Exercises

1) Give four examples of application (uses) of HMMs in computational biology. Gene finding, conserved/regulatory/chromatin regions, CNV, homologs

2) What is the probability of x = AATTCG under the CpG island Markov chain and under the non-CpG island Markov chain (described in the slides)? next slide

3) How do you avoid overflow – errors caused by operations on really small numbers? add log values

4) What are two disadvantages of using a sliding window with a cutoff to identify CpG islands? cutoff, window size

5) In HMMs, the labels (states) are hidden/observed, and the emissions are hidden/observed?

CpG island

	A	G	C	T
A	0.19	0.27	0.40	0.14
G	0.17	0.33	0.36	0.14
C	0.19	0.36	0.25	0.20
T	0.10	0.34	0.38	0.19

x = AATTCG

P(x) = 0.36*0.38*0.19*0.14*0.19*0.16

P(x) = 1.10622e-04

Non-CpG island

	A	G	C	T
A	0.34	0.23	0.18	0.25
G	0.30	0.25	0.20	0.25
C	0.38	0.04	0.26	0.33
T	0.22	0.26	0.21	0.31

x = AATTCG

P(x) = 0.04*0.21*0.31*0.25*0.34*0.31

P(x) = 6.86e-05

6) What is the probability of AACG with hidden states OOII under the following HMM:

A	I	O
I	0.8	0.2
O	0.2	0.8

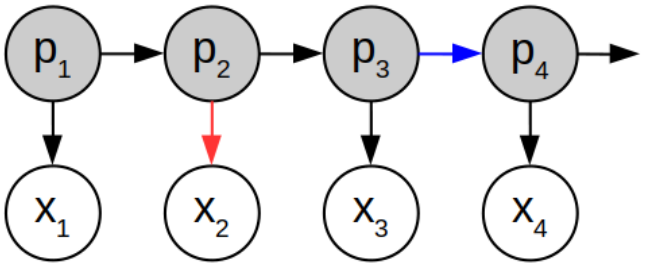
E	A	G	C	T
I	0.1	0.4	0.4	0.1
O	0.25	0.25	0.25	0.25

AACG
OOII

P = initial *
0.8*0.2*0.8 (transitions)
0.25*0.25*0.4*0.4 (emissions)
If initial = 0.5, P = 0.00064

7) In the diagram below, we observed x1-x4 but not p1-p4:

a) does P(p3) depend on p2? yes
b) does P(p3) depend on x3? yes
c) does P(p3) depend on x2? yes
d) does P(p3) depend on x4? yes
e) does P(p3|p2) depend on x2? no



8) Fill in the last column using viterbi and A and E from prior slides.

9) Whats the most likely path?

A	F	L
F	0.6	0.4
L	0.4	0.6

E	H	T
F	0.5	0.5
L	0.8	0.2

FFFL

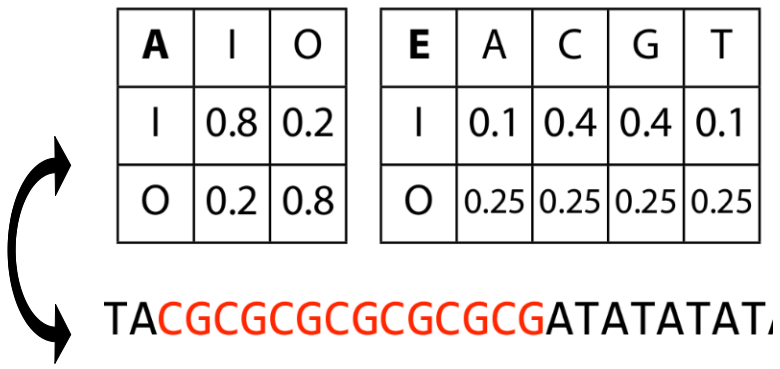
x	T	H	T	H
$S_{F,i}$	E(T|F) A(F) $S_{F,1}$=.25	E(H|F)=.5 $S_{F,1}$ A(F|F) $S_{L,1}$A(F|L) $S_{F,2}$=.075	E(T|F)=.5 $S_{F,2}$ A(F|F) $S_{L,2}$A(F|L) $S_{F,3}$=.0225	$S_{F,4}$ = E x max{$S_{k,3}$ x A} E(H|F) = 0.5 $S_{F,3}$ x A(F|F) = .0225*.6 max $S_{L,3}$ x A(F|L) = .0096*.4 $S_{F,4}$ = .00675
$S_{L,i}$	E(T|L) A(L) $S_{L,1}$=.1	E(H|L) = .8 $S_{F,1}$ A(L|F) $S_{L,1}$A(L|L) $S_{L,2}$=.08	E(T|L) = .2 $S_{F,2}$ A(L|F) $S_{L,2}$A(L|L) $S_{L,3}$= .0096	$S_{L,4}$ = E x max{$S_{k,3}$ x A} E(H|L) = 0.8 $S_{F,3}$ x A(L|F) = .0225*.4 max $S_{L,3}$ x A(L|L) = .0096*.6 $S_{L,4}$ = .0072 (max)

Today's objectives

- HMM, formulation of models
- Forward/backward
- Gene finding
- Profile HMMs

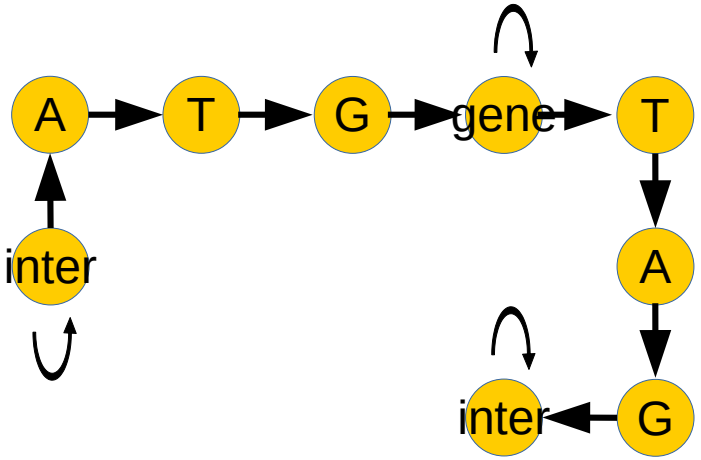
Constructing HMMs

- Labeling problems are common in genomics
- HMMs provide a general solution to labeling problems because of their flexibility



A	I	O
I	0.8	0.2
O	0.2	0.8

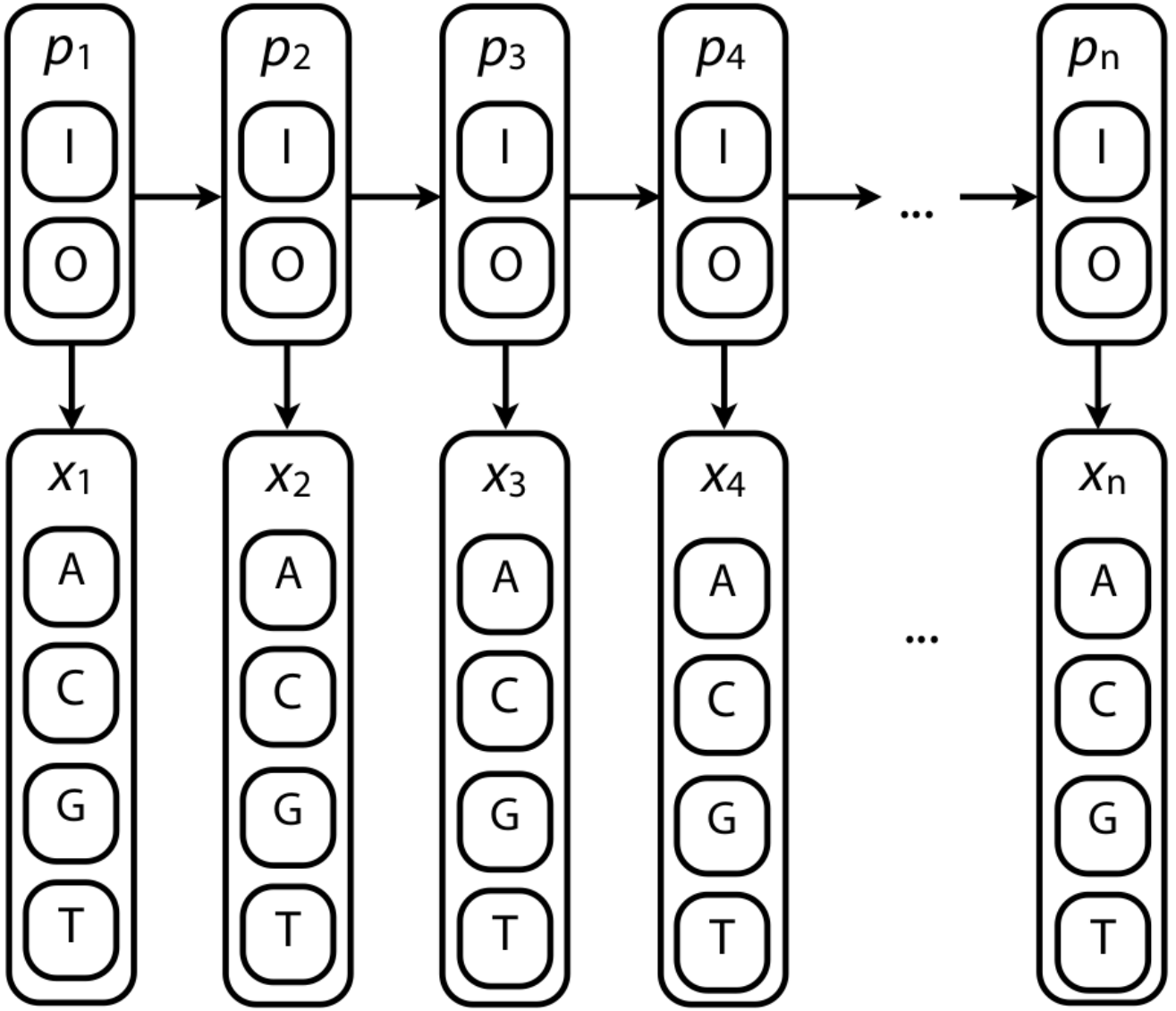
E	A	C	G	T
I	0.1	0.4	0.4	0.1
O	0.25	0.25	0.25	0.25



- HMMs can have huge number of parameters
- HMMs trained on known examples gene emits {A, G, C, T}
- HMMs learned using Baum-Welsh inter emist {A, G, C, T}

