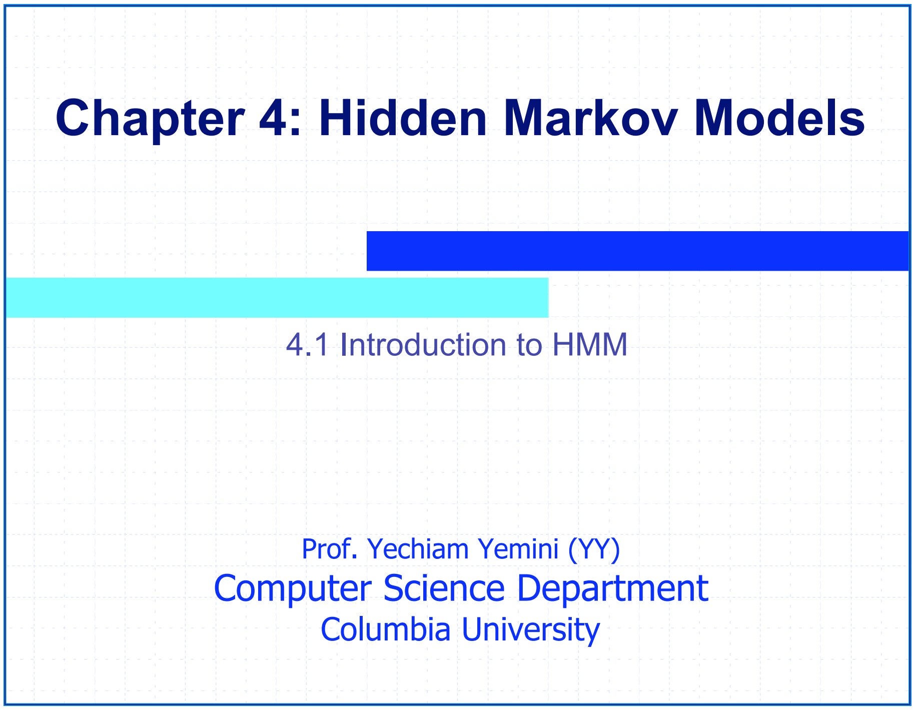
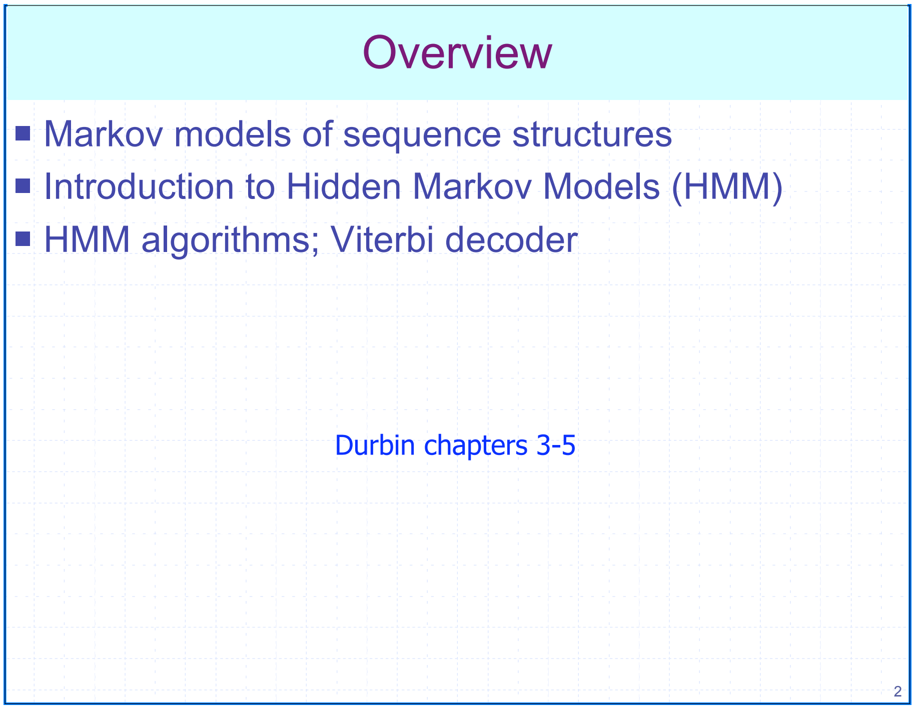


