Problem 1:

- Input: Small sequence S
- Output: Is S from a CpG island?
  - Build Markov models: M+ and M —
  - Then compare

Markov Models

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

—	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

How to distinguish?

- Compute

$$S(x) = \log\left(\frac{P(x \mid M+)}{P(x \mid M-)}\right) = \sum_{i=1}^{L} \log\left(\frac{p_{x(i-1)x_i}}{m_{x(i-1)x_i}}\right) = \sum_{i=1}^{L} r_{x(i-1)x_i}$$

r=p/m	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.573	0.393	-0.679

Score(`GCAC`)
= .461-.913+.419
< 0.
`GCAC` not from CpG island.

Score(`GCTC`)
= .461-.685+.573
> 0.
`GCTC` from CpG island.

Problem 1:

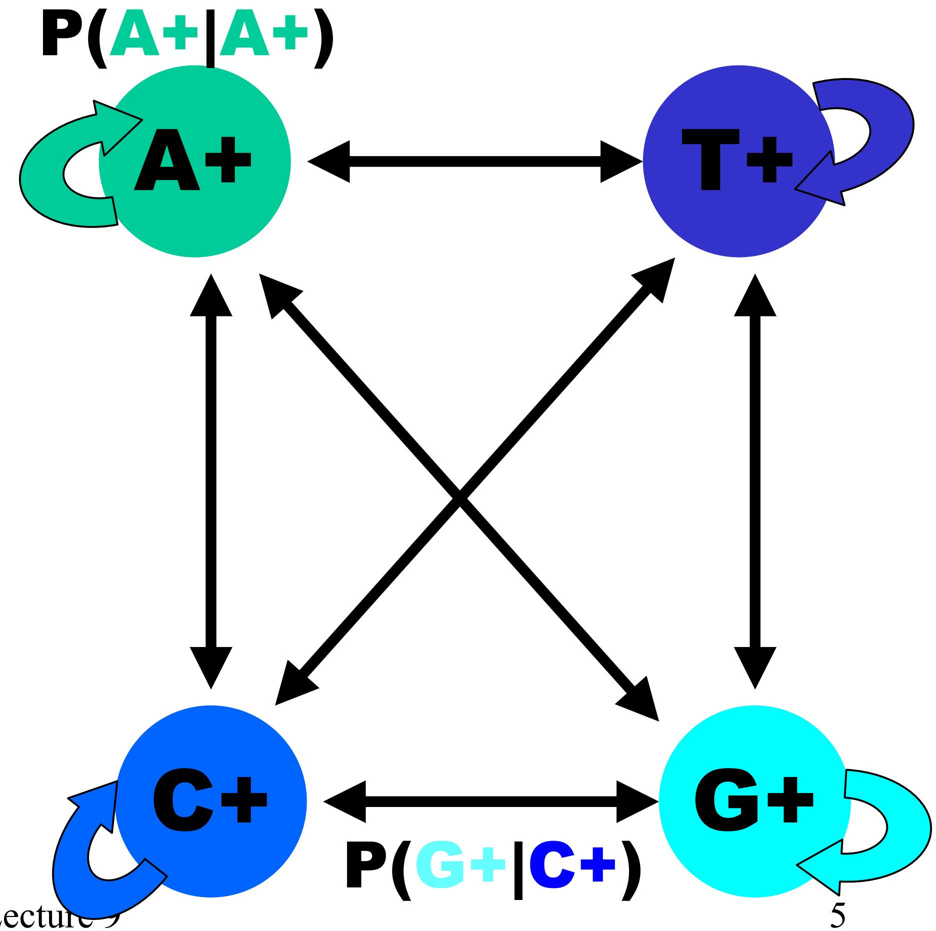
- Input: Small sequence S
- Output: Is S from a CpG island?
  - Build Markov Models: M+ & M-
  - Then compare

Problem 2:

- Input: Long sequence S
- Output: Identify the CpG islands in S.
  - Markov models are inadequate.
  - Need Hidden Markov Models.