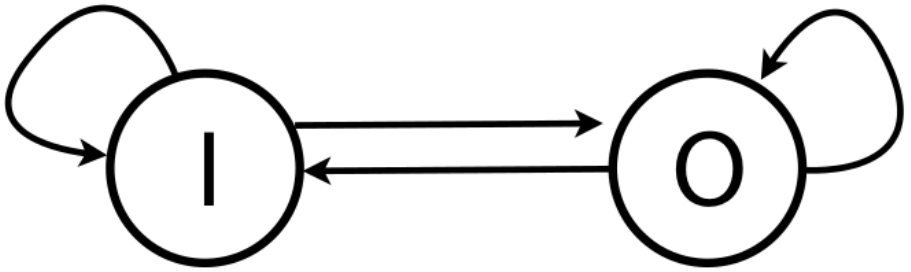
CpG island HMM

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



CpG island HMM training

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



inside outside

A	I	O
I		
O		

Transition matrix

E	A	C	G	T
I				
O				

Emission matrix

Estimate as fraction of positions where we transition from inside to outside

Estimate as fraction of nucleotides inside islands that are C

Viterbi output

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

x: ATATATACGCGCGCGCGCGATATATATATATATA

p: OOOOOOOIIIIIIIIIIIIIIOOOOOOOOOOOOO

(from Viterbi)

Viterbi output 2

Example 3 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

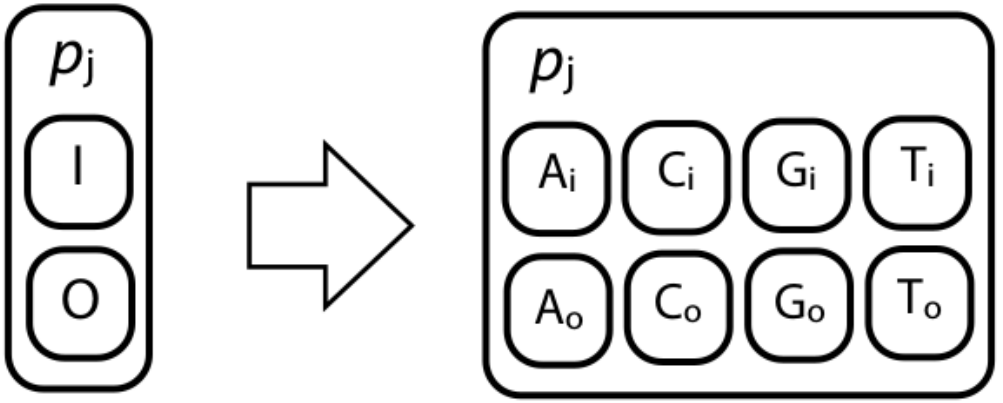
x: ATATATACCCCCCCCCCCCCCATATATATATATA

p: OOOOOOOIIIIIIIIIIIIIIOOOOOOOOOOOO

(from Viterbi)

Oops - not a CpG island!

Second HMM try

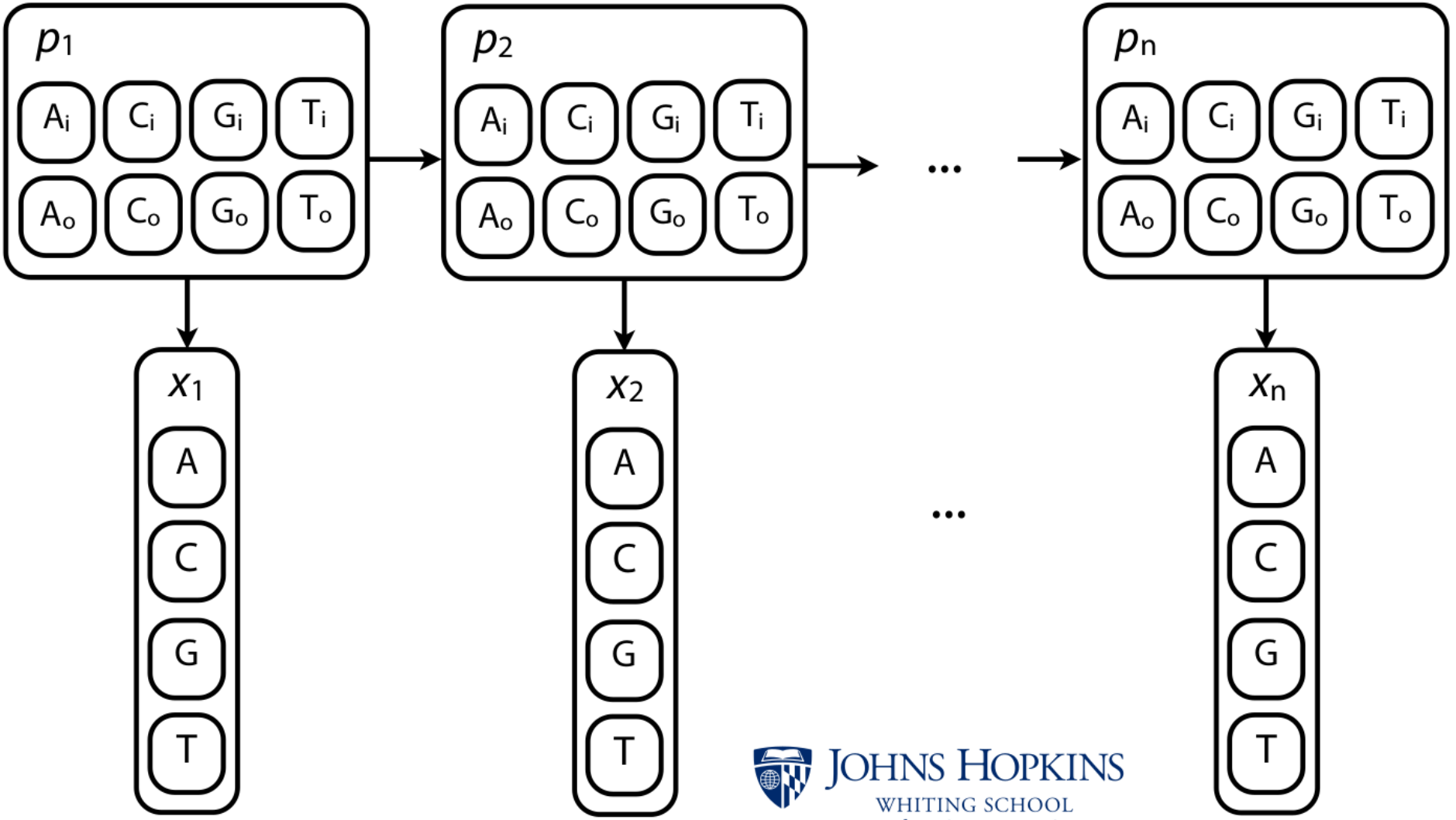


$Q = \{ \mathrm{I}, \mathrm{O} \}$

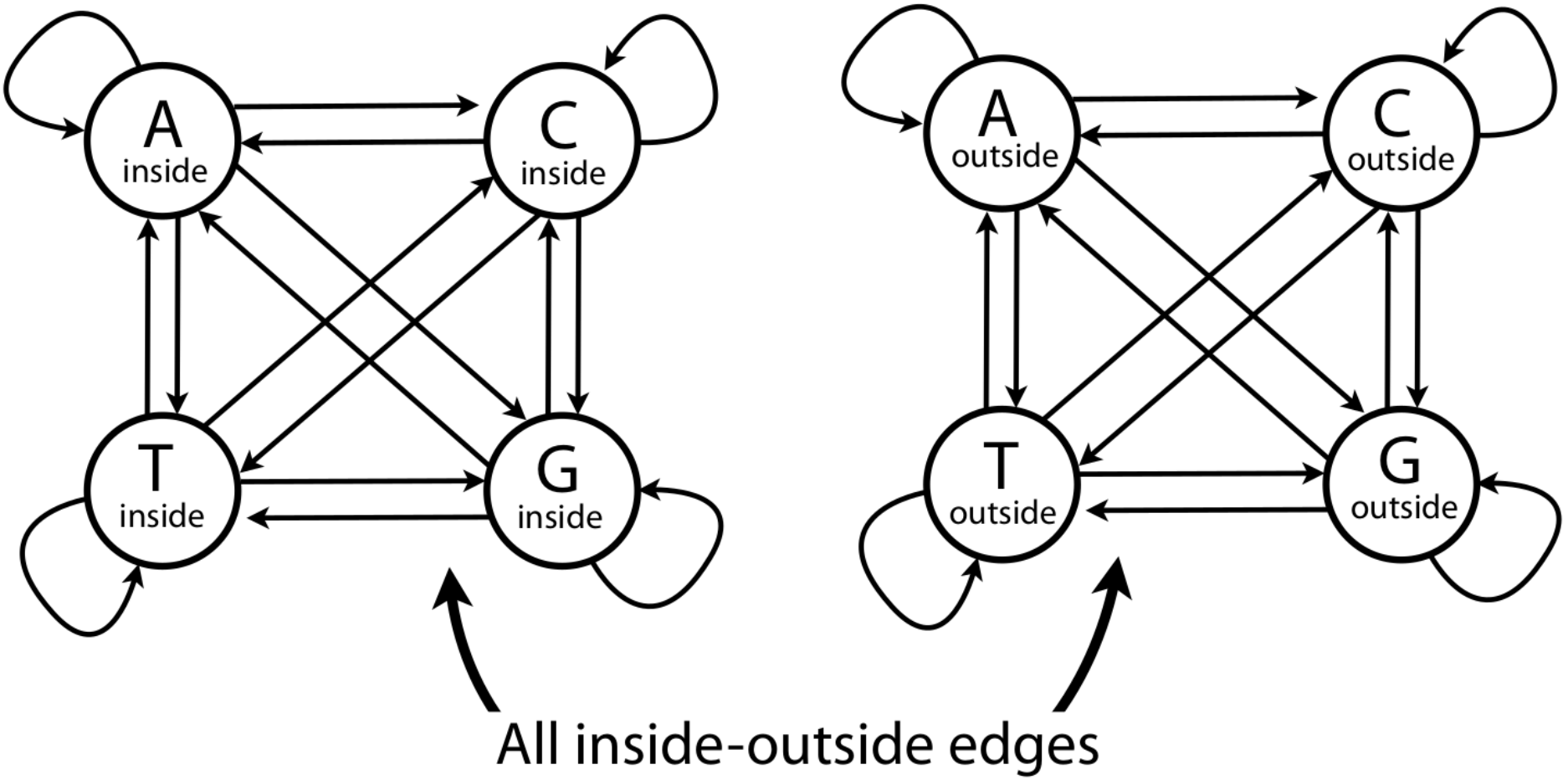
$Q = \{ \mathrm{I}, \mathrm{O} \} \times \{ \mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T} \}$

