

# 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  $x = \text{AATTCTG}$  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

$x = \text{AATTCTG}$

$$P(x) = 0.36 \times 0.38 \times 0.19 \times 0.14 \times 0.19 \times 0.16$$

$$P(x) = 1.10622e-04$$

## 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

$x = \text{AATTCTG}$

$$P(x) = 0.04 \times 0.21 \times 0.31 \times 0.25 \times 0.34 \times 0.31$$

$$P(x) = 6.86e-05$$

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	

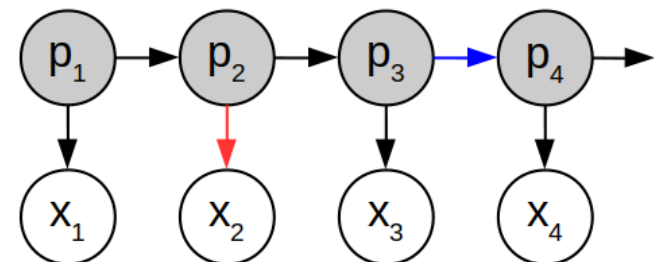
	E	A	G	C	T
I	0.1	0.4	0.4	0.1	
O	0.25	0.25	0.25	0.25	

AACG  
OOII

$P = \text{initial} * 0.8 * 0.2 * 0.8$  (transitions)  
 $0.25 * 0.25 * 0.4 * 0.4$  (emissions)  
 If initial = 0.5,  $P = 0.00064$

7) In the diagram below, we observed  $x_1$ - $x_4$  but not  $p_1$ - $p_4$ :

- a) does  $P(p_3)$  depend on  $p_2$ ? **yes**
- b) does  $P(p_3)$  depend on  $x_3$ ? **yes**
- c) does  $P(p_3)$  depend on  $x_2$ ? **yes**
- d) does  $P(p_3)$  depend on  $x_4$ ? **yes**
- e) does  $P(p_3|p_2)$  depend on  $x_2$ ? **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

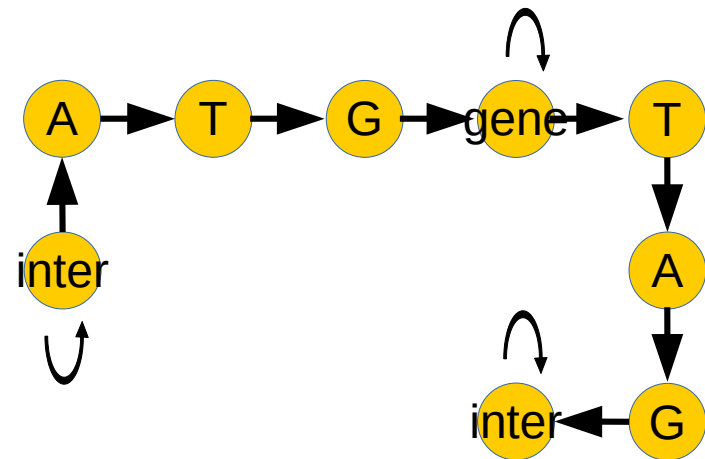
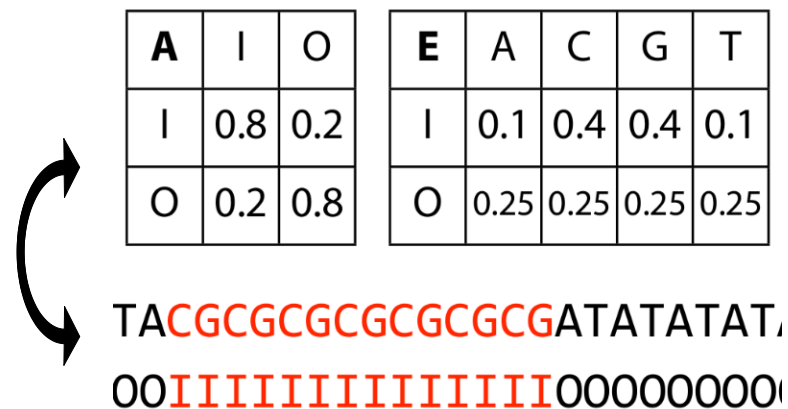
X	T	H	T	H
$S_{F,i}$	$E(T F)$ $A(F)$ $S_{F,1} = .25$	$E(H F) = .5$ $\underline{S_{E,1}} \underline{A(F F)}$ $S_{L,1} A(F L)$ $S_{F,2} = .075$	$E(T F) = .5$ $\underline{S_{E,2}} \underline{A(F F)}$ $S_{L,2} A(F L)$ $S_{F,3} = .0225$	$S_{F,4} = E \times \max\{S_{k,3} \times A\}$ $E(H F) = 0.5$ $\underline{S_{E,3}}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

# Today's objectives

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

# Constructing HMMs

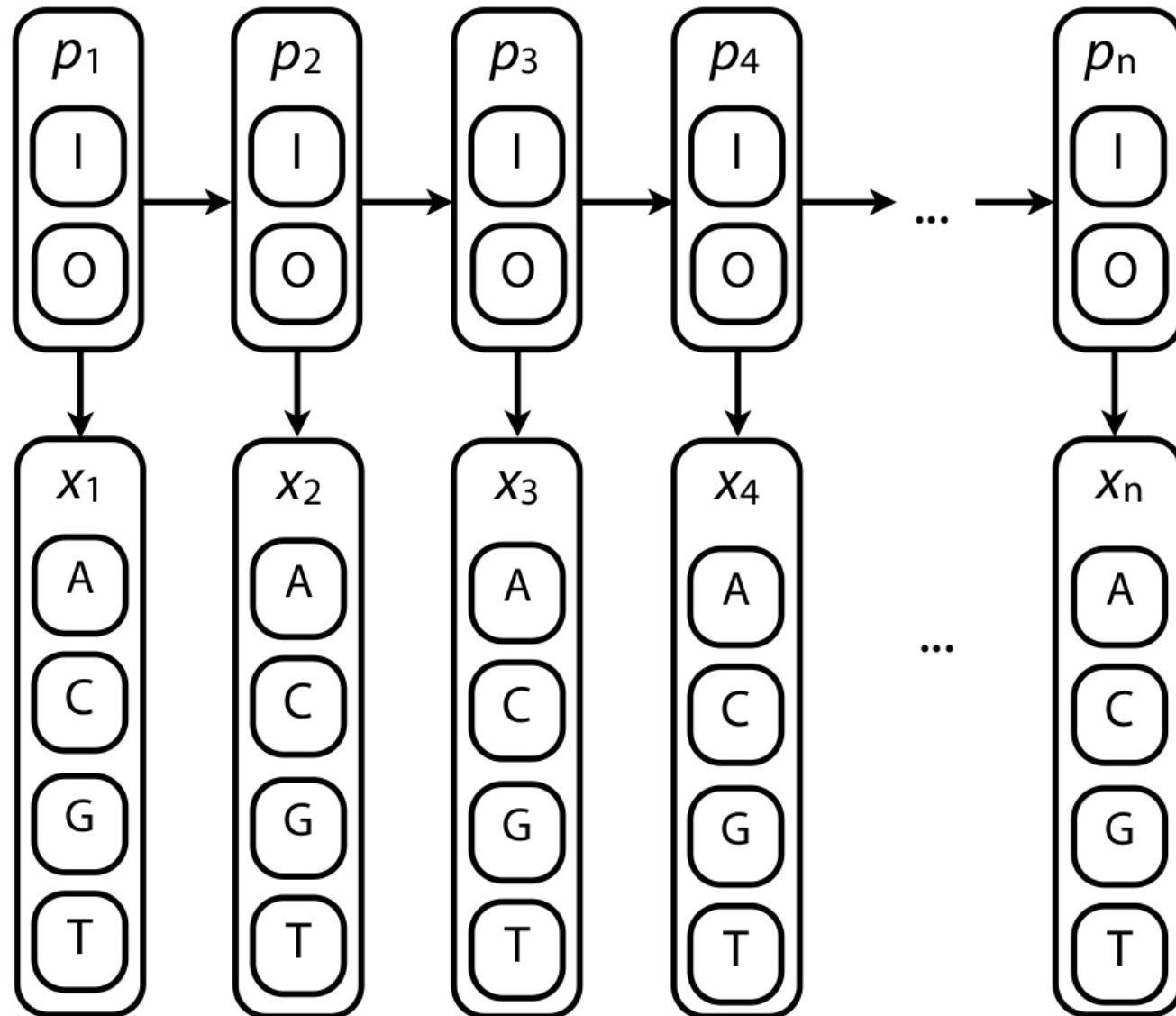
- Labeling problems are common in genomics
- HMMs provide a general solution to labeling problems because of their flexibility



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

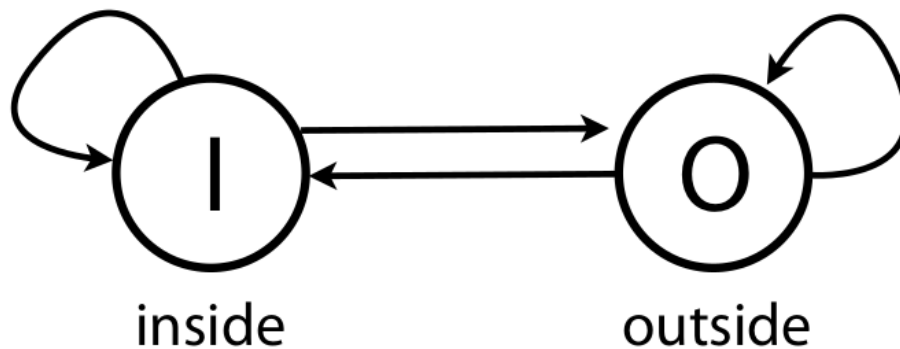
# CpG island HMM

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



# CpG island HMM training

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



<b>A</b>	<b>I</b>	<b>O</b>
<b>I</b>		●
<b>O</b>		

Transition matrix

<b>E</b>	<b>A</b>	<b>C</b>	<b>G</b>	<b>T</b>
<b>I</b>		●		
<b>O</b>				

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:

<b>A</b>	<b>I</b>	<b>O</b>
<b>I</b>	0.8	0.2
<b>O</b>	0.2	0.8

<b>E</b>	<b>A</b>	<b>C</b>	<b>G</b>	<b>T</b>
<b>I</b>	0.1	0.4	0.4	0.1
<b>O</b>	0.25	0.25	0.25	0.25

x: ATATATACGCGCGCGCGCGCGATATATATATATA

p: 00000000IIIIIIIIIIIIII0000000000000000

(from Viterbi)