Chapter 4: Hidden Markov Models

4.1 Introduction to HMM

Prof. Yechiam Yemini (YY)
Computer Science Department
Columbia University



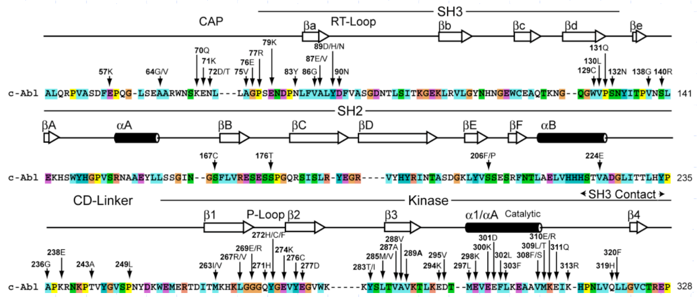
Overview

- Markov models of sequence structures
- Introduction to Hidden Markov Models (HMM)
- HMM algorithms; Viterbi decoder

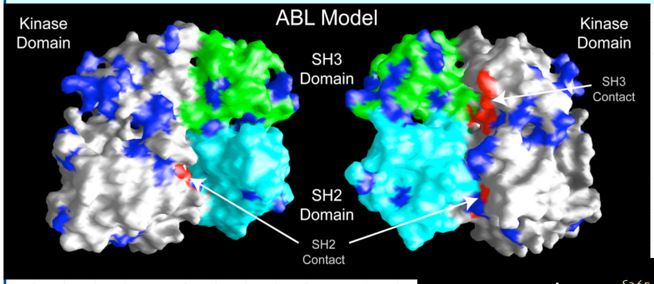
Durbin chapters 3-5

The Challenges

- Biological sequences have modular structure
  - Genes ➔ exons, introns
  - Promoter regions ➔ modules, promoters
  - Proteins ➔ domains, folds, structural parts, active parts
- How do we identify informative regions?
  - How do we find & map genes
  - How do we find & map promoter regions