Second HMM try

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



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



Second HMM try

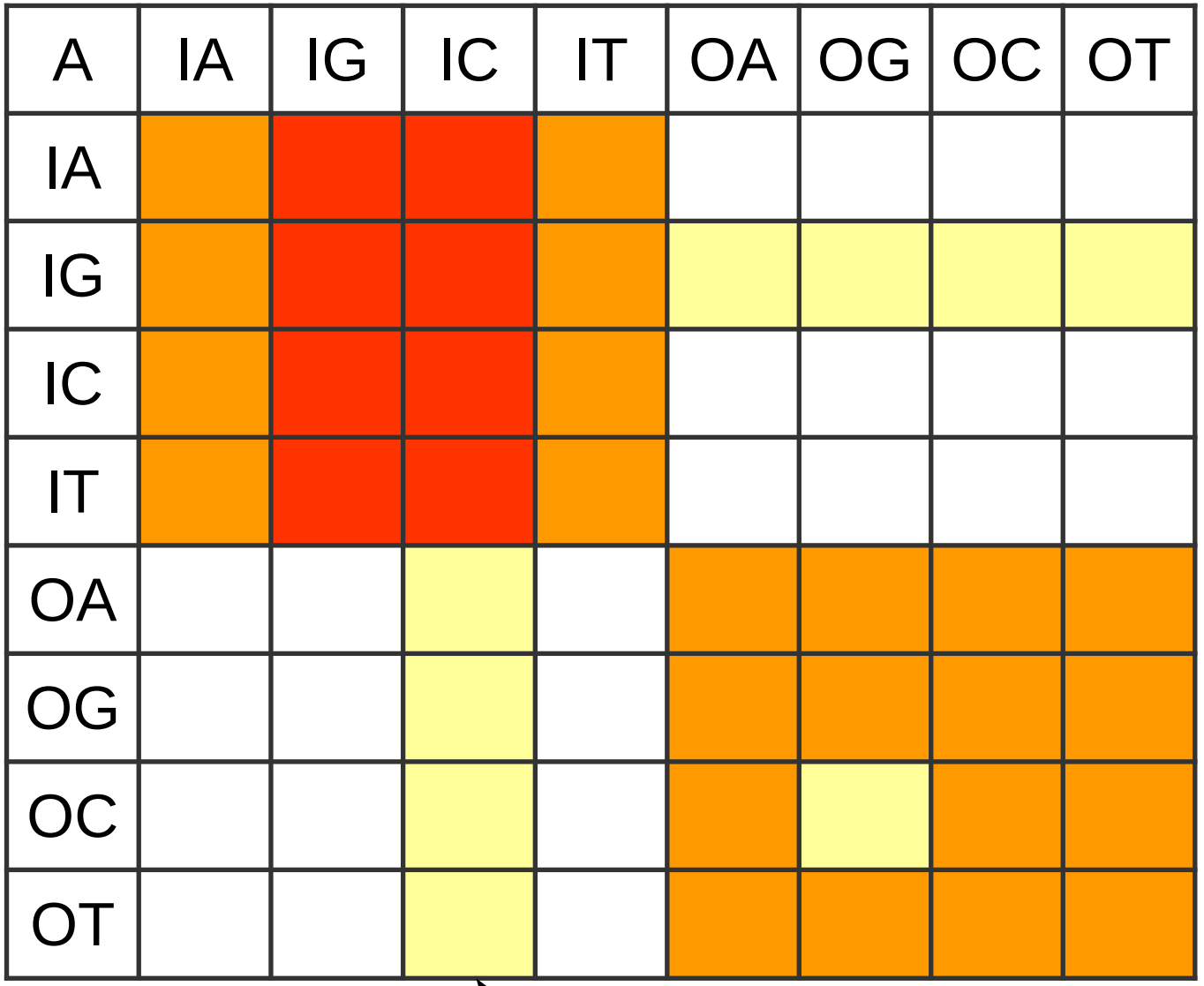
Idea 2: $Q = \{ A_i, C_i, G_i, T_i, A_o, C_o, G_o, T_o \}$, $\Sigma = \{ A, C, G, T \}$

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								

Estimate $P(C_i \mid T_i)$ as fraction of all dinucleotides where first is an inside T, second is an inside C

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

Filling in with real data



red: highest probability
Orange: high probability
Yellow: low probability
White: zero

CpG islands end with a
CG not just G

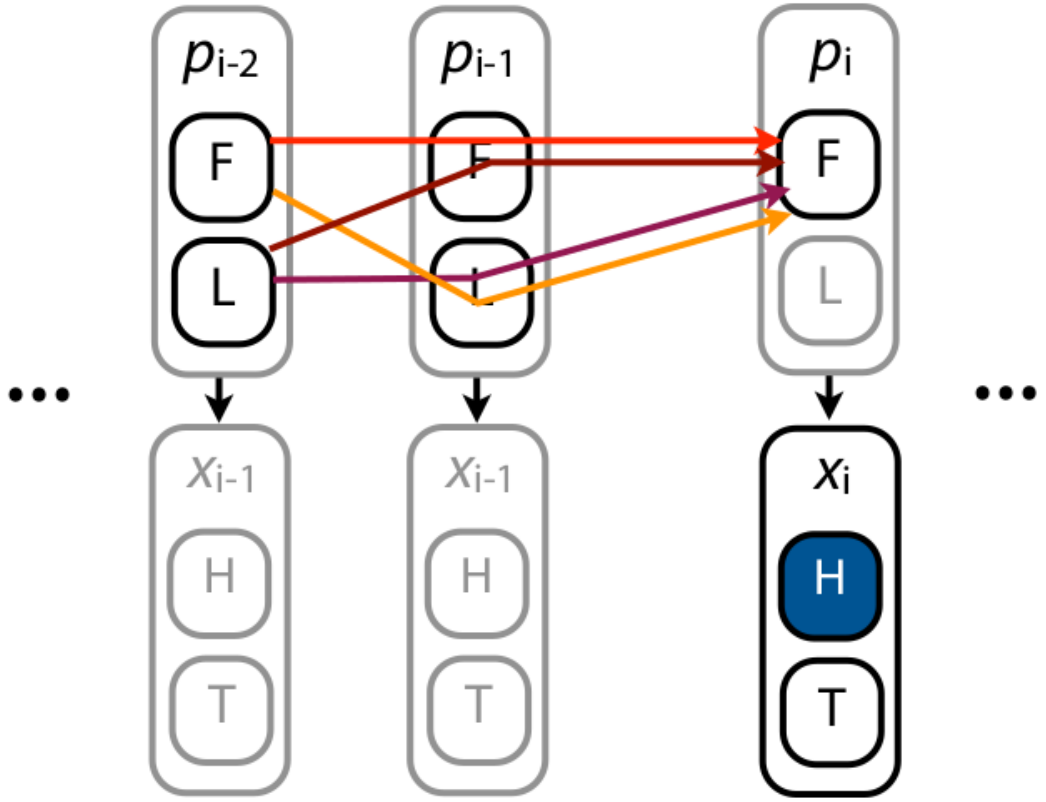
Positive:
High GC content inside
CG rare outside

Negative:
End on G start on C

CpG islands start with a
CG not just C

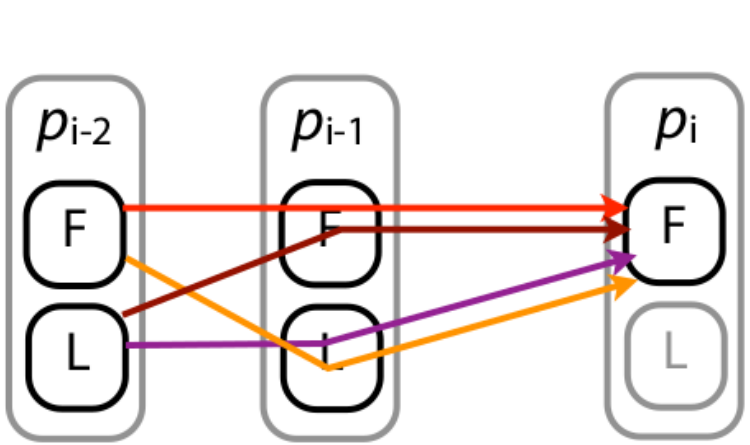
Higher order HMMs

For higher-order HMMs, Viterbi $s_{k,i}$ no longer depends on just the previous state assignment

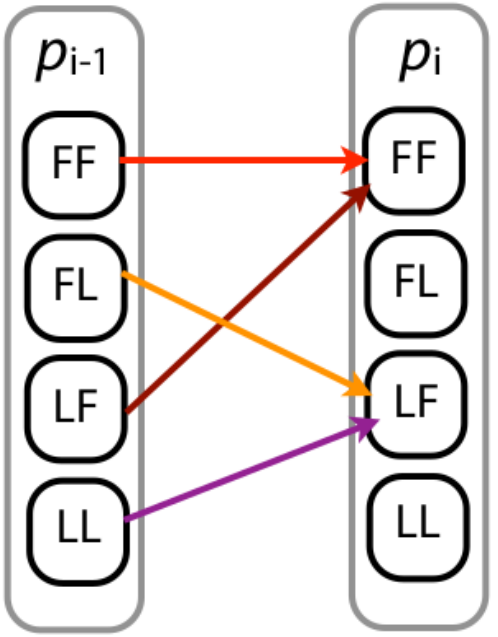


Composite states

Now *one* state encodes the last *two* "loadedness"es of the coin



Equivalent



$Q = \{ F, L \}$

```
FFLFFF
FF      p1
 FL     p2
  LF    p3
   FF   p4
    FF  p5
HHTTHH
```

$Q = \{ F, L \} \times \{ F, L \}$

Forward and Backward Algorithm

What is the joint probability of p and x?