# Viterbi output 2

Example 3 using HMM idea 1:

<b>A</b>	<b>I</b>	<b>O</b>
<b>I</b>	0.8	0.2
<b>O</b>	0.2	0.8

<b>E</b>	<b>A</b>	<b>C</b>	<b>G</b>	<b>T</b>
<b>I</b>	0.1	0.4	0.4	0.1
<b>O</b>	0.25	0.25	0.25	0.25

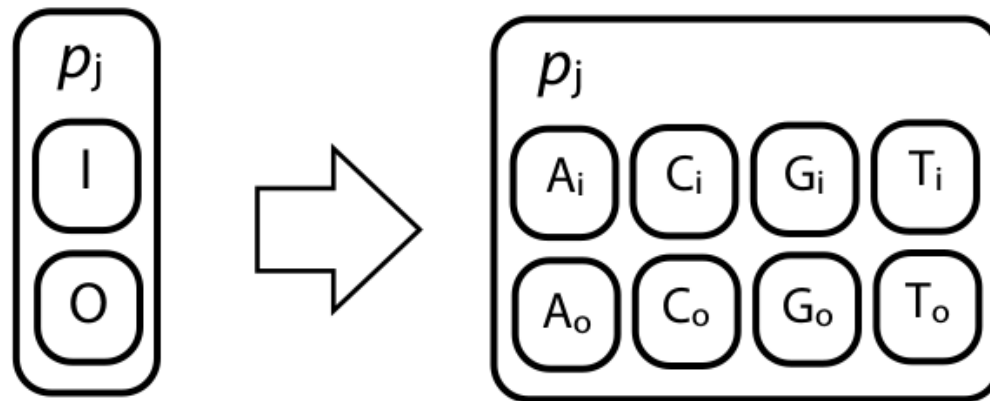
x: ATATATACCCCCCCCCCCCCCATATATATATATA

p: 00000000IIIIIIIIIIIIII0000000000000000

(from Viterbi)

Oops - not a CpG island!

# Second HMM try

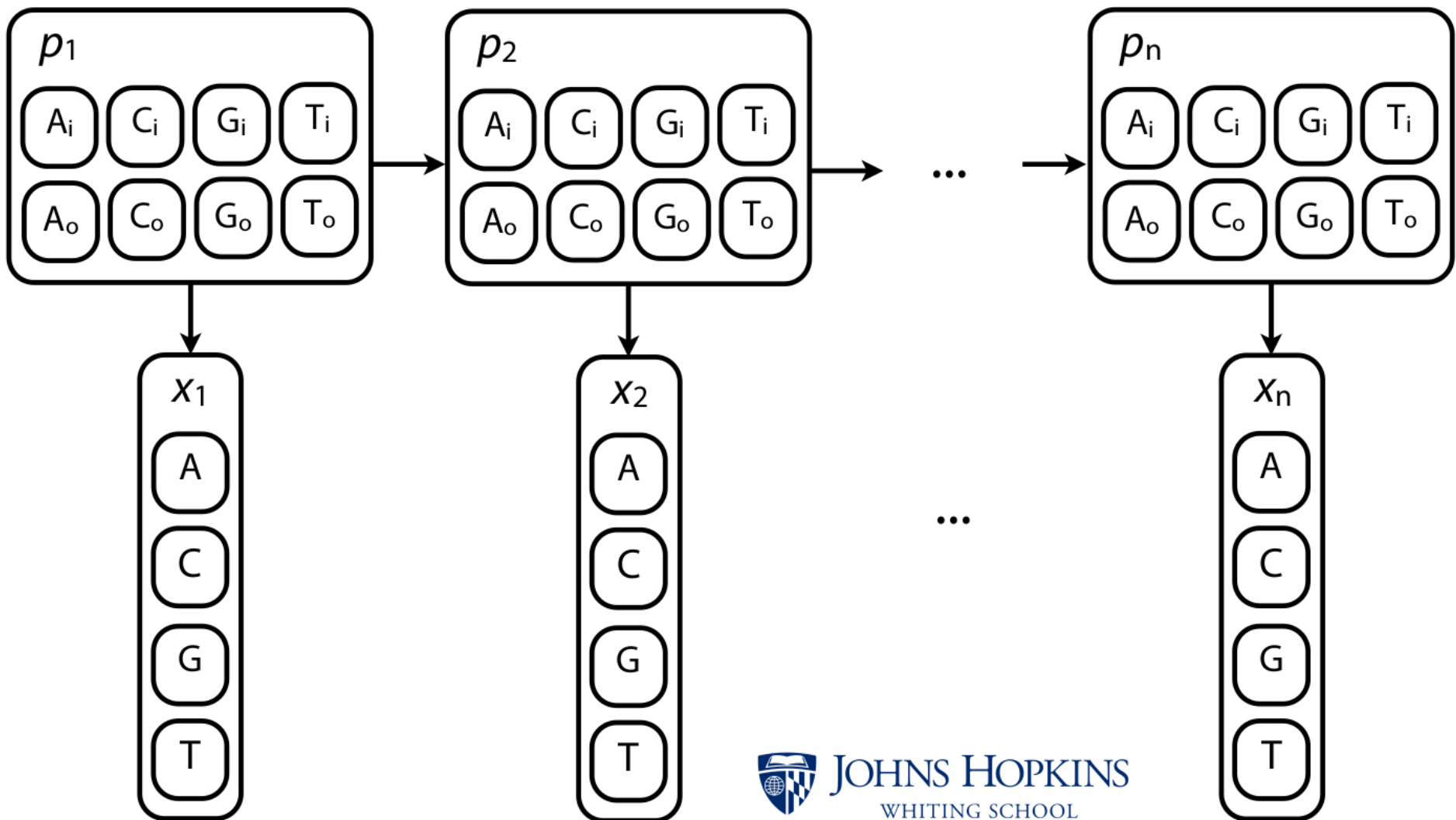


$$Q = \{I, O\}$$

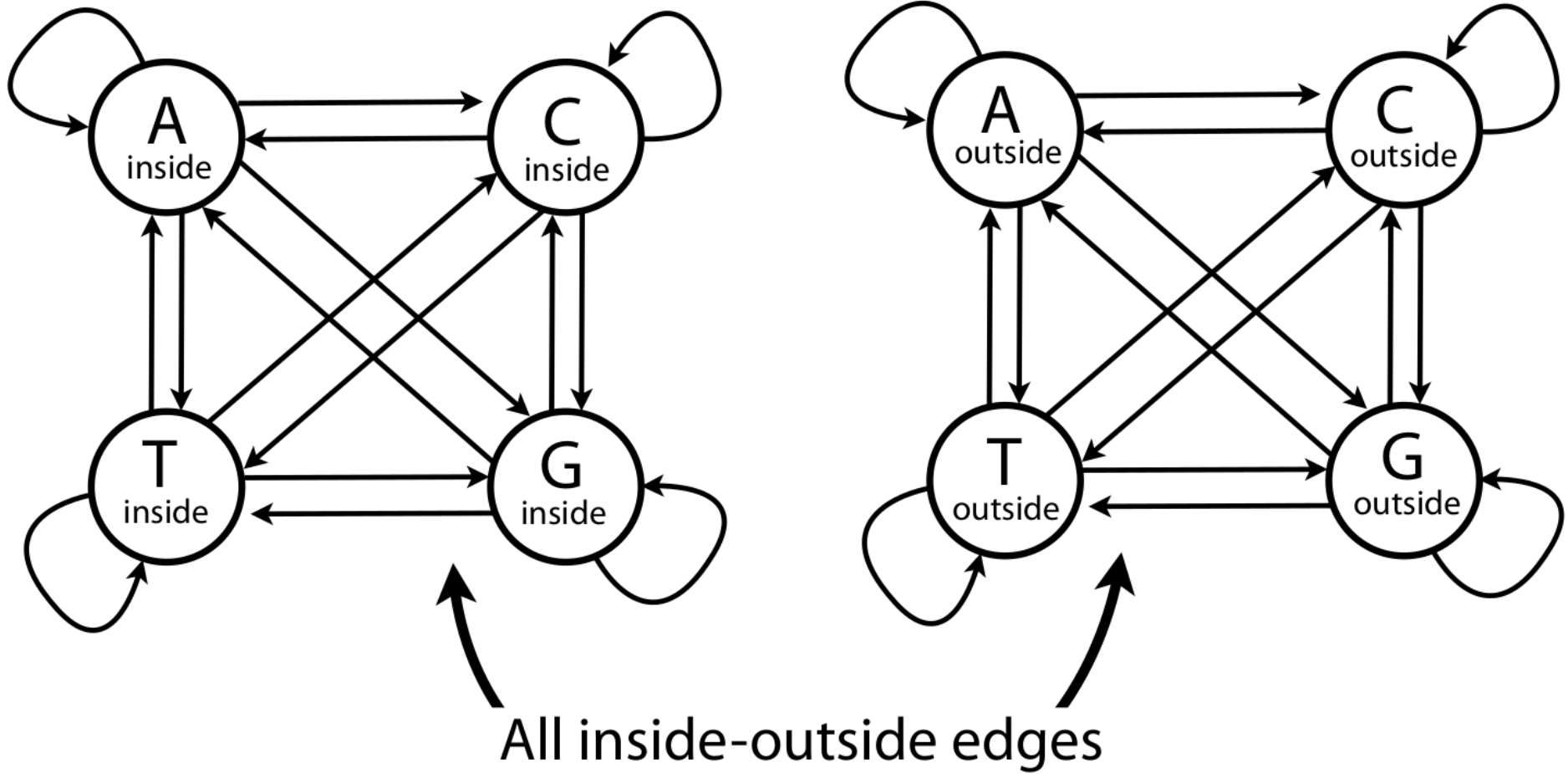
$$Q = \{I, O\} \times \{A, C, G, T\}$$

# Second HMM try

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



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

<b>A</b>	$A_i$	$C_i$	$G_i$	$T_i$	$A_o$	$C_o$	$G_o$	$T_o$
$A_i$								
$C_i$								
$G_i$								
$T_i$								
$A_o$								
$C_o$								
$G_o$								
$T_o$								