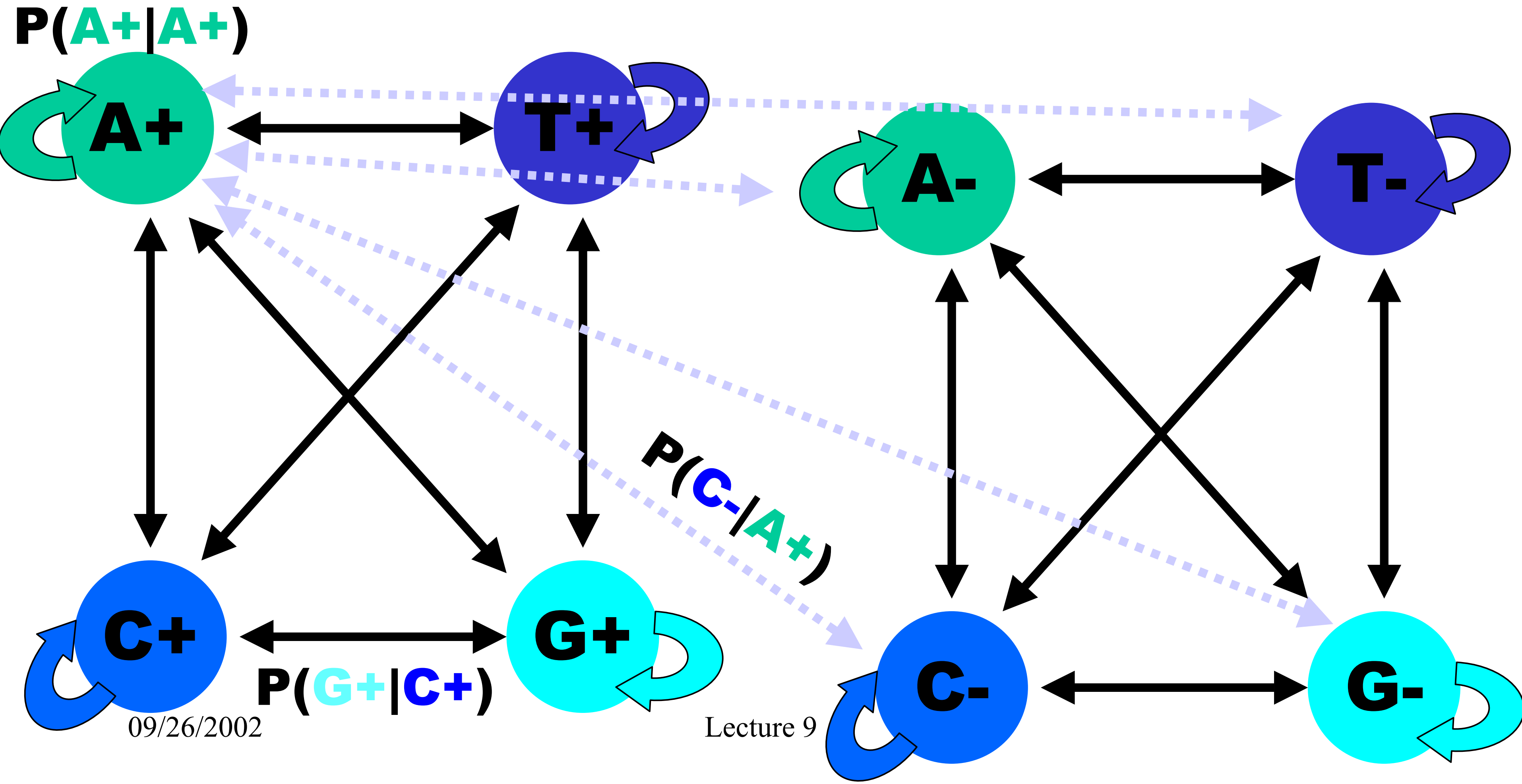
Markov Models

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182



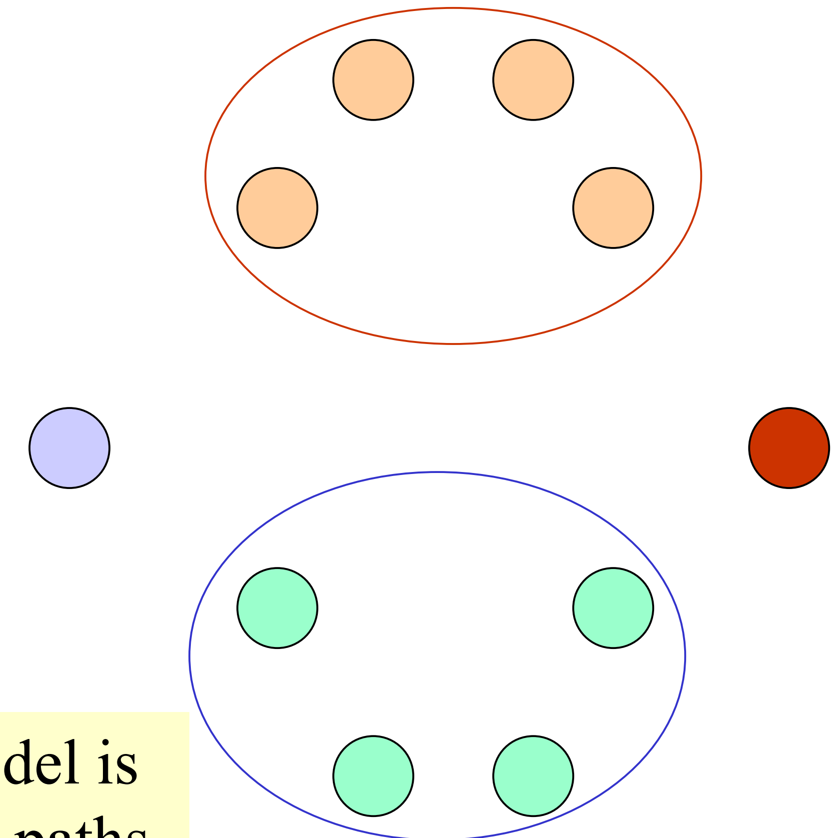
CpG Island + in an ocean of –
First order Hidden Markov Model

MM=16, HMM= 64 transition probabilities (adjacent bp)



Hidden Markov Model (HMM)

- States
- Transitions
- Transition Probabilities
- Emissions
- Emission Probabilities



- What is hidden about HMMs?

Answer: The path through the model is hidden since there are many valid paths.

How to Solve Problem 2?

- Solve the following problem:

Input: Hidden Markov Model M,

parameters Θ, emitted sequence S

Output: Most Probable Path Π

How: Viterbi's Algorithm (Dynamic Programming)

Define $\Pi[i,j]$ = MPP for first j characters of S ending in state i

Define $P[i,j]$ = Probability of $\Pi[i,j]$

- Compute state i with largest $P[i,j]$.

Problem 3: LIKELIHOOD QUESTION

- Input: Sequence S, model M, state i