$P(p_1,..., p_n, x_1,...., x_n)$

What is the most likely path? (decoding = viterbi algorithm)

$p^* = \text{argmax } P(p_1,..., p_n \mid x_1,...., x_n)$

What is the probability p is in state t and emitting $x_1 ... x_i$

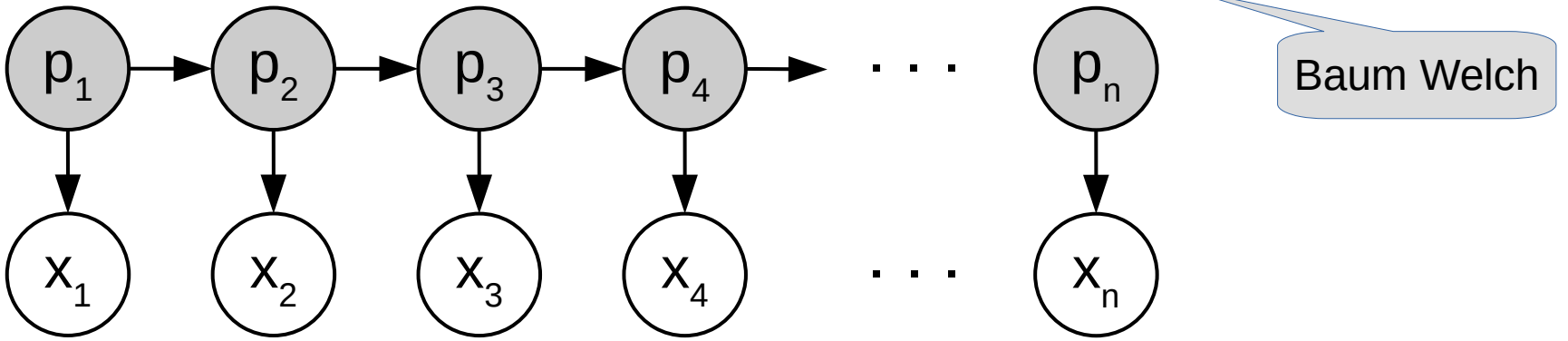
$P(p_i = t, x_1,...., x_i)$ – forward algorithm

What is the probability of emitting $x_{i+1} ... x_n$ given $p_i = t$?

$P(x_{i+1}...x_n \mid p_i = t)$ – backward algorithm

What is the conditional probability of hidden state p at site i

$P(p_i \mid x_1,...x_n)$ -- forward and backward algorithm



Forward algorithm

What is the probability that p_i = state t

$P(p_i=t, x_1 \ldots x_i)$ – forward algorithm

$$F(t,i) = E_t[x_i] \sum_{s \in Q} F(s, i-1) A[s,t]$$

Emission

Recursion Probability

Transition

Sum instead of max (viterbi)

Backward algorithm

What is the probability of emitting x_i ... x_n given p_i = t?

$P(x_{i+1}...x_n \mid p_i = t)$ – backward algorithm

$$B(t,i) = \sum_{s \in Q} E_s[x_{i+1}] B(s, i+1) A[t,s]$$

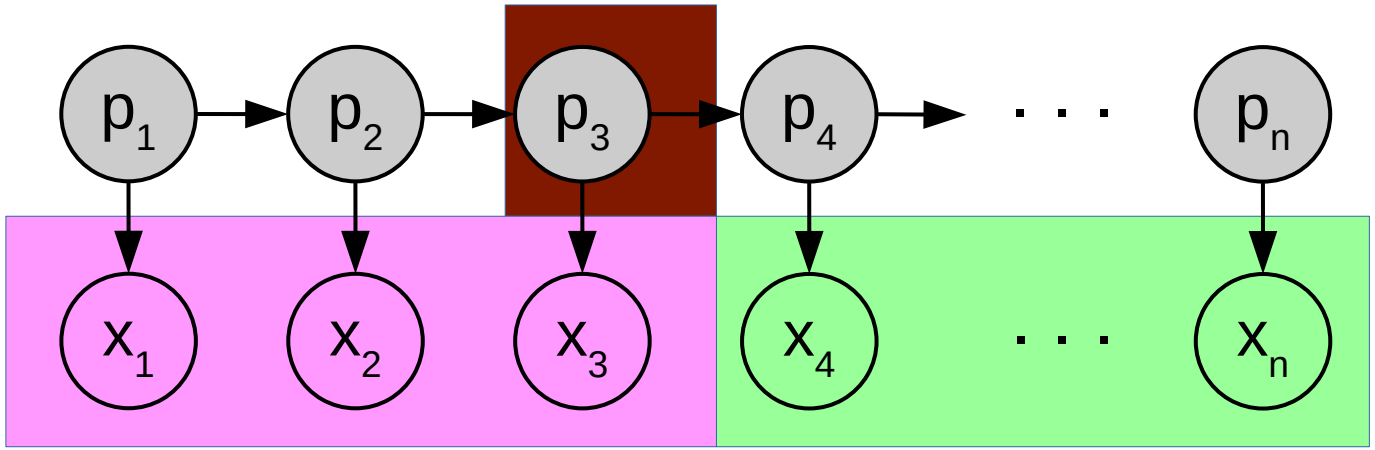
Emission

Recursion Probability

Transition

Sum instead of max (viterbi)

Forward-Backward Algorithm



Forward $P(p_3, x_1 \ldots x_3)$ Backward $P(x_4 \ldots x_n \mid p_3)$

Foward/backward

$P(p_3 \mid x_1 \ldots x_n)$ proportional to $P(p_3, x_1 \ldots x_n)$

$P(p_3, x_1 \ldots x_n) = P(p_3, x_1 \ldots x_3) P(x_4 \ldots x_n \mid p_3)$

$= F(t,i) B(t,i)$

What is the conditional probability of hidden state p at site i

$P(p_i \mid x)$ -- forward and backward algorithm

Baum-Welch algorithm

Pseudocounts: an amount (small constant) added to the number of observed cases in order to change the expected probability in a model of those data, when not known to be zero. (avoids P = 0)
Training when rates are unknown

The Baum–Welch algorithm uses the Expectation Maximization (EM) algorithm to find the maximum likelihood estimate of the parameters of a HMM given a set of observed feature vectors.

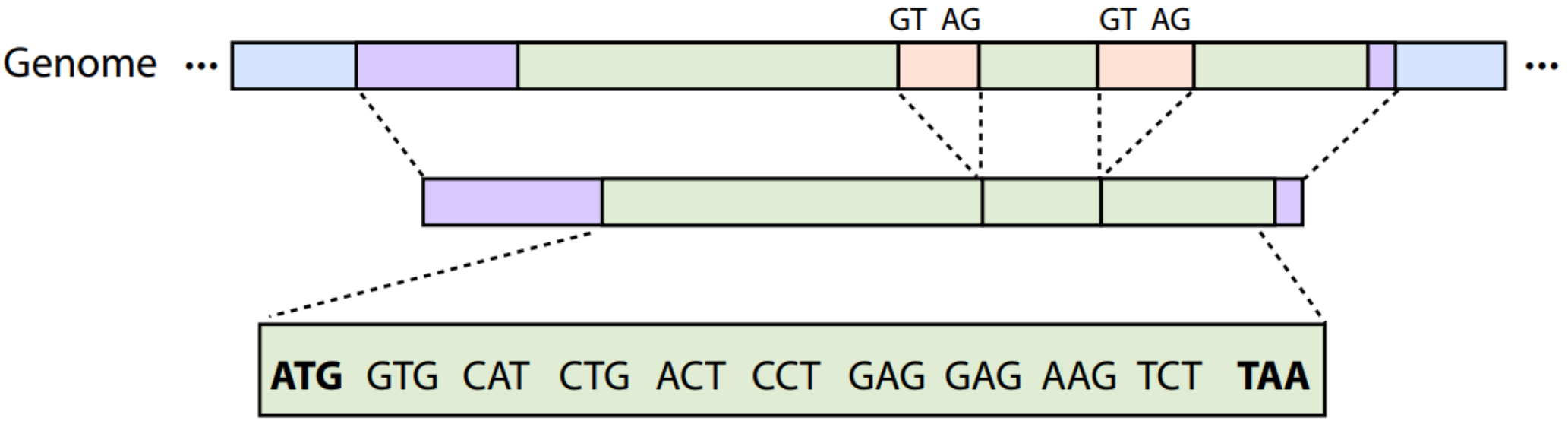
Baum-Welch Algorithm
Initialize by picking arbitrary model parameters
Recurrence:
- Set all A and E variables to their pseudocount values
- For each sequence j = 1..n:
  - Calculate F(t,i) for sequence j using forward
  - Calculate B(t,i) for sequence j using backward
  - Add contribution of sequence j to A and E
- Calculate new model parameters
- Calculate log likelihood

Termination: change in log likelihood is small or max iterations

Eukaryotic genes: a challenge

HMM gene finder

Parts: non-genes, exons (both coding and non-coding portions), introns



Sequence signals: acceptors, donors, branch sites, pyrimidine-rich sites, nucleotide compositions

Nucleotide composition and codon bias

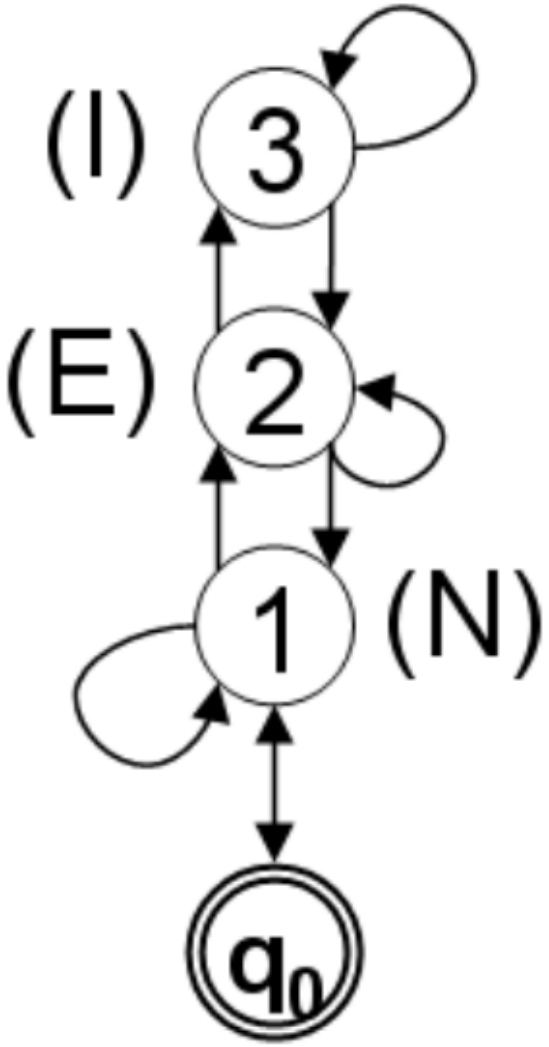
HMM for genes

Attempt 1:

Emissions are nucleotides

I = intron
E = exon
N = intergenic (between genes)

q0 is a *start state*; guarantees we start in the N (intergenic) state



Each with their own nucleotide frequencies

Model captures:

Genes, exons and introns

Does not capture:

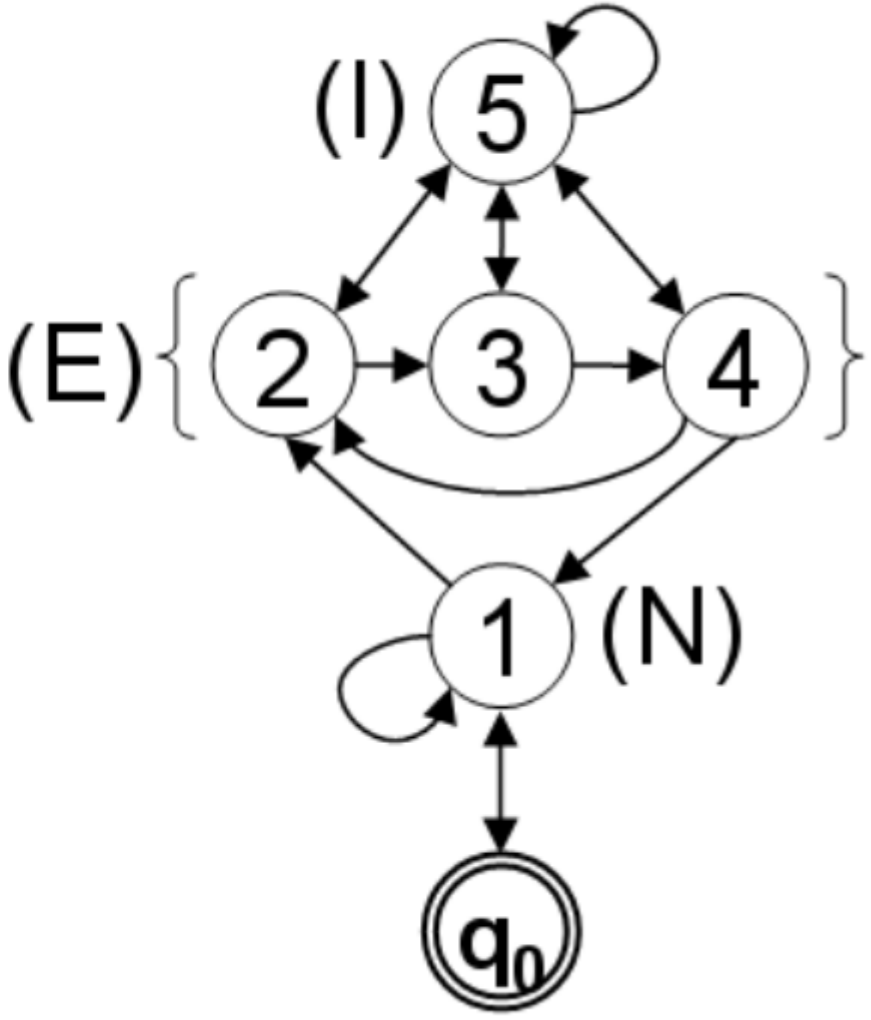
Start/stop codons, acceptors/donors, codons

Problem
Codon bias
Sum of exons not multiple of 3

Attempt 2:

Three exon states
additionally capture *codons*

Problems
Sum exons = 3
11 234 255534 255234
Splice junction (AG)
Start and Stop codons



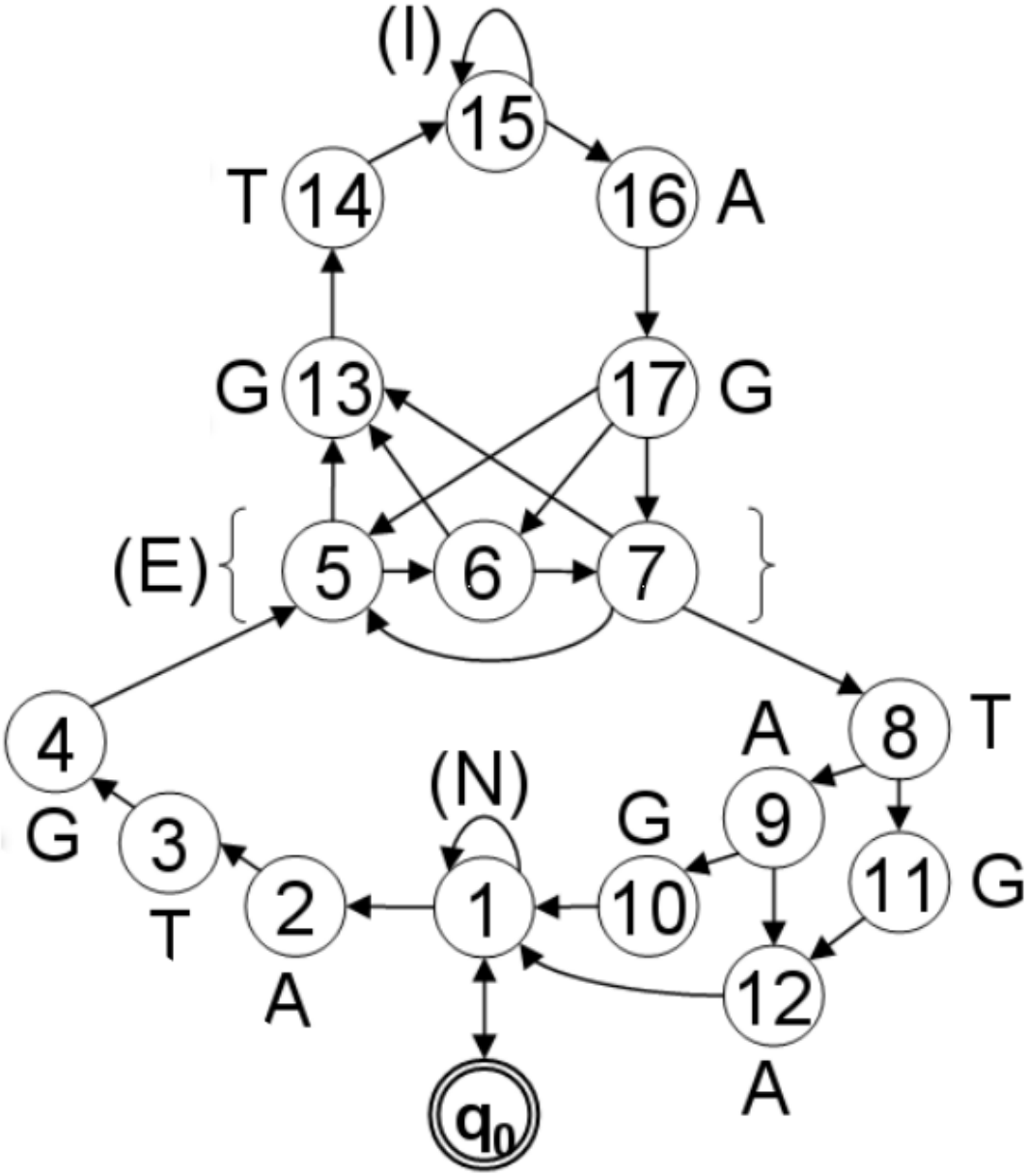
Attempt 3:

States 2-4 capture start codon

13, 14 capture donor

16, 17 capture acceptor

8-12 capture stop codons



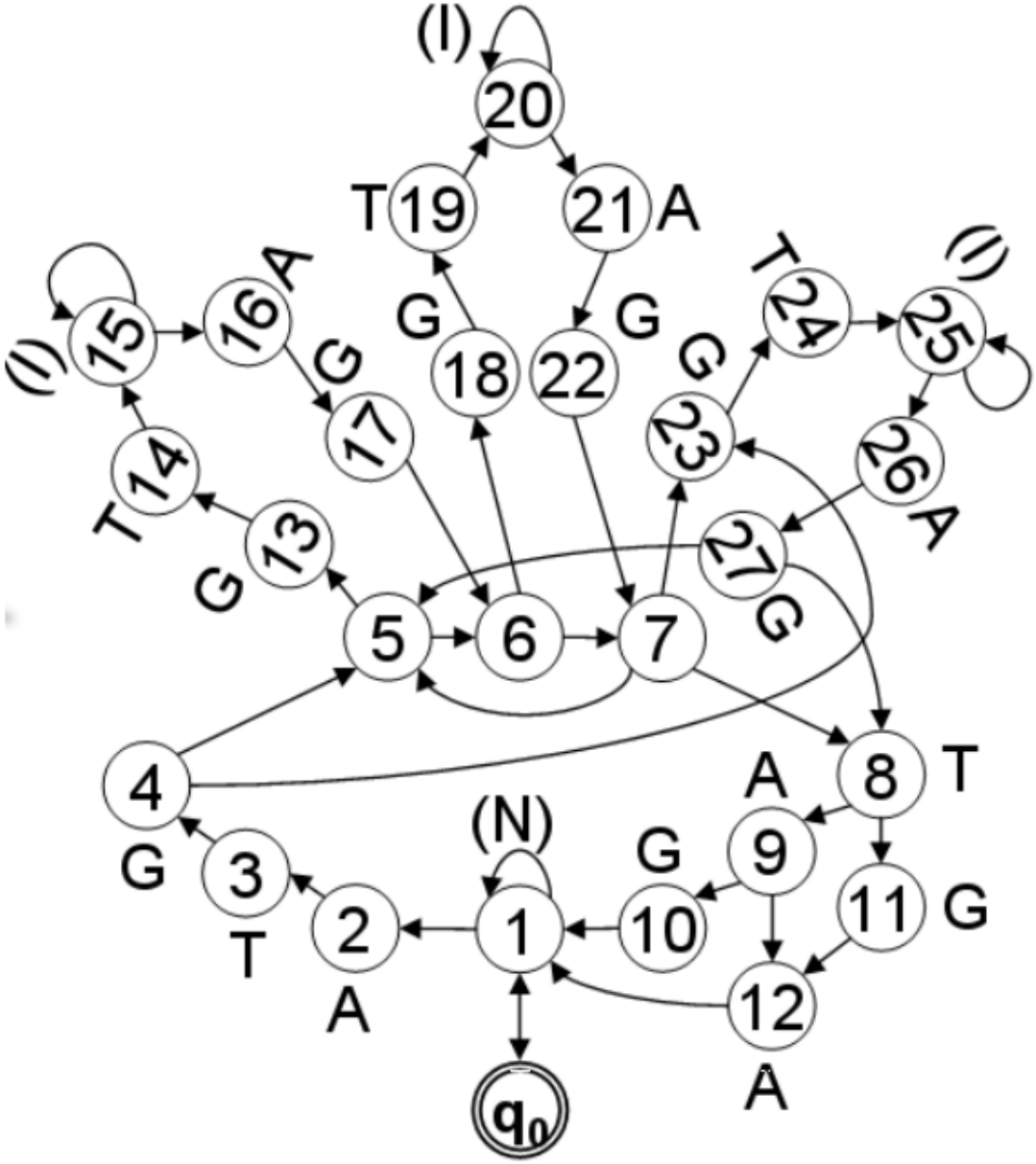
Why not arrow from 5 to 8?
Codon would be 1TGA|1TAG|1TAA – not a stop
Do exons end 2nd position in coding and start at 3rd?
Not always