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

<b>E</b>	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

A	IA	IG	IC	IT	OA	OG	OC	OT
IA	Orange	Red	Red	Orange	White	White	White	White
IG	Orange	Red	Red	Orange	Yellow	Yellow	Yellow	Yellow
IC	Orange	Red	Red	Orange	White	White	White	White
IT	Orange	Red	Red	Orange	White	White	White	White
OA	White	White	Yellow	White	Orange	Orange	Orange	Orange
OG	White	White	Yellow	White	Orange	Orange	Orange	Orange
OC	White	White	Yellow	White	Orange	Yellow	Orange	Orange
OT	White	White	Yellow	White	Orange	Orange	Orange	Orange

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

↖ CpG islands end with a CG not just G

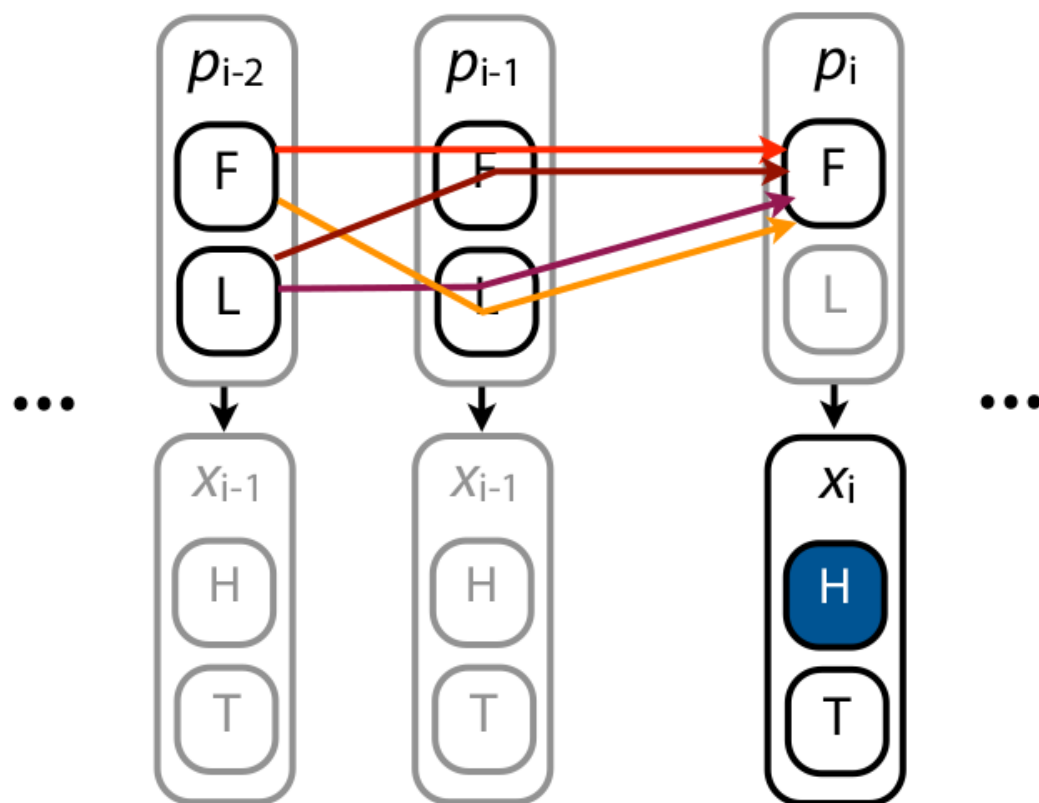
**Positive**:  
High GC content inside  
CG rare outside

**Negative**:  
End on G start on C

↖ CpG islands start with a CG not just C

# Higher order HMMs

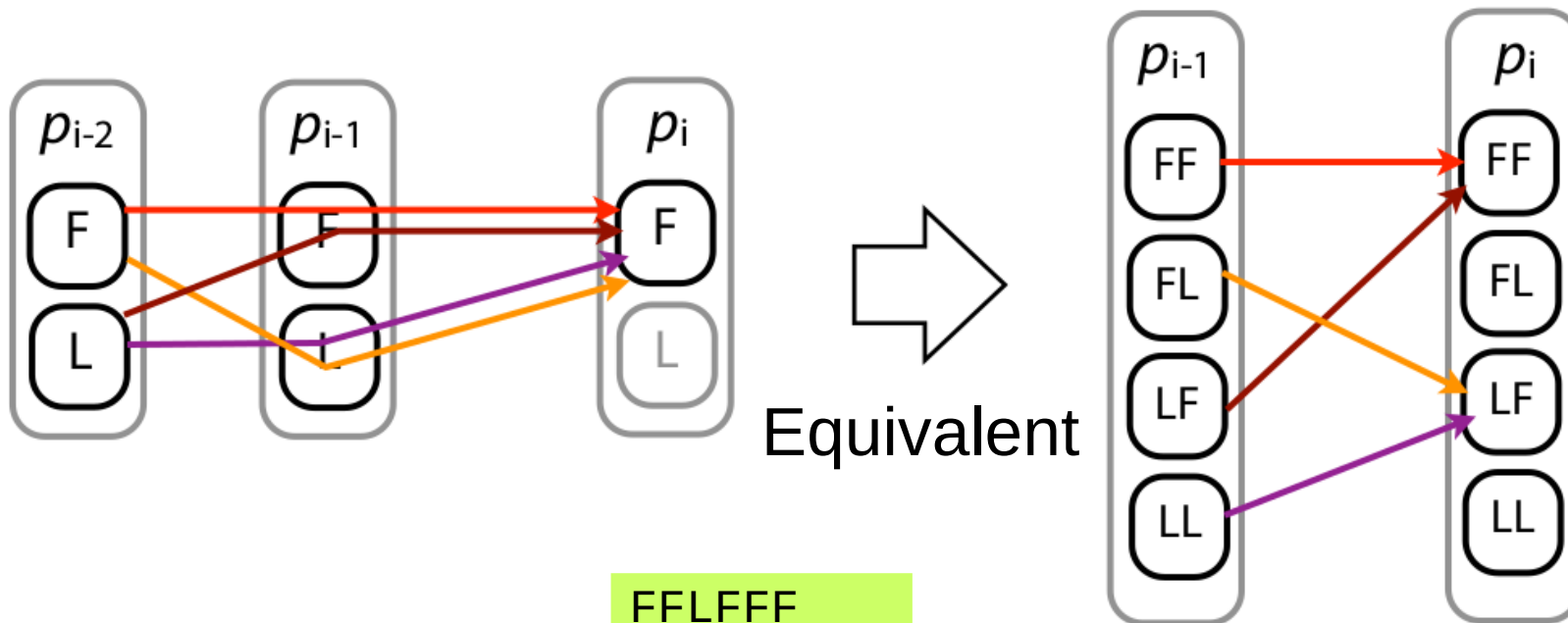
For higher-order HMMs, Viterbi  $\mathbf{s}_{k,i}$  no longer depends on just the previous state assignment





# 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  $x$ ?

$$P(p_1, \dots, p_n, x_1, \dots, x_n)$$

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

$$p^* = \operatorname{argmax} P(p_1, \dots, p_n | x_1, \dots, x_n)$$

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

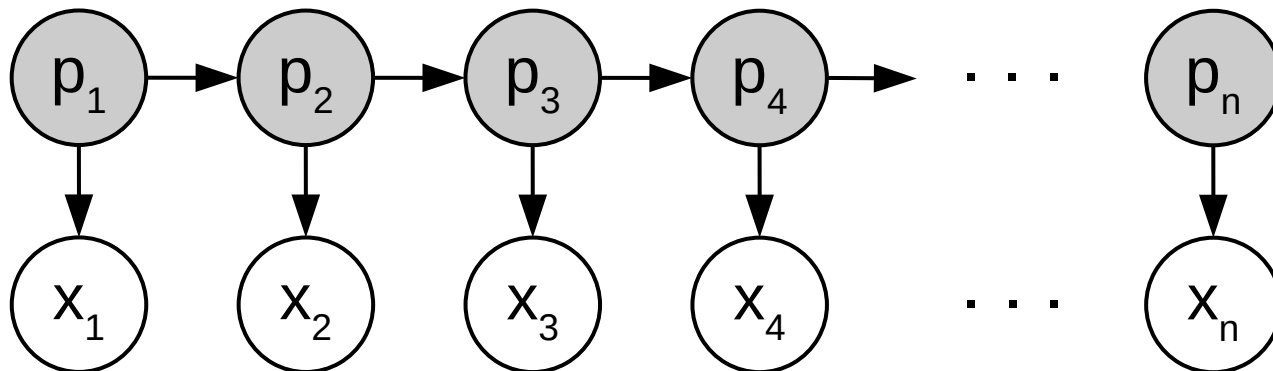
$$P(p_i = t, x_1, \dots, x_i) - \text{forward algorithm}$$

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

$$P(x_{i+1} \dots x_n | p_i = t) - \text{backward algorithm}$$

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

$$P(p_i | x_1, \dots, x_n) -- \text{forward and backward algorithm}$$



Baum Welch

# Forward algorithm

What is the probability that  $p_i = \text{state } t$

$P(p_i=t, x_1 \dots x_i)$  – 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  $x_i \dots x_n$  given  $p_i = t$ ?

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

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

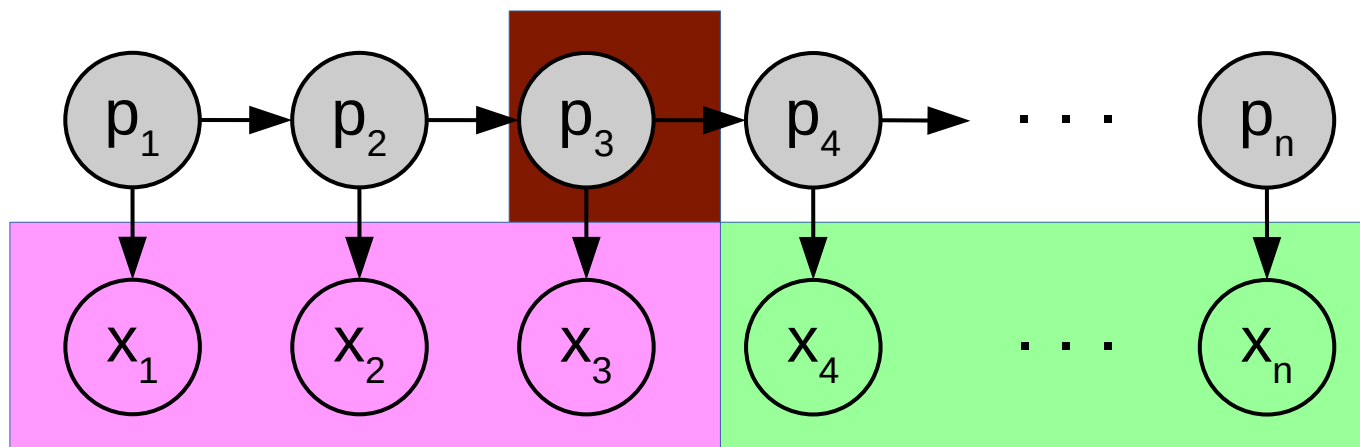
Emission

Recursion  
Probability

Transition

Sum instead of max (viterbi)