- Output: Compute the probability of reaching state i with sequence S using model M
  - Backward Algorithm (DP)

Problem 4: LIKELIHOOD QUESTION

- Input: Sequence S, model M
- Output: Compute the probability that S was emitted by model M
  - Forward Algorithm (DP)

Problem 5: <u>LEARNING QUESTION</u>

- Input: model structure M, Training Sequence S
- Output: Compute the parameters Θ
- Criteria: ML criterion
  - maximize $P(S \mid M, \Theta)$ HOW???

Problem 6: <u>DESIGN QUESTION</u>

- Input: Training Sequence S
- Output: Choose model structure M, and compute the parameters Θ
  - No reasonable solution
  - Standard models to pick from

Iterative Solution to the LEARNING QUESTION (Problem 5)

- Pick initial values for parameters Θ_0
- <u>Repeat</u>

    Run training set *S* on model M

    Count # of times transition i $\Rightarrow$ j is made

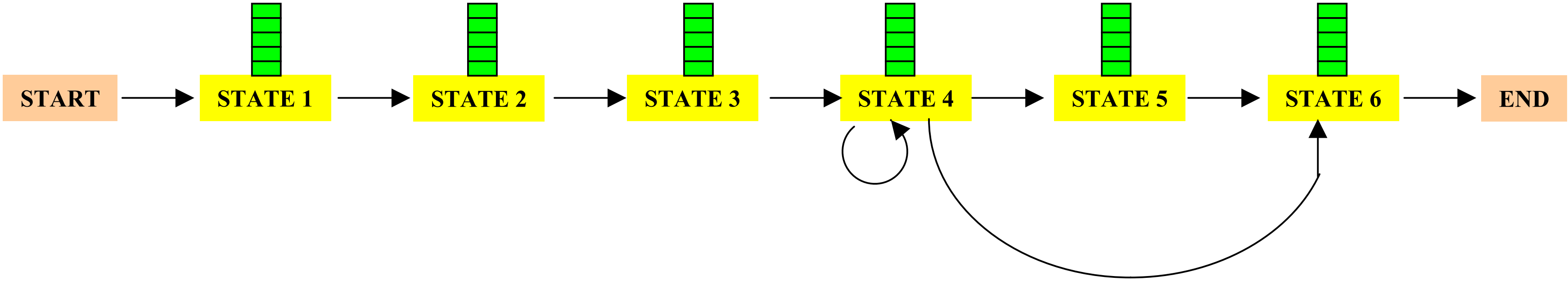
    Count # of times letter x is emitted from state i