Attempt 4:

Additional copies of the accepter/intron/donor loop allow us to pick up where we left off in reading frame

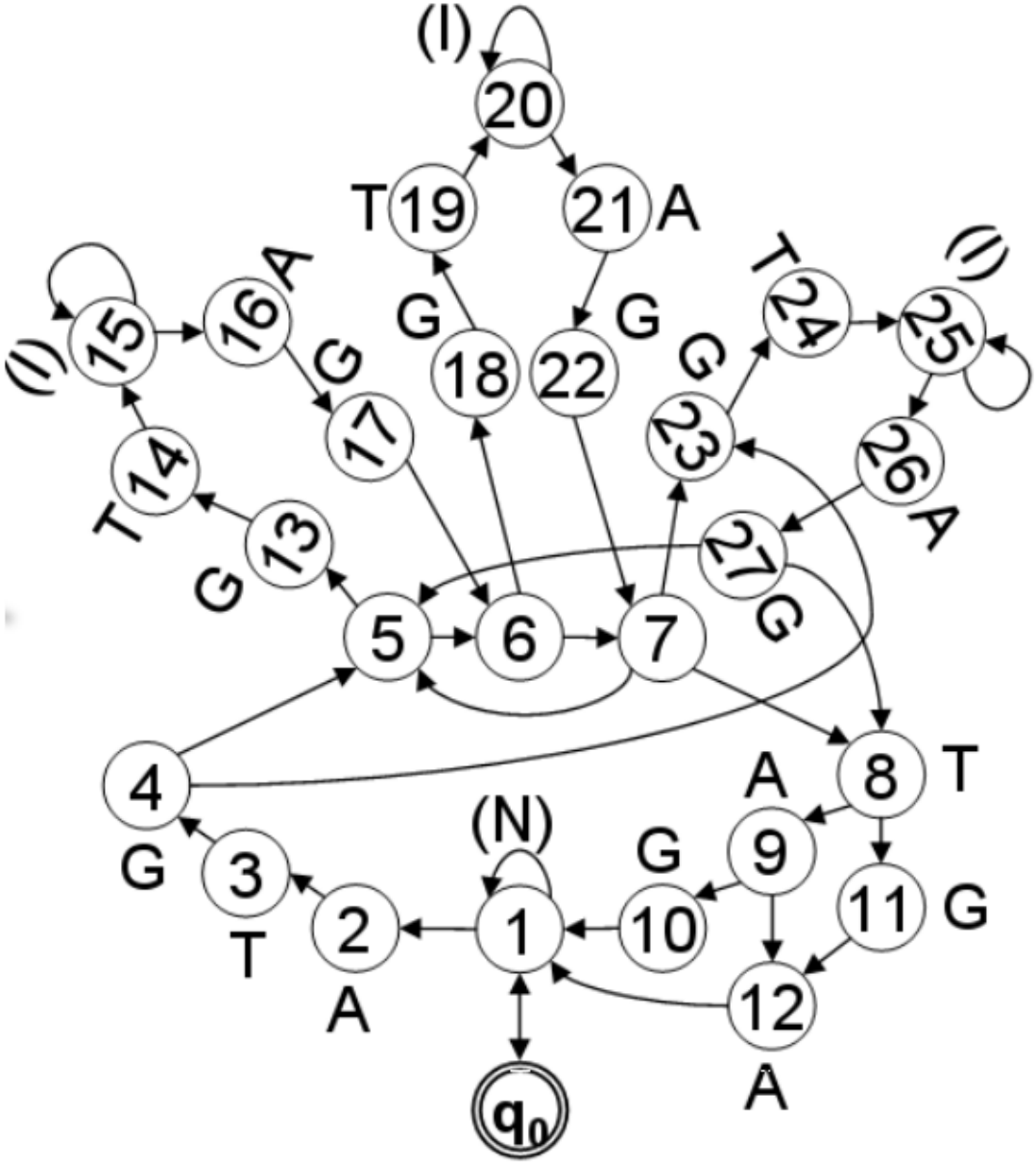
Recall transition probability matrix has $|Q|^2$ elements

27 states → 729 transition probs

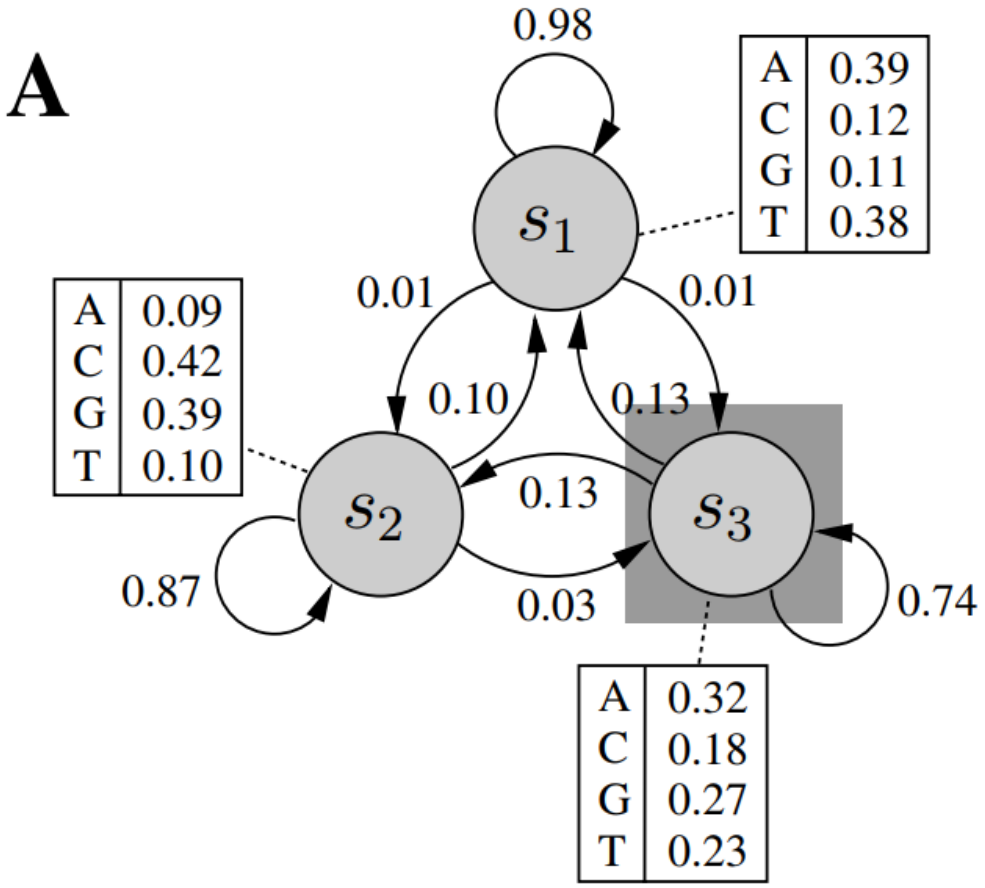


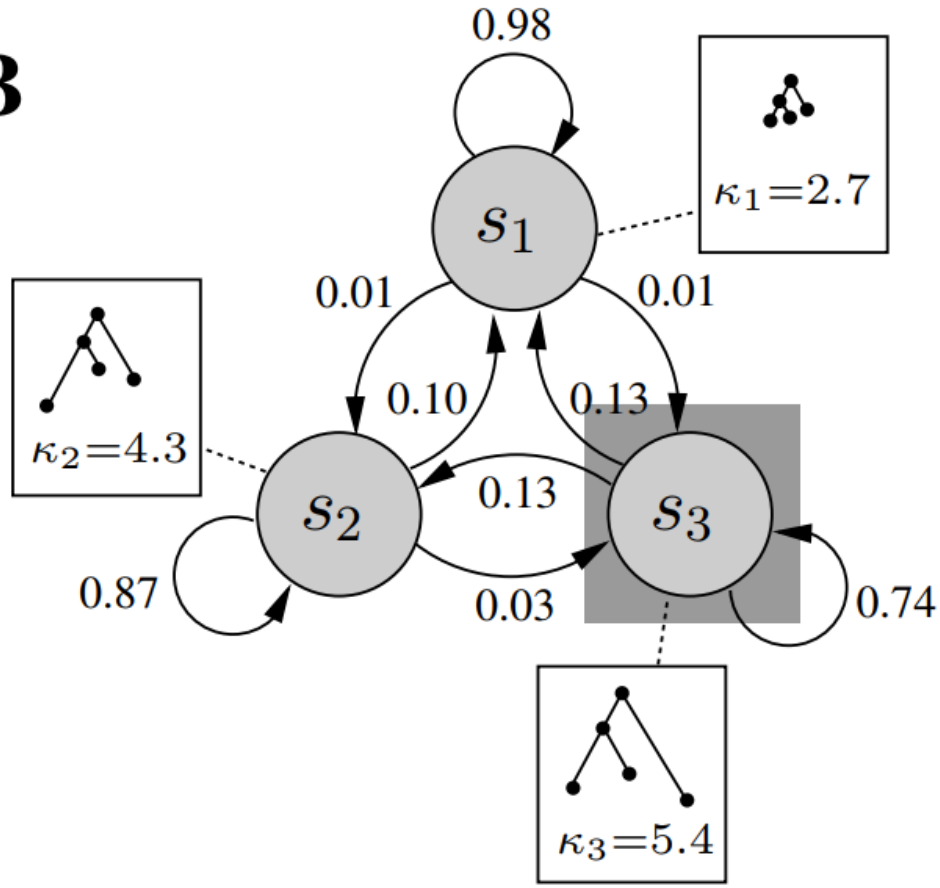
How is codon bias incorporated?