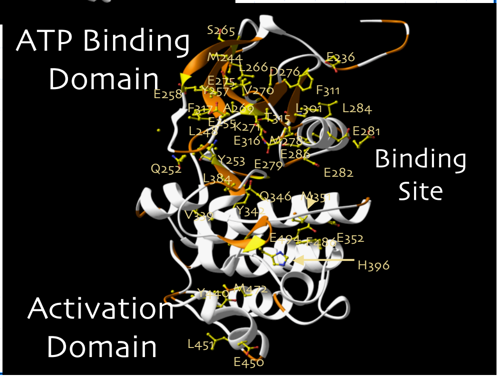


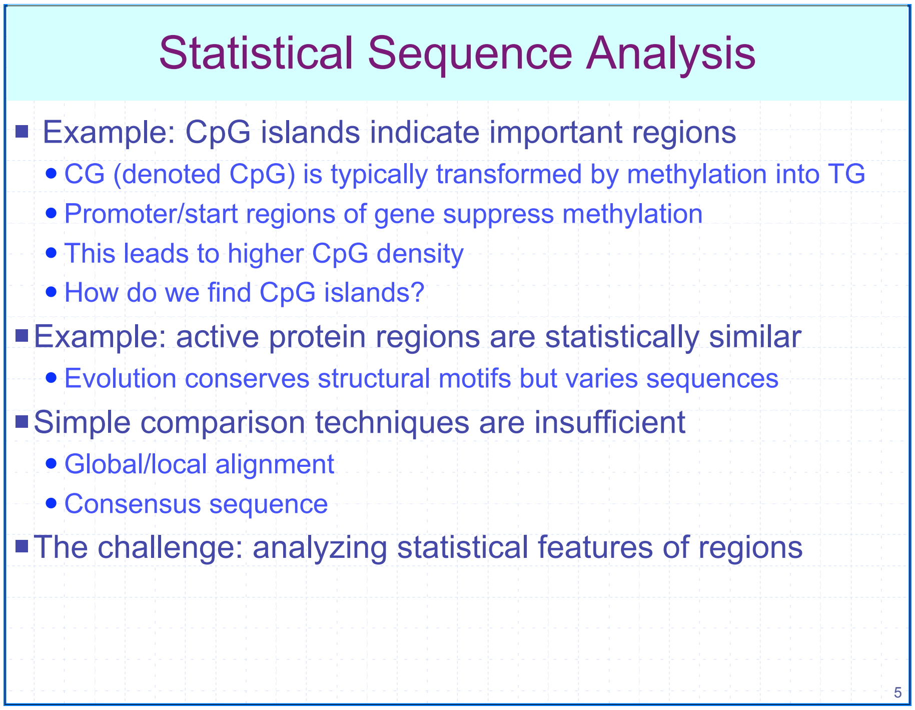
Mapping Protein Regions



Interfaces

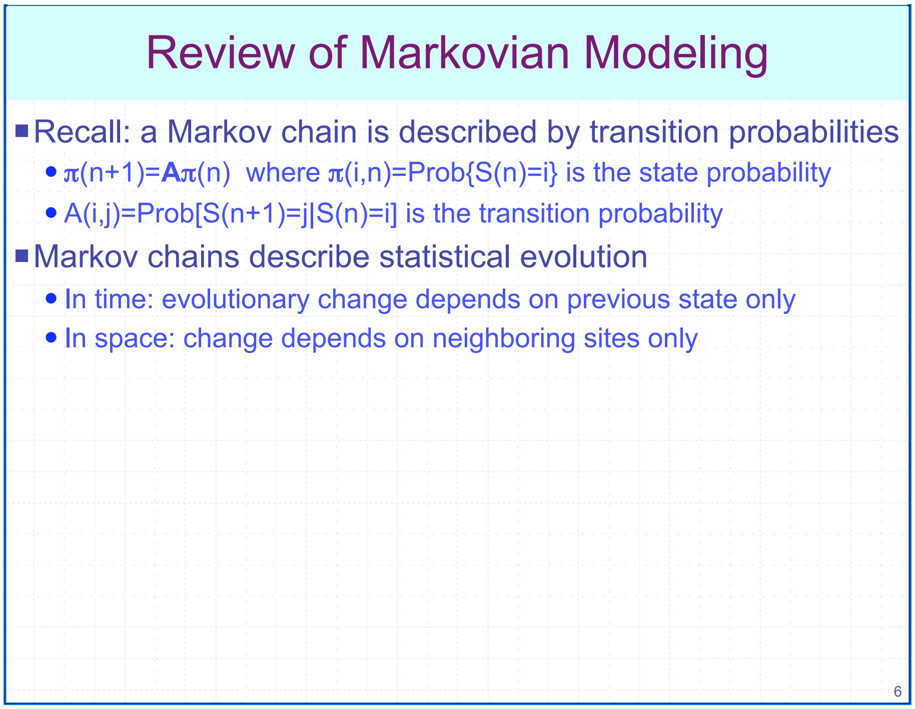
Active components





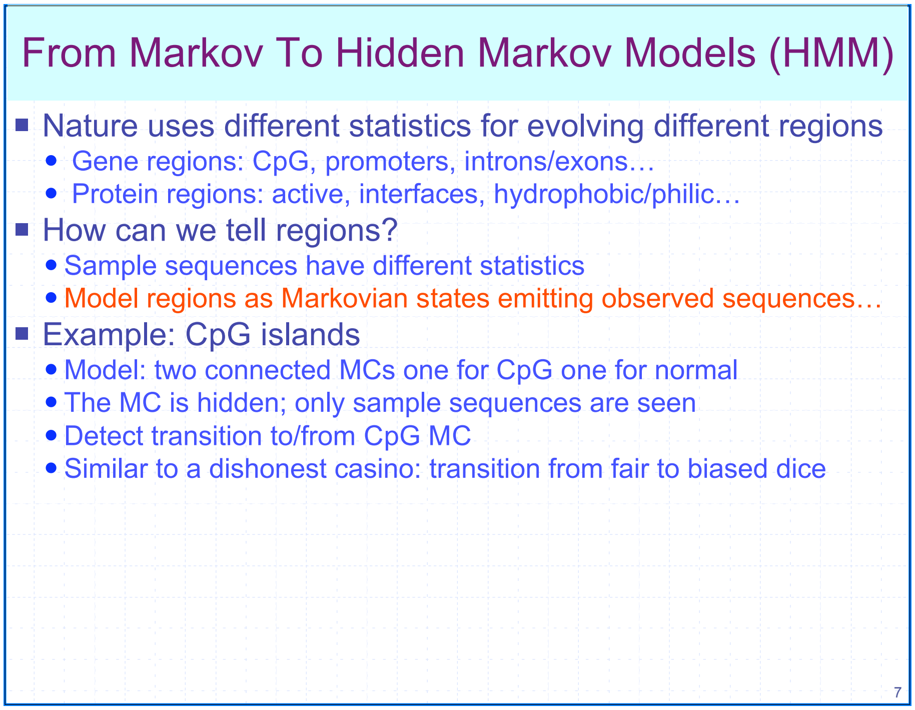
Statistical Sequence Analysis

- Example: CpG islands indicate important regions
  - CG (denoted CpG) is typically transformed by methylation into TG
  - Promoter/start regions of gene suppress methylation
  - This leads to higher CpG density
  - How do we find CpG islands?
- Example: active protein regions are statistically similar
  - Evolution conserves structural motifs but varies sequences