# Forward-Backward Algorithm



Forward  $P(p_3, x_1 \dots x_3)$     Backward  $P(x_4 \dots x_n | p_3)$

Forward/backward

$P(p_3 | x_1 \dots x_n)$  proportional to  $P(p_3, x_1 \dots x_n)$

$$\begin{aligned} P(p_3, x_1 \dots x_n) &= P(p_3, x_1 \dots x_3) P(x_4 \dots x_n | p_3) \\ &= F(t, i) B(t, i) \end{aligned}$$

What is the conditional probability of hidden state  $p$  at site  $i$   
 $P(p_i | x)$  -- 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

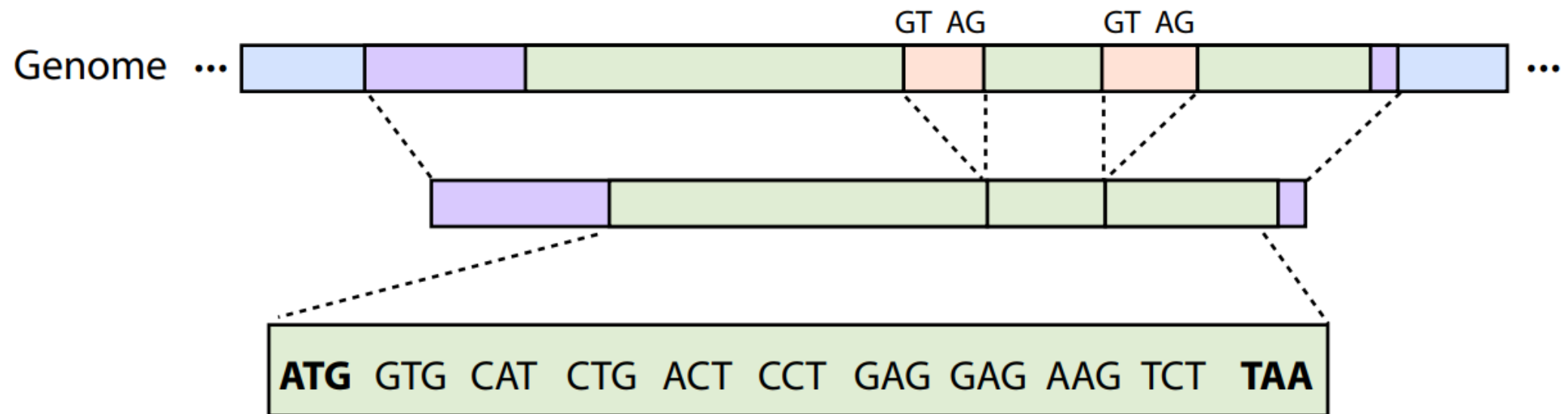
# Eukaryotic genes: a challenge

ATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACCTCCTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTC  
ATCACTTAGACCTCACCTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGG  
GCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTTACATTTGCTTCTGACACAACCTGTGTTCACTAGCAACCTCAAA  
CAGACACC**ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTT**  
**GGTGGTGAGGCCCTGGGCAG**GTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGA  
CAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAG**GCTGCTGGTG**  
**GTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGT**  
**GAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCA**  
**CACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGG**GTGAGTCTATGGGACGCTTGATGTTTT  
CTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAG  
ACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTTCTTTTATTTGCTGTTTATAACAATTGTTTTCTTTT  
GTTTAATTCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAA  
AGGAAATATCTCTGAGATACATTAAGTAACCTAAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAATAT  
ATGTGTGCTTATTTGCATTTATGGGTTAAAGTGTAATTTACATAATCATTATACATATTTGCATTTGTAATTTTAAAA  
AATGCTTTCTTCTTTTAATATACTTTTTTGTATTCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTTCAAGGGCAATAA  
TGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCT  
CTGCATATAAATATTTCTGCATATAAATTGTAACCTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTA  
CCATTCTGCTTTTATTTTATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCA  
TACCTCTTATCTTCTCCACAG**CTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACC**  
**CCACCAGTGCAAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCACAAGTATCACTAAGCTCGCTT**  
TCTTGCTGTCCAATTTCTATTAAAGGTTCTTTGTTCCCTAAGTCCAACCTACTAACTGGGGGATATTATGAAGGGCCTT  
GAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAATGATGTATTAAATTATTTCTGAATATTTTACTA  
AAAAGGGAATGTGGGAGGTCAGTGCATTTAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGGGAAAATACACTATA

Homo sapiens hemoglobin, beta (HBB)

# 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:

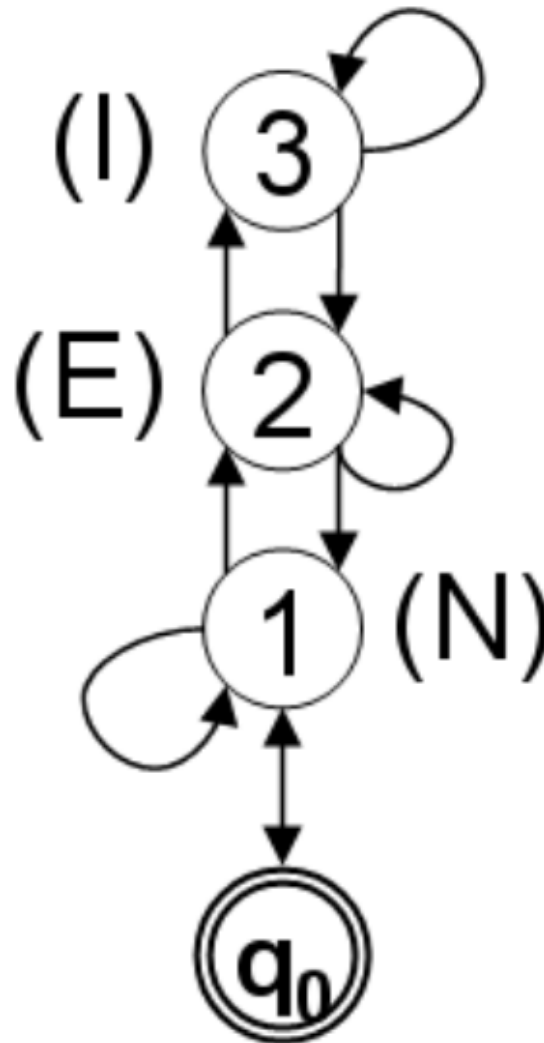
Emissions are nucleotides

I = intron

E = exon

N = intergenic  
(between genes)