States 5, 6, 7
7 = third position
7 emits GC (bias) more than AT



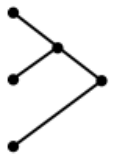
Phylogenetic Hidden Markov Model: Markov chains in time and space



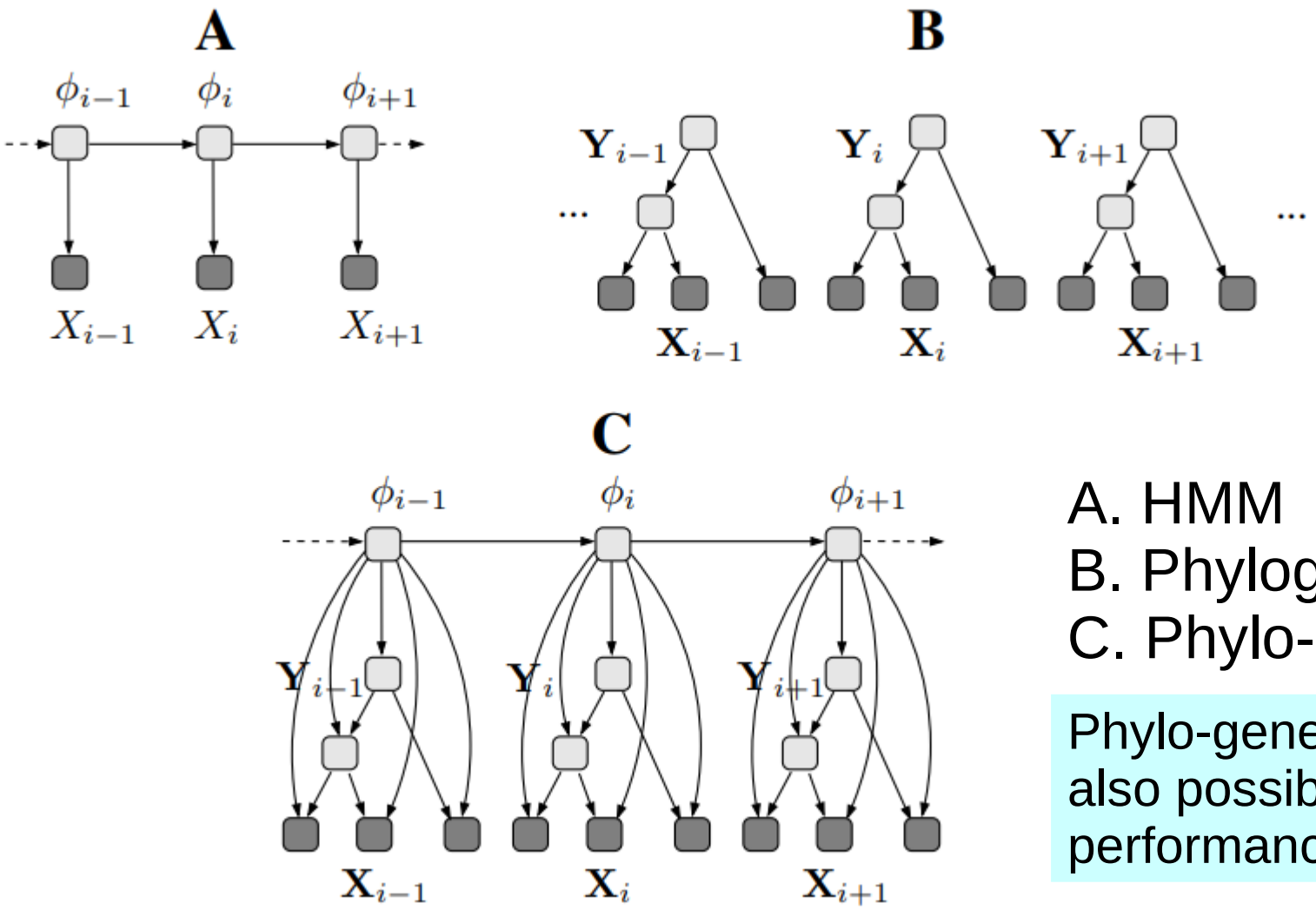


$$\mathbf{X} = \texttt{TAACGGCAGA}\ldots$$

$$\mathbf{X} = \begin{matrix}\texttt{TAACGGCAGA}\ldots \\ \texttt{TTAGGCAAGG}\ldots \\ \texttt{AAGGCGCCGA}\ldots\end{matrix}$$



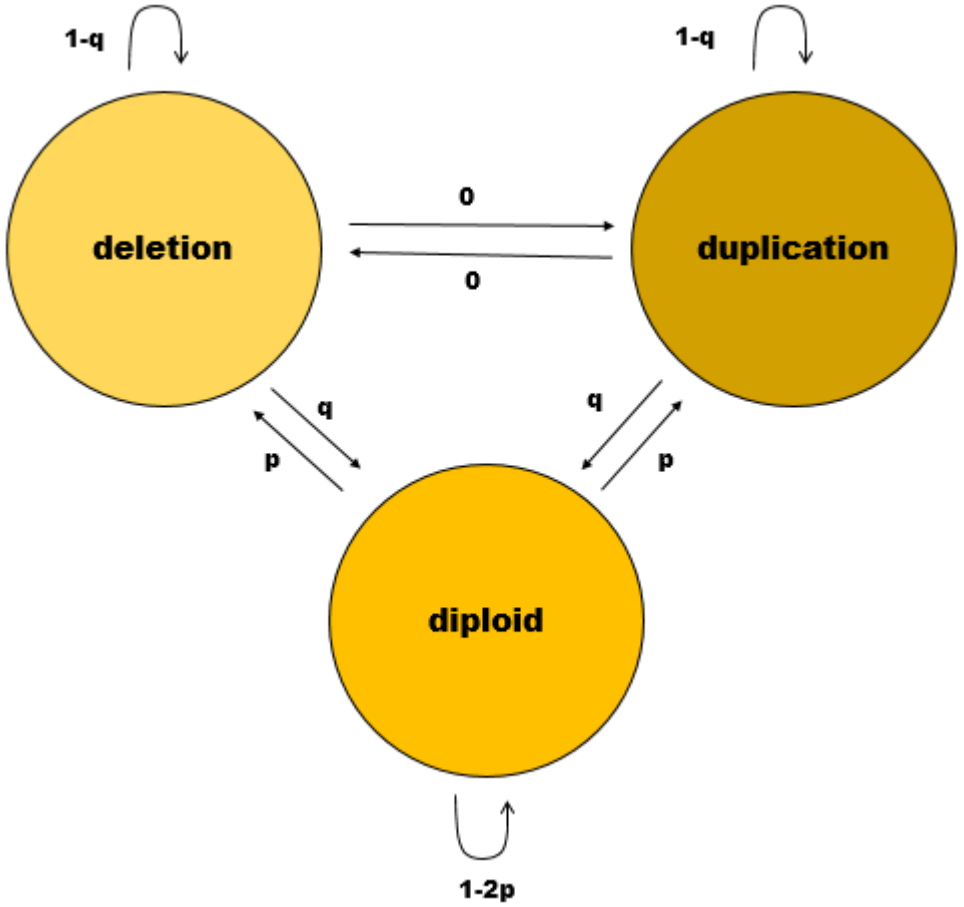
HMM in time and space

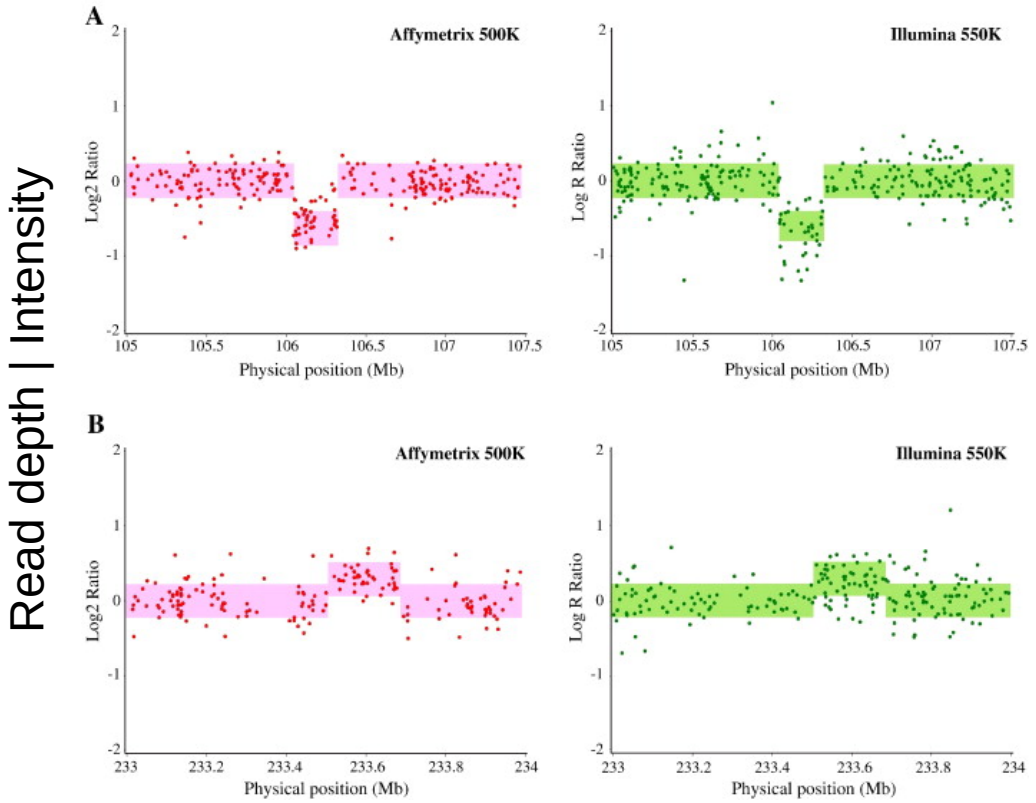


A. HMM
B. Phylogenetic model
C. Phylo-HMM

Phylo-gene finding HMM is also possible and improves performance

Copy number variation





Physical position on chromosome

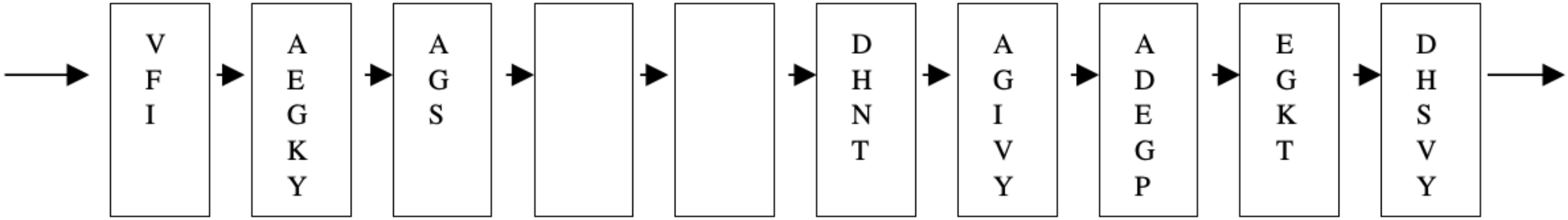
Emissions: read counts or log2(ratio) is discretized

