_i								
A_o								
C_o								
G_o								
T_o								

Transition matrix

E	A	C	G	T
A_i				
C_i				
G_i				
T_i				
A_o				
C_o				
G_o				
T_o				

Emission matrix

Hidden Markov Model

ATCGTCTCGATATAGCGCGGTACGCGATTAACGATATATATCGGCGTGCCGTACG
OOOOOOOOOOOOOOIIIIIIIIIIIOOOOOOOOOOOOOOIIIIIIIIIIIII

A	A_i	C_i	G_i	T_i	A_o	C_o	G_o	T_o
A_i								
C_i								
G_i								
T_i								
A_o								
C_o								
G_o								
T_o								

Transition matrix

E	A	C	G	T
A_i				
C_i				
G_i				
T_i				
A_o				
C_o				
G_o				
T_o				

Emission matrix

Hidden Markov Model

ATCGTCTCGATATAGCGCGGTACGCGATTAACGATATATATCGGCGTGCCGTACG
OOOOOOOOOOOOOOIIIIIIIIIIIOOOOOOOOOOOOOOIIIIIIIIIIIII

A	A_i	C_i	G_i	T_i	A_o	C_o	G_o	T_o
A_i								
C_i								
G_i								
T_i								
A_o								
C_o								
G_o								
T_o								

Estimate $P(C_i \mid T_i)$ as
T_i C_i s divided by # T_i s

Transition matrix

E	A	C	G	T
A_i				
C_i				
G_i				
T_i				
A_o				
C_o				
G_o				
T_o				

Emission matrix

Hidden Markov Model

ATCGTCTCGATATAGCGCGGTACGCGATTAACGATATATATCGGCGTGCCGTACG
OOOOOOOOOOOOOOIIIIIIIIIIIOOOOOOOOOOOOOOIIIIIIIIIIIII

A	A_i	C_i	G_i	T_i	A_o	C_o	G_o	T_o
A_i								
C_i								
G_i								
T_i								
A_o								
C_o								
G_o								
T_o								

Estimate $P(C_i \mid T_i)$ as
T_i C_i s divided by # T_i s

Transition matrix

E	A	C	G	T
A_i	1	0	0	0
C_i	0	1	0	0
G_i	0	0	1	0
T_i	0	0	0	1
A_o	1	0	0	0
C_o	0	1	0	0
G_o	0	0	1	0
T_o	0	0	0	1