    Update parameters Θ

- <u>Until</u> (some stopping condition)

Simple Models

- Helps to model simple sequence features.
  - single sequences e.g. **TTGACA** or **TATATT** [??]
  - sets of sequences e.g. [**AT**] **C** [**GC**] **TC** [**AGC**]
  - sets of sequences with inserts e.g. **GCA** [**AT**] [**AT**]* **AG**
  - & deletes too, e.g. **TATA** [**G –**] **T**



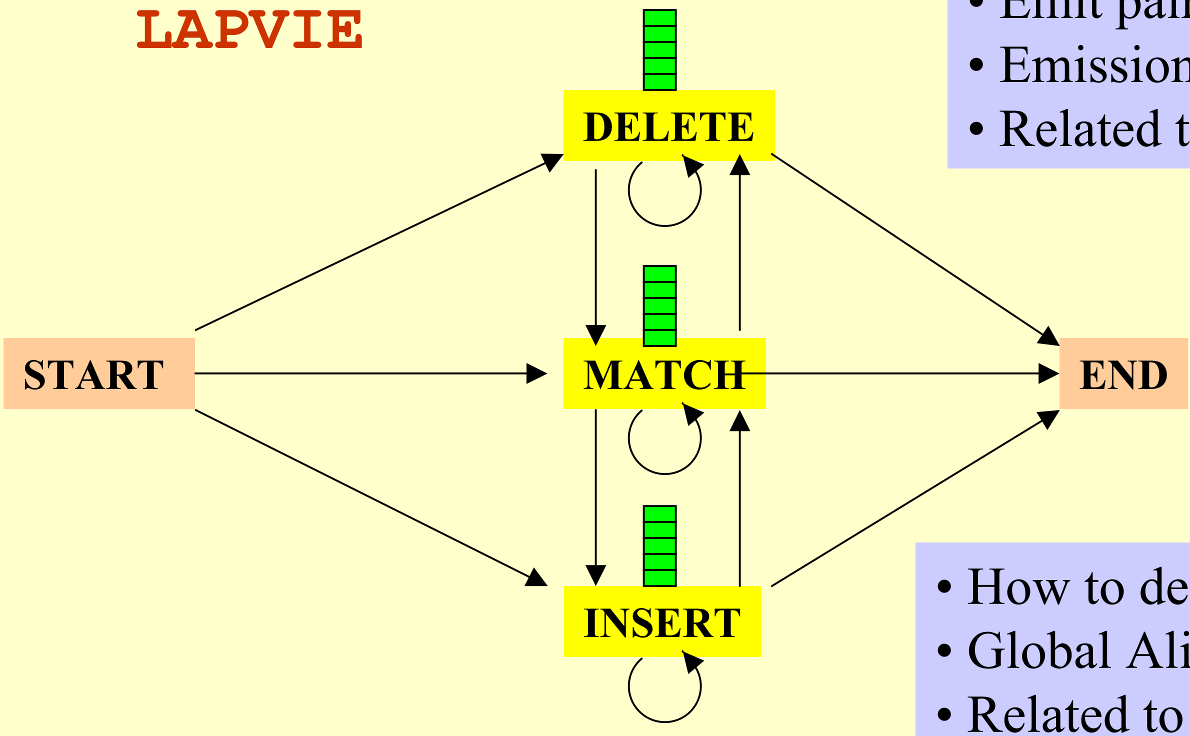
  - long sequences with a sequence of domains **H-B-T-B-H**

How to model Pairwise Sequence Alignment

LEAPVE

LAPVIE

Pair HMMs
- Emit pairs of synbols
- Emission probs?
- Related to Sub. Matrices



- How to deal with InDels?
- Global Alignment? Local?
- Related to Sub. Matrices