$q_0$  is a *start state*;  
guarantees we start in  
the N (intergenic) state



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*

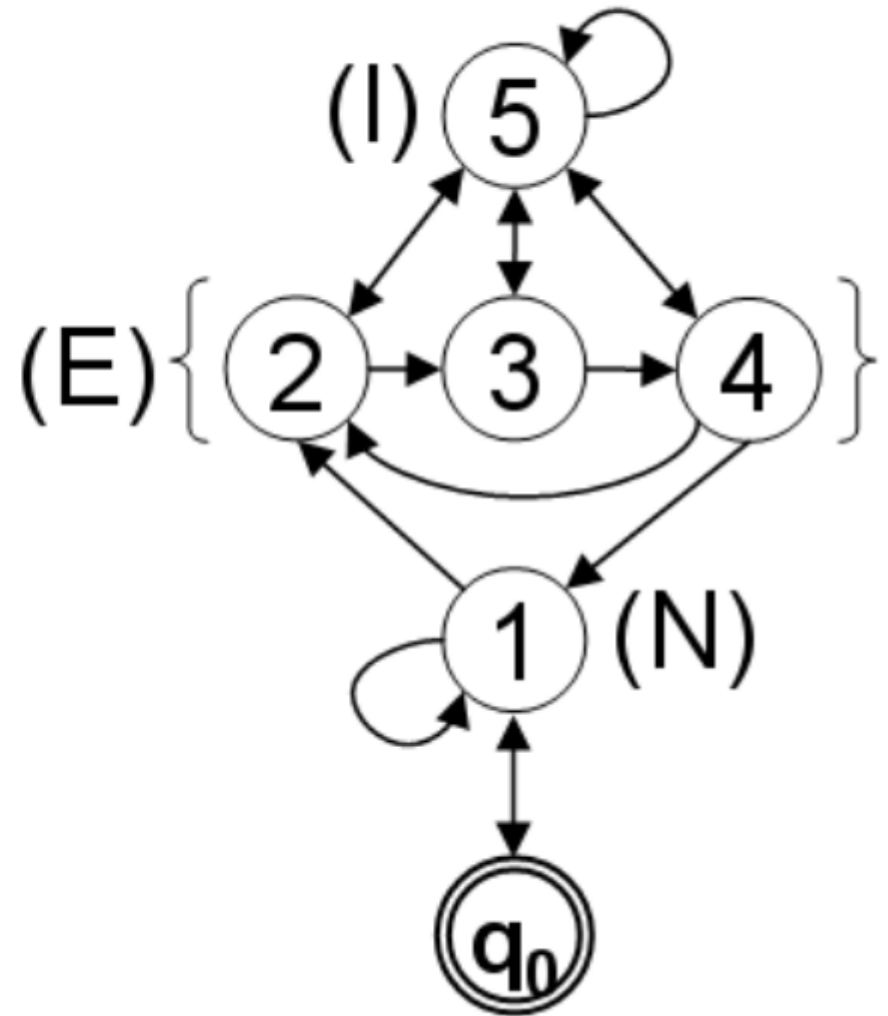
Problems

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

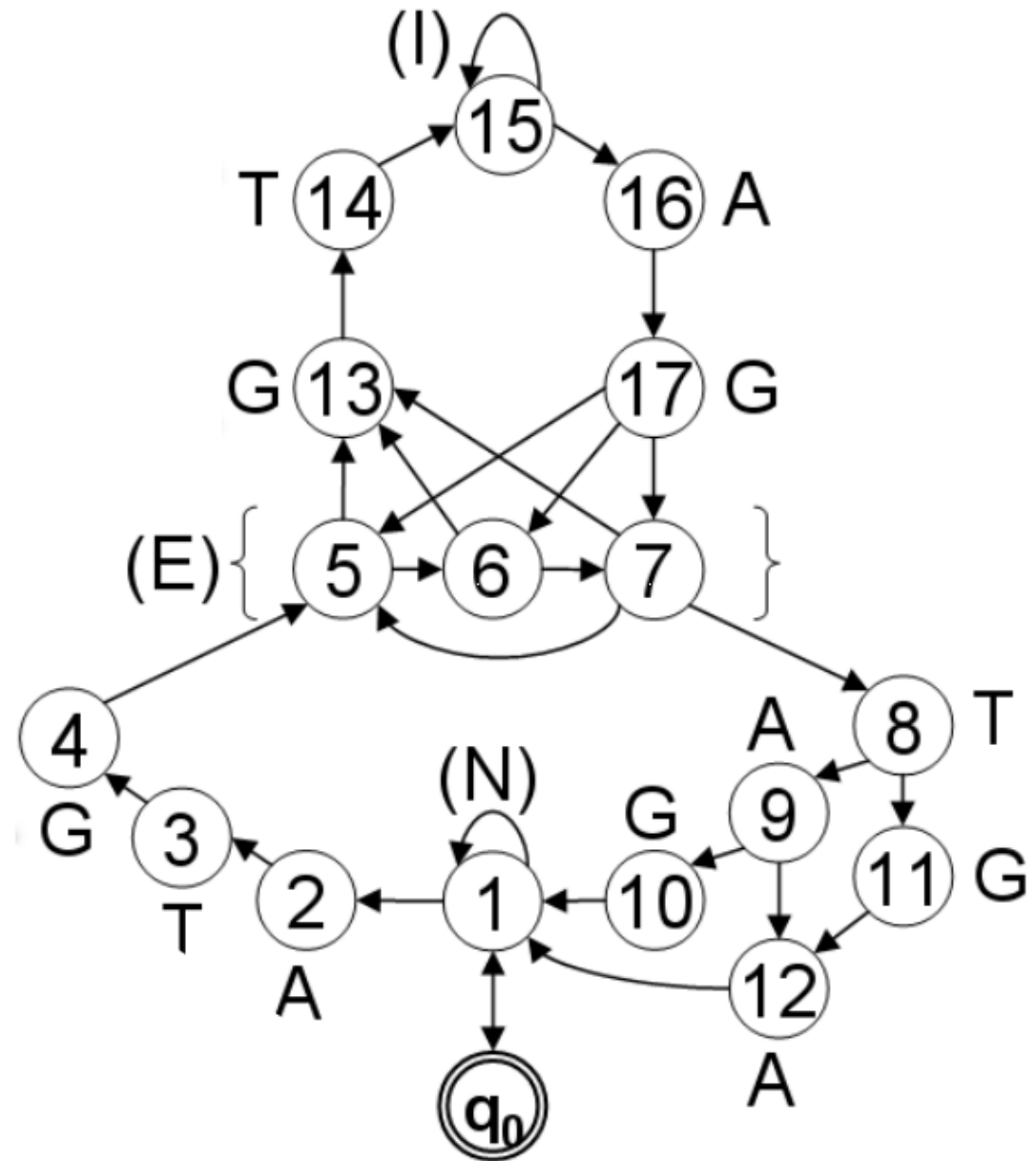
8-12 capture stop codons

Why not arrow from 5 to 8?

Codon would be 1TGA|  
1TAG|1TAA – not a stop

Do exons end 2<sup>nd</sup> position in  
coding and start at 3<sup>rd</sup>?

Not always

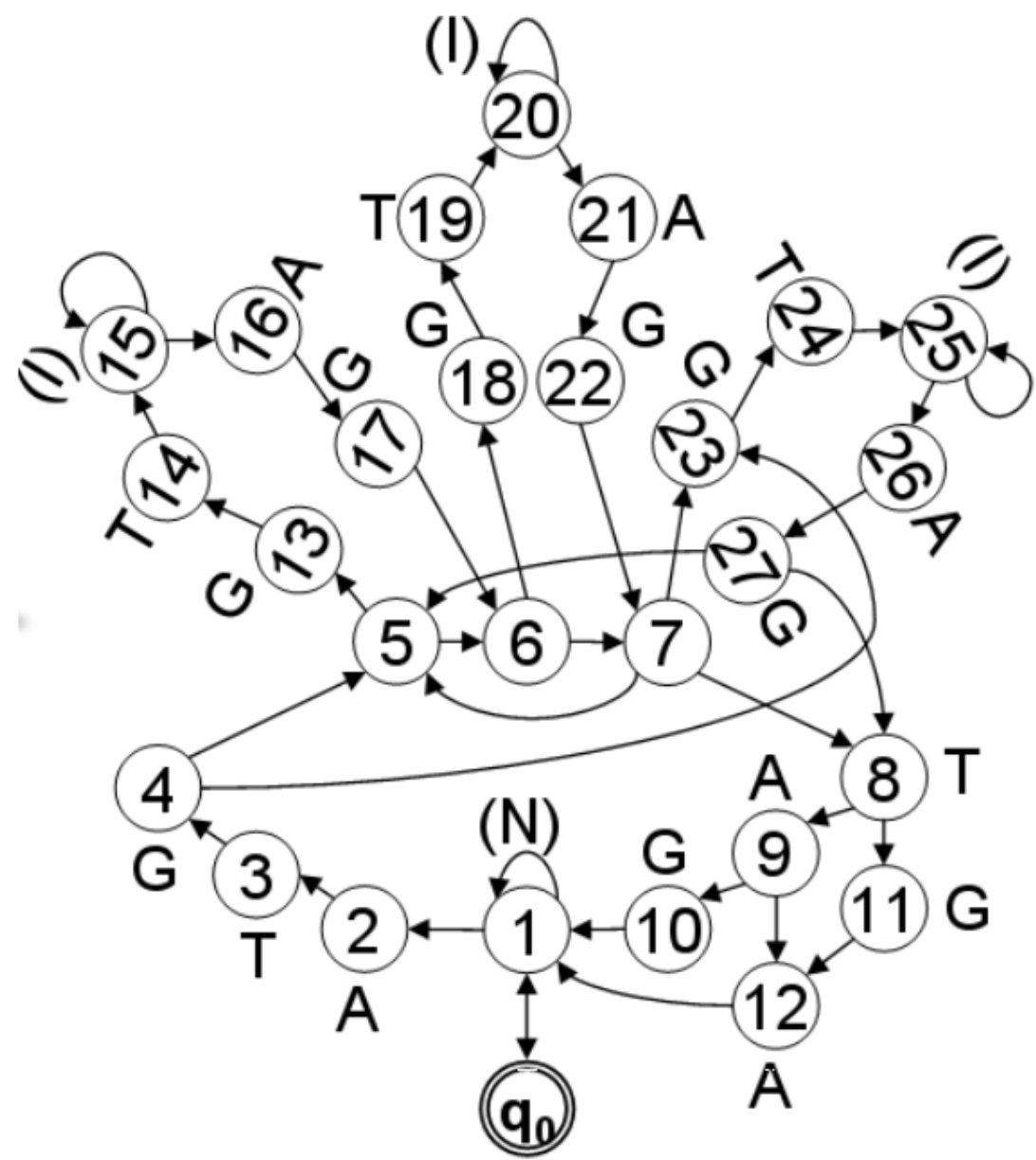


Attempt 4:

Additional copies of the acceptor/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  $\rightarrow$  729 transition probs