- Simple comparison techniques are insufficient
  - Global/local alignment
  - Consensus sequence
- The challenge: analyzing statistical features of regions



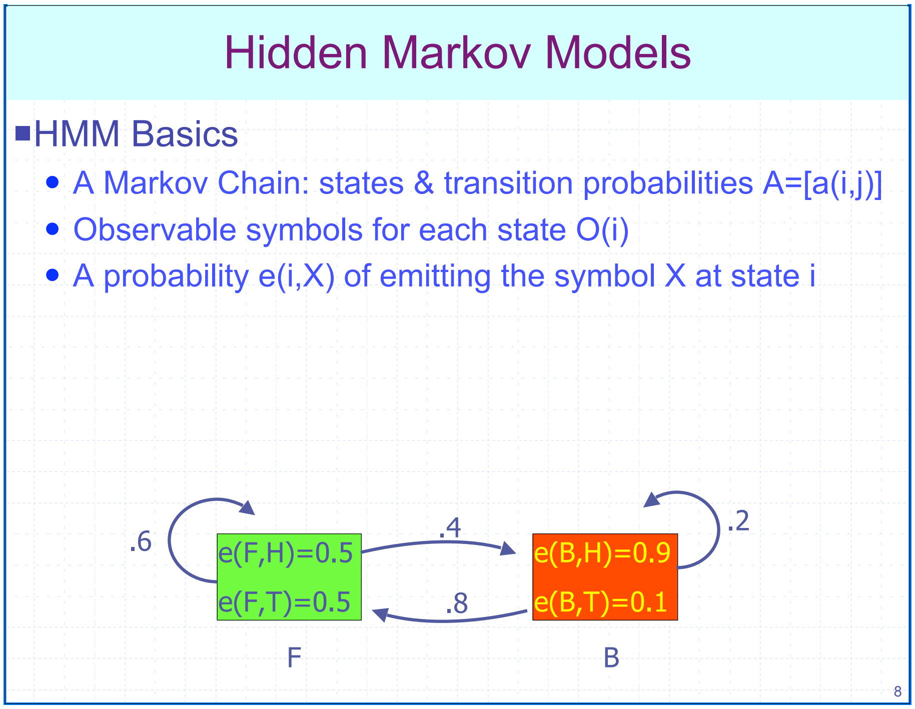
Review of Markovian Modeling

- Recall: a Markov chain is described by transition probabilities
  - $\pi(n+1)=\mathbf{A}\pi(n)$ where $\pi(i,n)=\text{Prob}\{S(n)=i\}$ is the state probability
  - $A(i,j)=\text{Prob}[S(n+1)=j|S(n)=i]$ is the transition probability
- Markov chains describe statistical evolution
  - In time: evolutionary change depends on previous state only
  - In space: change depends on neighboring sites only