Emission matrix

Hidden Markov Model

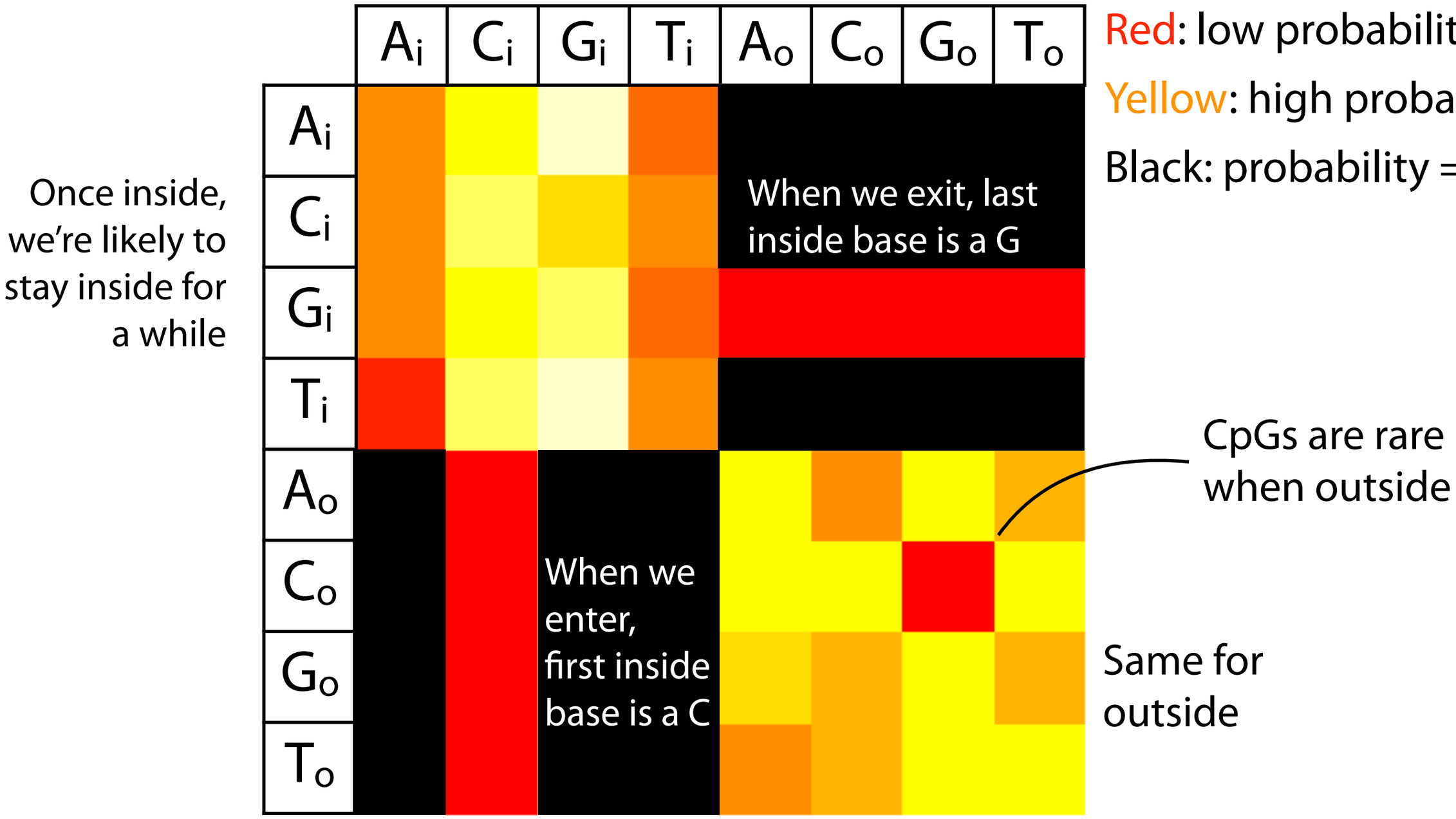
Trained transition matrix:

```
          A        C        G        T        a        c        g        t
A: 0.185441,0.276405,0.400914,0.137241,0.000000,0.000000,0.000000,0.000000
C: 0.189582,0.359051,0.253240,0.198127,0.000000,0.000000,0.000000,0.000000
G: 0.171904,0.328635,0.354250,0.139893,0.000785,0.001479,0.001857,0.001198
T: 0.094102,0.341636,0.376867,0.187395,0.000000,0.000000,0.000000,0.000000
a: 0.000000,0.000038,0.000000,0.000000,0.294811,0.194641,0.286965,0.223544
c: 0.000000,0.000076,0.000000,0.000000,0.326810,0.294085,0.061724,0.317305
g: 0.000000,0.000057,0.000000,0.000000,0.257128,0.233486,0.294244,0.215085
t: 0.000000,0.000031,0.000000,0.000000,0.179562,0.232469,0.294630,0.293308
```

Uppercase = inside, lowercase = outside

Hidden Markov Model

Trained transition matrix:



Hidden Markov Model

Viterbi result; lowercase = *outside*, uppercase = *inside*:

```
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCG
CGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCG
CGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
```

Hidden Markov Model

Viterbi result; lowercase = *outside*, uppercase = *inside*:

```
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatgggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggg
gggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggg
ggggggggggggggggggggggggggggggggggggggggggggggggggggggatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatat
```