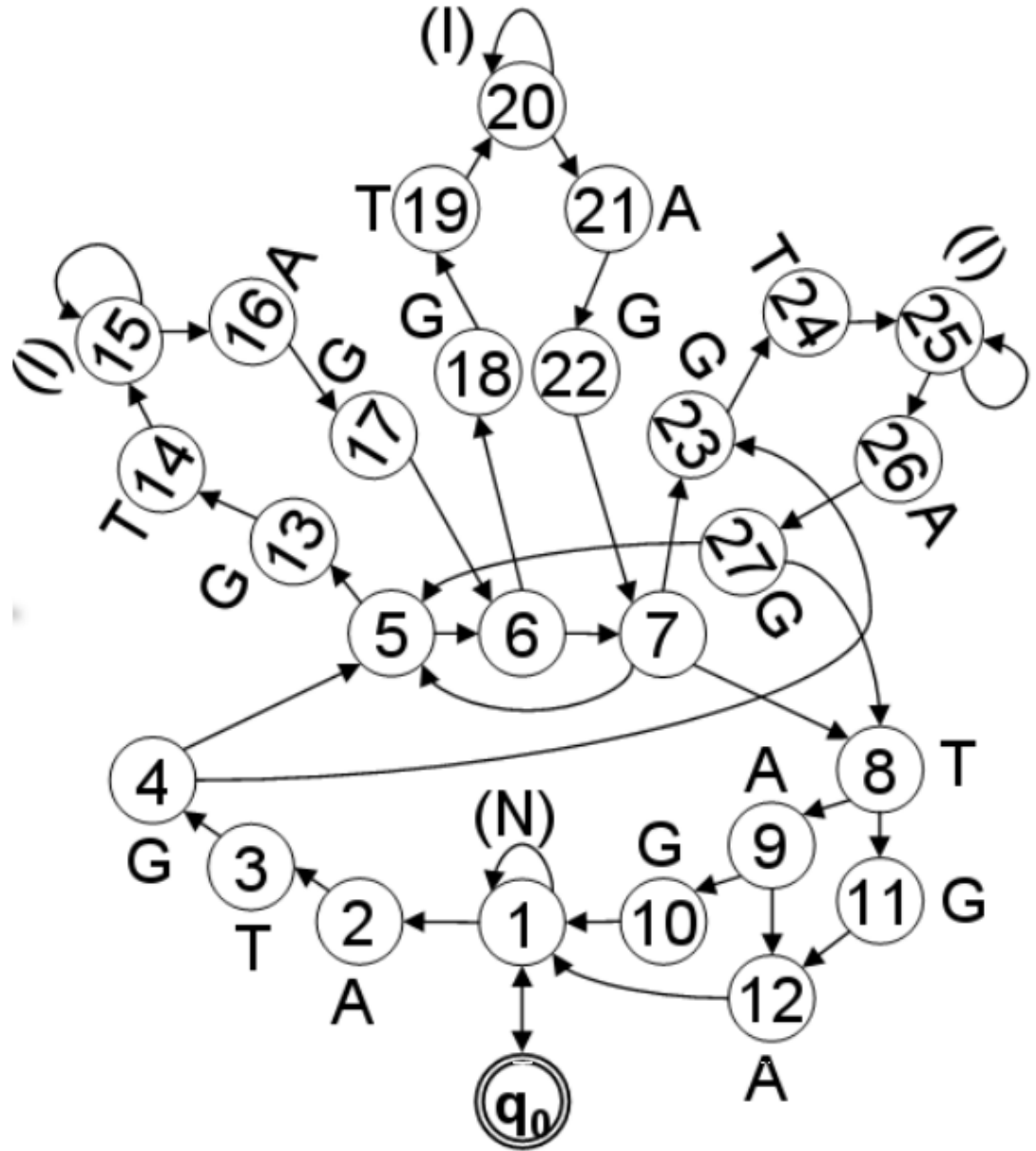


# How is codon bias incorporated?

## States 5, 6, 7

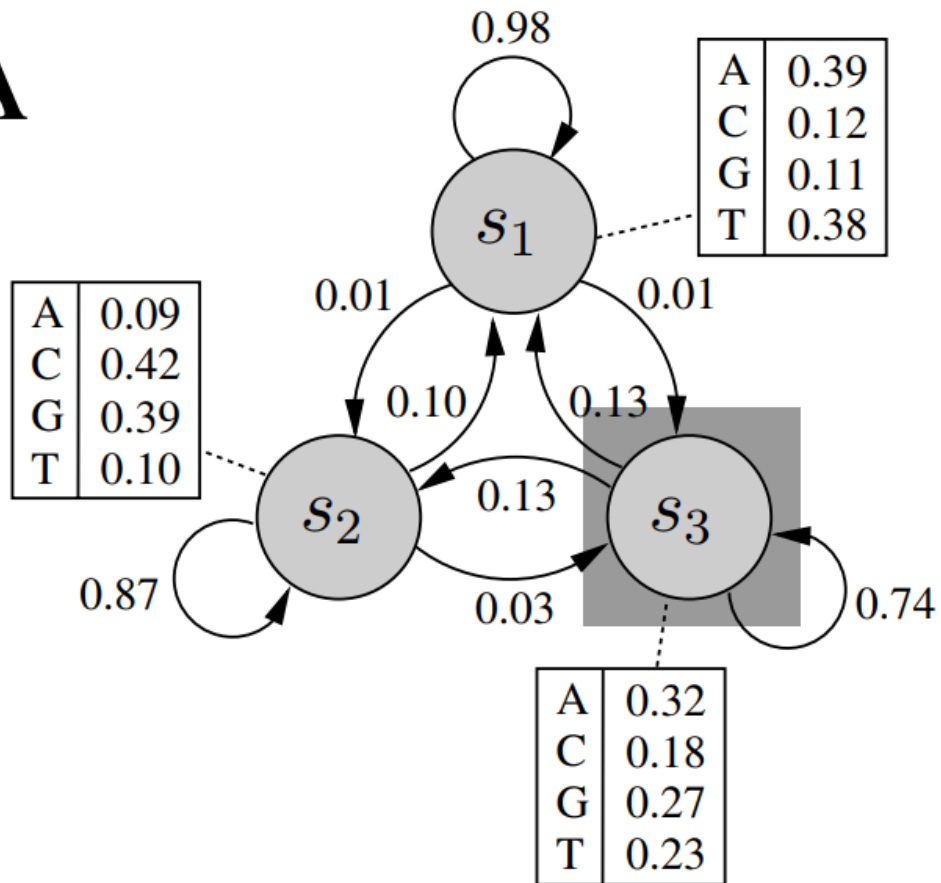
7 = third position

7 emits GC (bias) more than AT



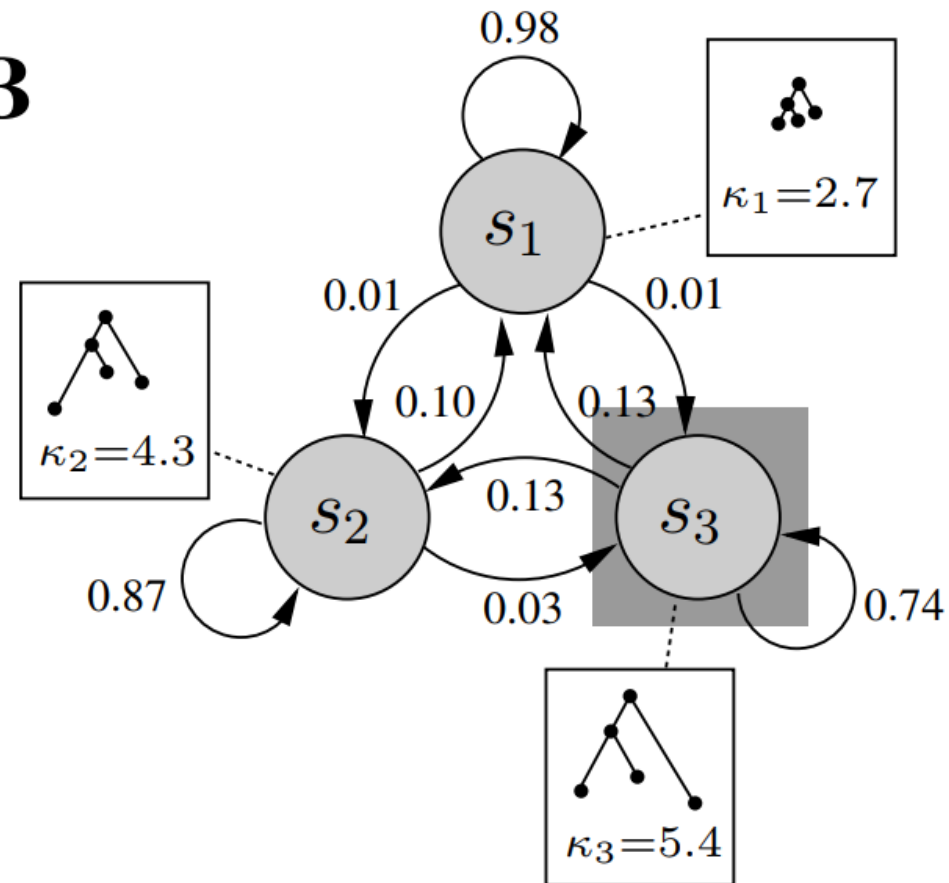
# Phylogenetic Hidden Markov Model: Markov chains in time and space

**A**



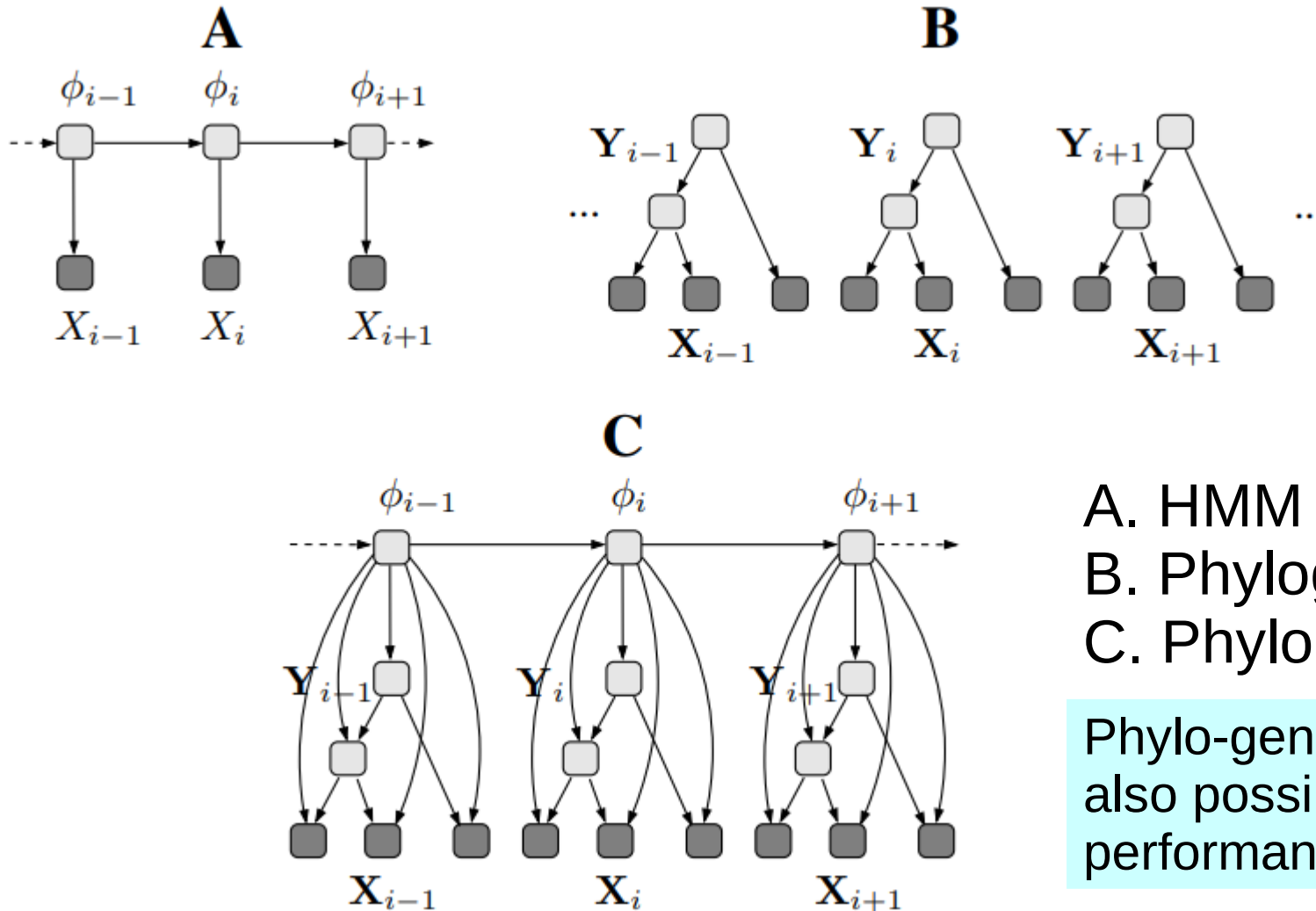
$\mathbf{X} = \text{TAACGGCAGA} \dots$

**B**



$\mathbf{X} = \begin{matrix} \text{TAACGGCAGA} \dots \\ \text{TTAGGCAAGG} \dots \\ \text{AAGGCGCCGA} \dots \end{matrix} \dots$