Profile HMMs

Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using profile HMMs (HMMER software)

```
HBA_HUMAN     ...VGA--HAGEY...
HBB_HUMAN     ...V----NVDEV...
MYG_PHYCA     ...VEA--DVAGH...
GLB3_CHITP    ...VKG------D...
GLB5_PETMA    ...VYS--TYETS...
LGB2_LUPLU    ...FNA--NIPKH...
GLB1_GLYDI    ...IAGADNGAGV...
"Matches":       ***   *****
```

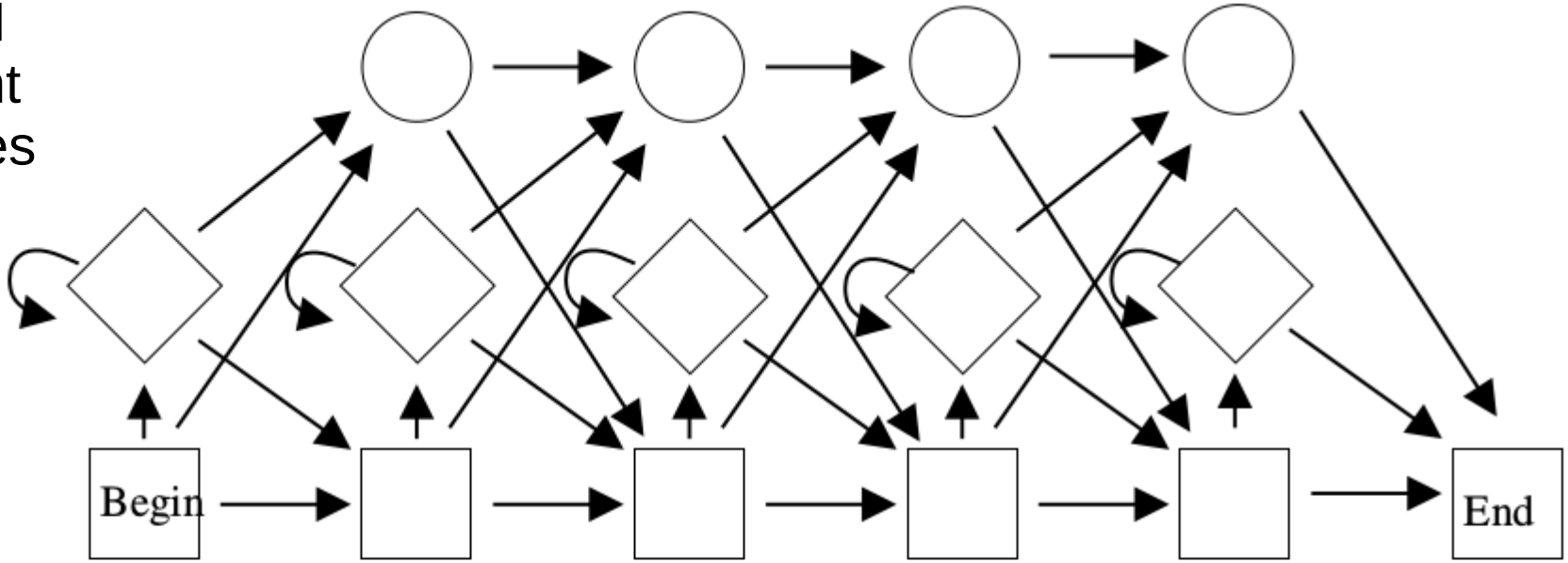
- Improve genome annotations
- Better remote homology
- Better handle variable domain architectures

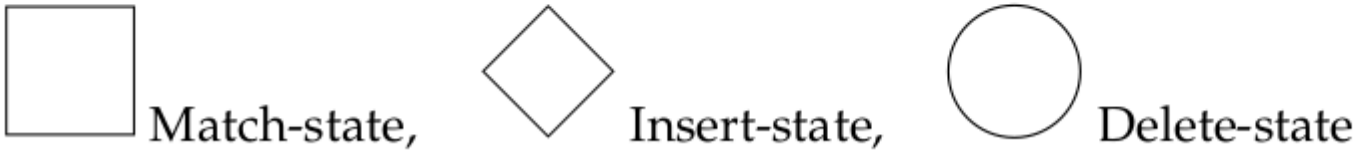


Profile HMM is equivalent to position specific scoring matrix (PSSM)

Profile HMMs

- Make close alignment
- Train HMM
- Find distant homologues

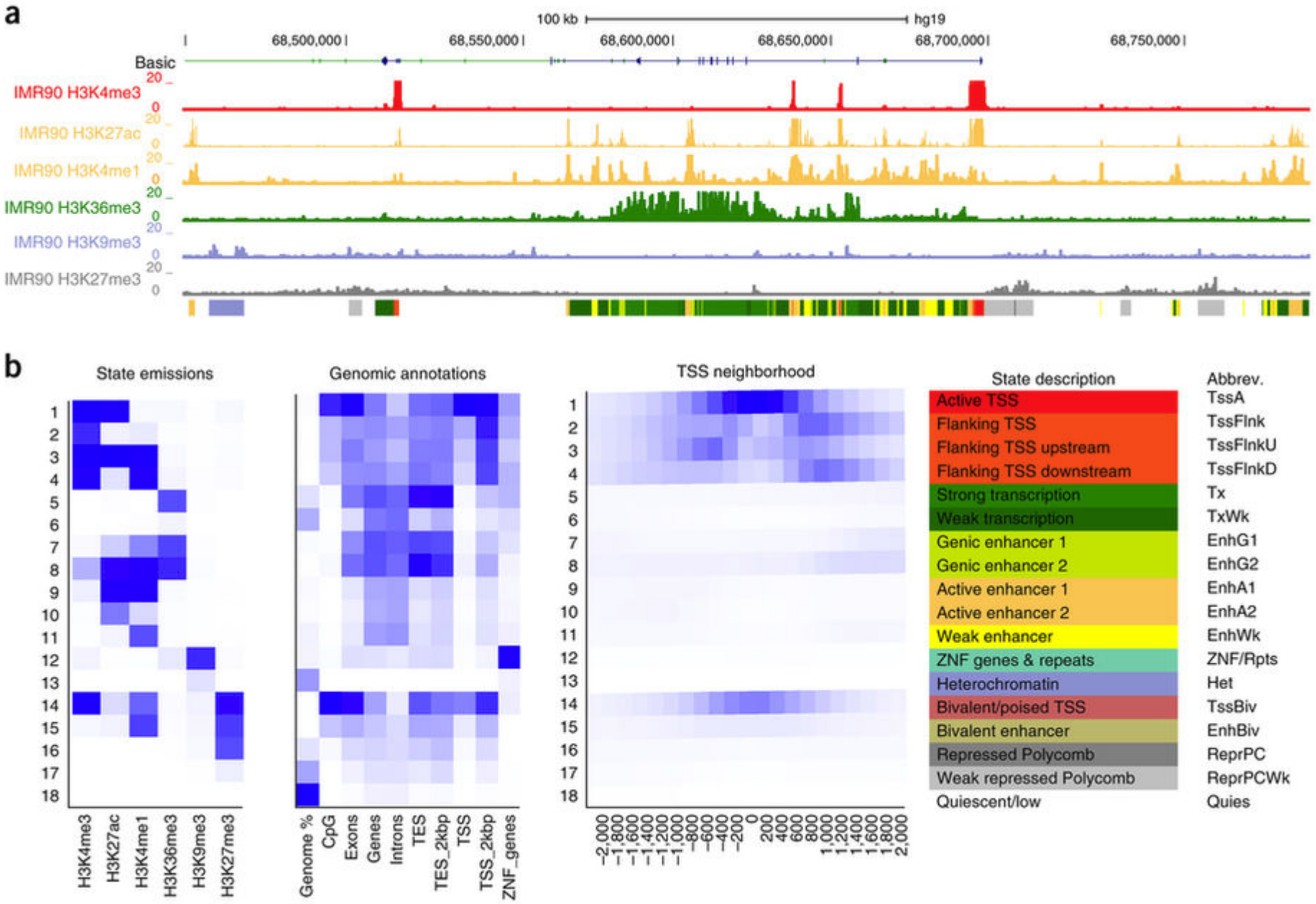




Delete-states are silent and do no have emissions
Insert and delete states account for aligned columns with gaps
Insert/delete if more than 50% sequences gapped.

Example of Hmmer alignment

Chromatin-HMM



Exercises

1) Fill in the A and E matrix given this training data

E	A	G	C	T
I				
O				

A	I	O
I		
O		

`Sequence:` `AAAAATCGGGATAT`

`Labels:` `OOOOOIIIIIOIOI`

2) Huntington's disease is caused by $(CAG)_n$ repeats. Propose an HMM that would identify such repeats.

3) Give one possible sequence and labels for the following HMM, assuming orange is possible, white is not possible, IA emits A, etc.

A	IA	IG	IC	IT	OA	OG	OC	OT
IA								
IG								
IC								
IT								
OA								
OG								
OC								
OT								

Exercises

1) What are pseudocounts used to avoid in the EM algorithm?

2) Why does this model not work well in identifying CpG islands:

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

3) Why do profile HMMs work better for distant homology searches?

4) Propose an HMM model (A and E matrix) for the following:

```
Sequence:     ATCGAAAATCGGGATATATATGACTTAATTCTCGTA

Labels:       OOOOOOOOOOOOIIIIIIIIOOOOOOOOOOOOOO
```