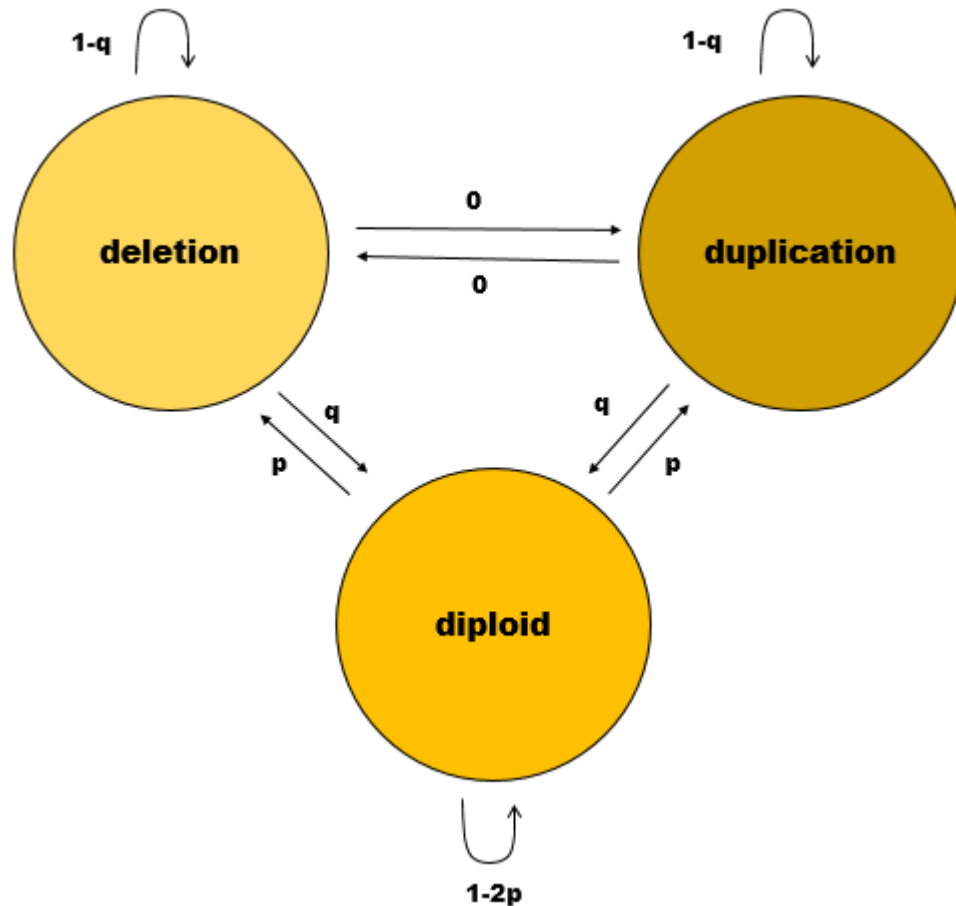
# HMM in time and space



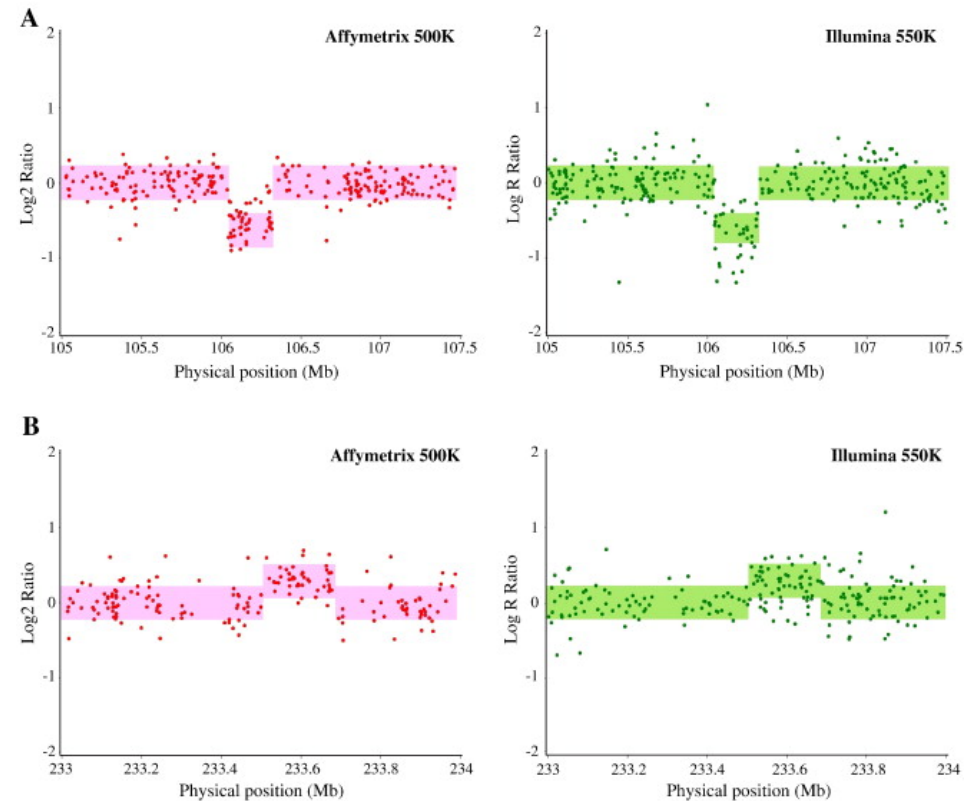
- A. HMM
- B. Phylogenetic model
- C. Phylo-HMM

Phylo-gene finding HMM is also possible and improves performance

# Copy number variation



Read depth | Intensity



Physical position on chromosome

Emissions: read counts or log2(ratio) is discretized

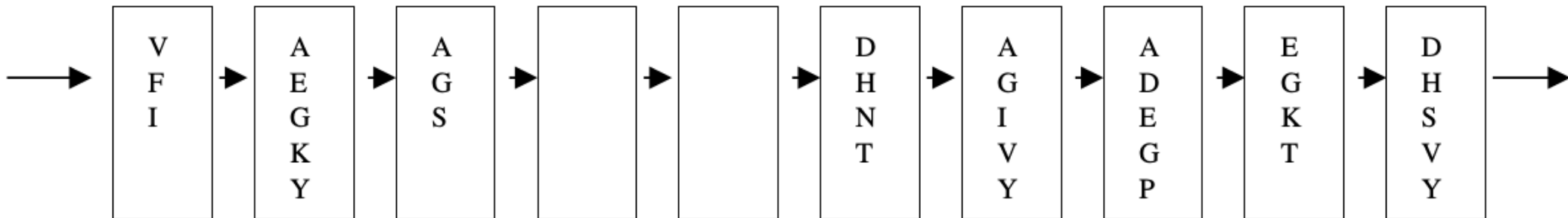


# Profile HMMs

**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":    ***  *****
```

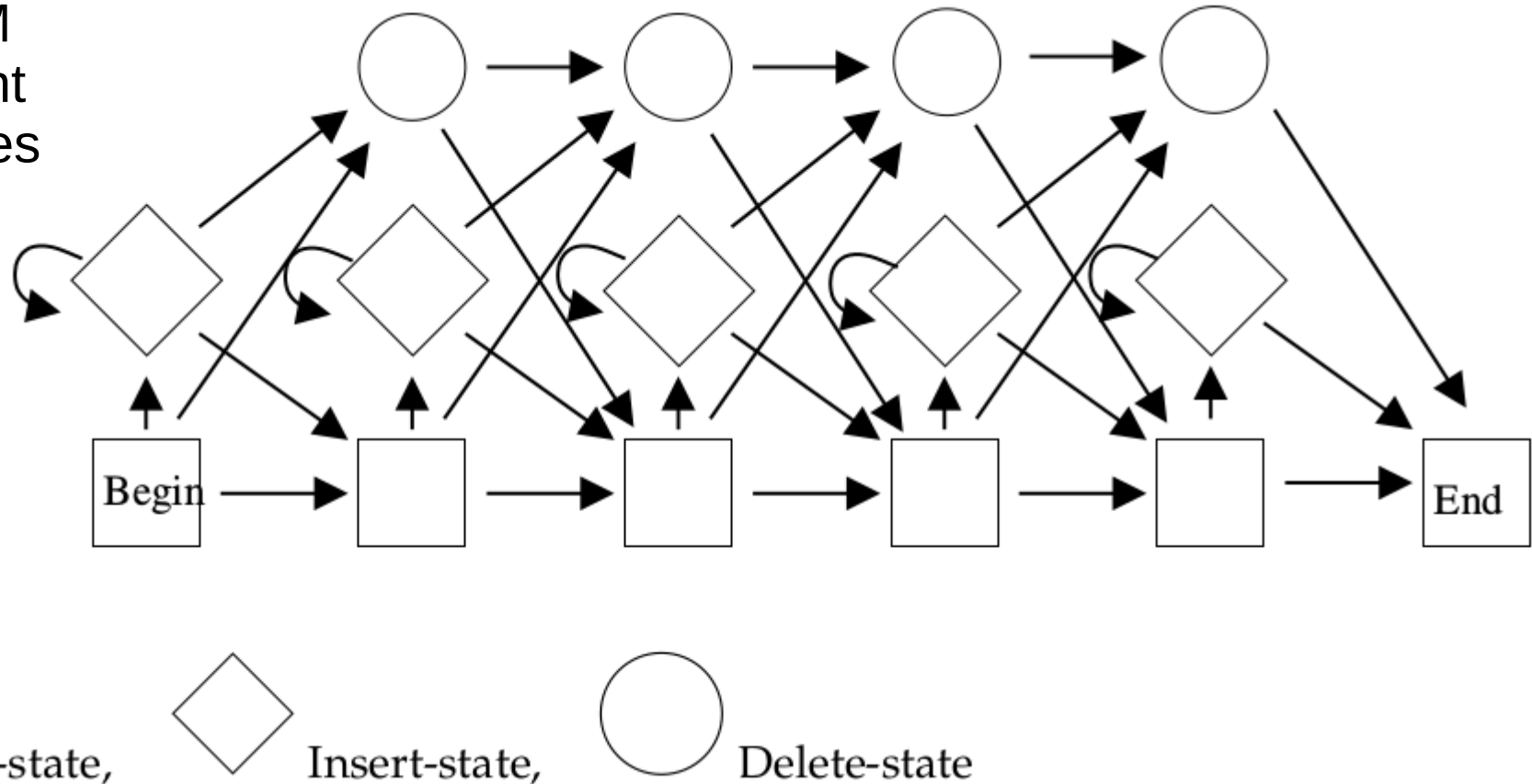
- Improve genome annotations
- Better remote homology
- Better handle variable domain architectures



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

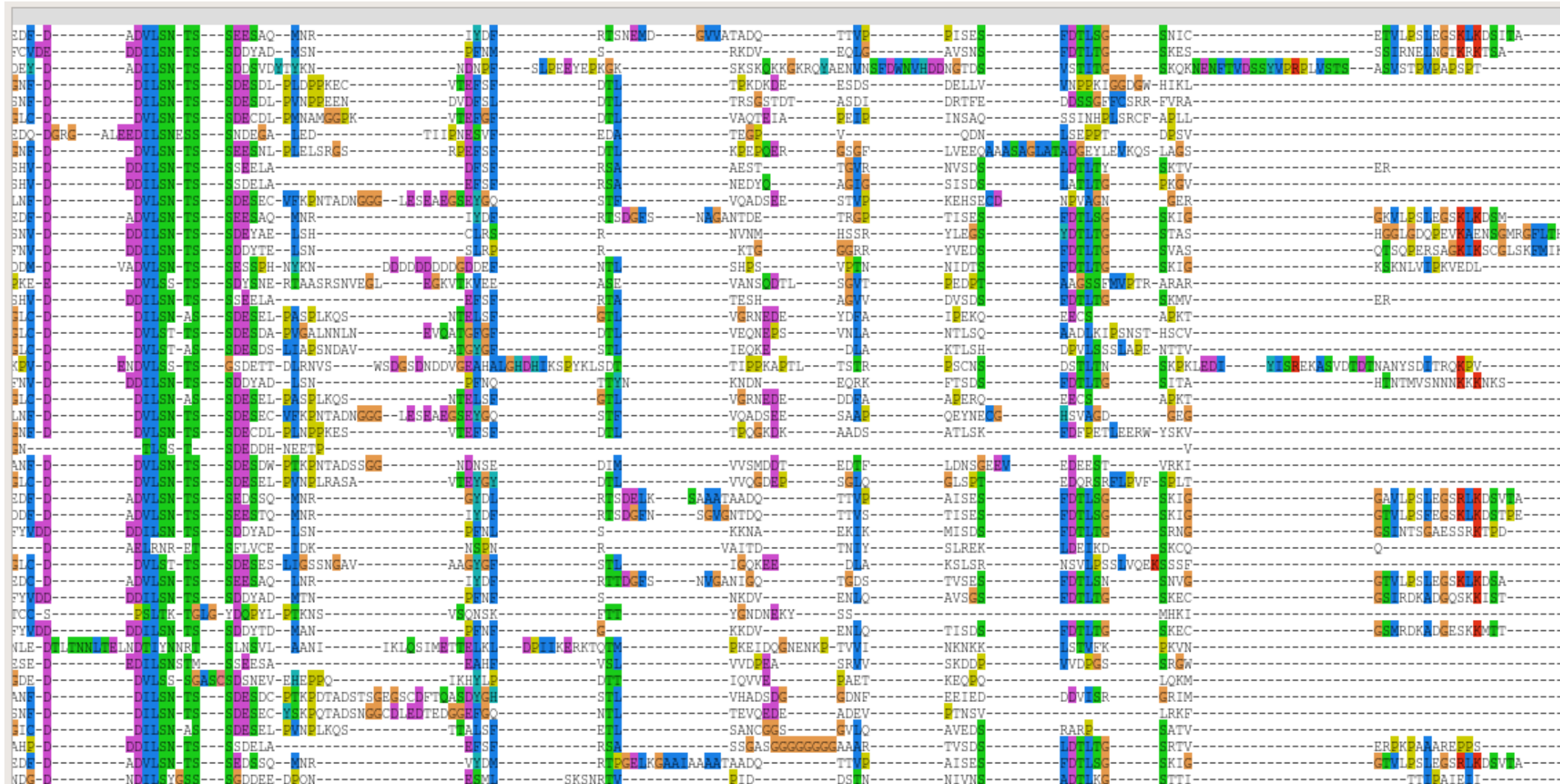
# Profile HMMs

- Make close alignment
- Train HMM
- Find distant homologues

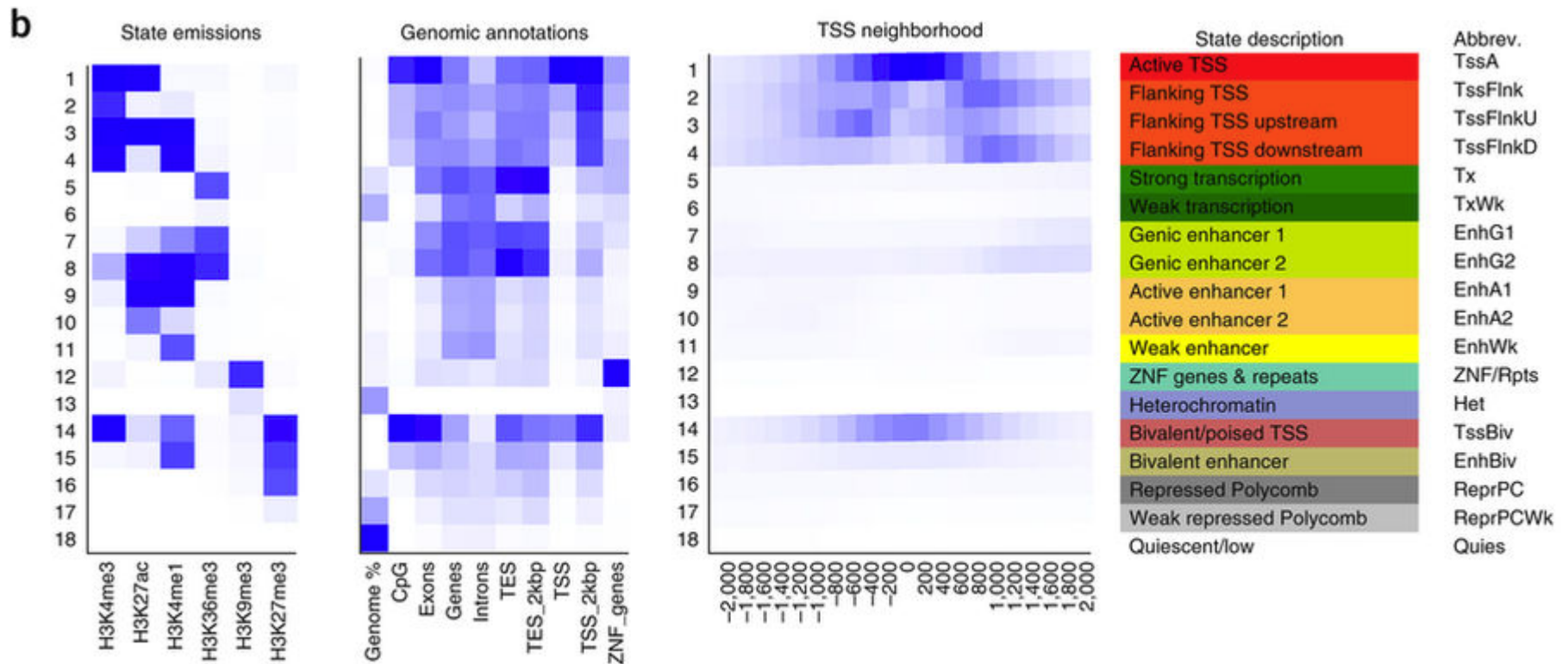
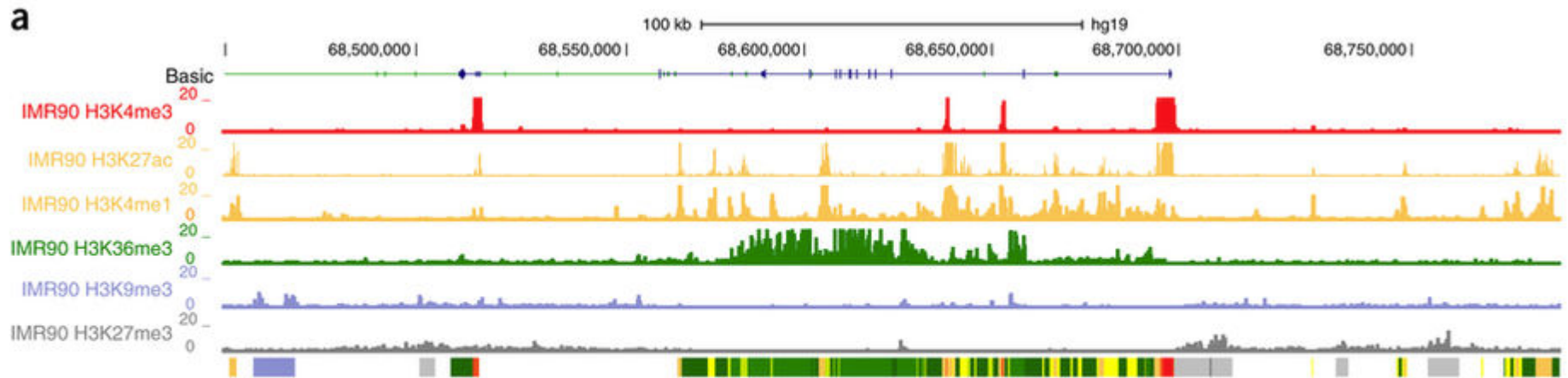


Delete-states are silent and do not have emissions  
Insert and delete states account for aligned columns with gaps  
Insert/delete if more than 50% sequences gapped.

# Example of Hmmer alignment



# Chromatin-HMM



# Exercises

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

E	A	G	C	T	A	I	O
I					I		
O					O		

Sequence: AAAAAATCGGGATAT

Labels: 00000IIIIIIOIOI

- 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	Orange	Orange					Orange	Orange
IG	Orange	Orange					Orange	Orange
IC								
IT								
OA	Orange	Orange					Orange	Orange
OG	Orange	Orange					Orange	Orange
OC	Orange	Orange			Orange	Orange	Orange	Orange
OT	Orange	Orange			Orange	Orange	Orange	Orange

# Exercises

- 1) What are pseudocounts used to avoid in the EM algorithm?
- 2) Why does this model not work well in identifying CpG islands:

<b>A</b>	<b>I</b>	<b>O</b>	<b>E</b>	<b>A</b>	<b>C</b>	<b>G</b>	<b>T</b>
I	0.8	0.2	I	0.1	0.4	0.4	0.1
O	0.2	0.8	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:         00000000000000IIIIIIIIII0000000000000000