From Markov To Hidden Markov Models (HMM)

- Nature uses different statistics for evolving different regions
  - Gene regions: CpG, promoters, introns/exons…
  - Protein regions: active, interfaces, hydrophobic/philic…
- How can we tell regions?
  - Sample sequences have different statistics
  - Model regions as Markovian states emitting observed sequences…
- Example: CpG islands
  - Model: two connected MCs one for CpG one for normal
  - The MC is hidden; only sample sequences are seen
  - Detect transition to/from CpG MC
  - Similar to a dishonest casino: transition from fair to biased dice

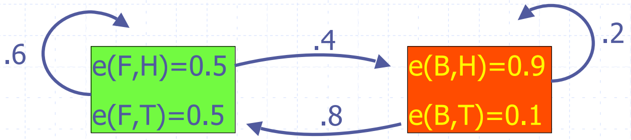


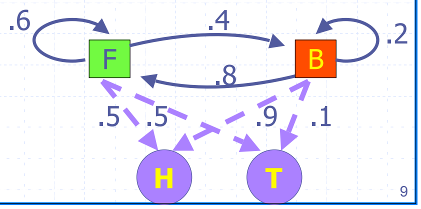
Hidden Markov Models

- HMM Basics
  - A Markov Chain: states & transition probabilities A=[a(i,j)]
  - Observable symbols for each state O(i)
  - A probability e(i,X) of emitting the symbol X at state i

Coin Example

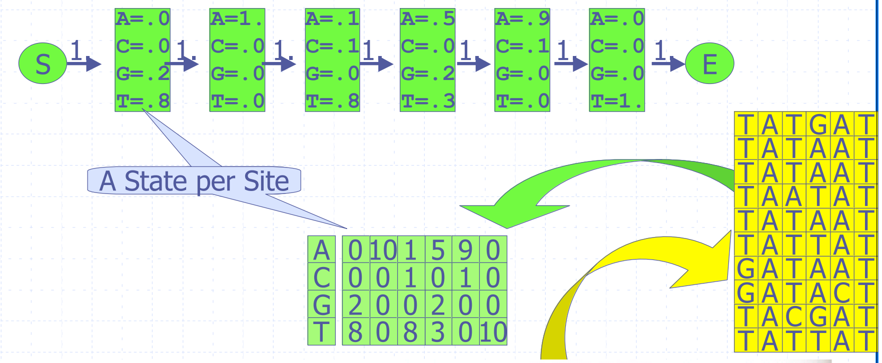
- Two states MC: {F,B} F=fair coin, B=biased
- Emission probabilities
  - Described in state boxes
  - Or through emission boxes
- Example: transmembrane proteins
  - Hydrophilic/hydrophobic regions





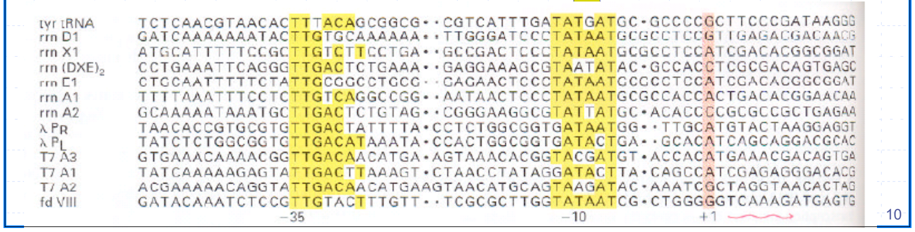
9

HMM Profile Example (Non-gapped)



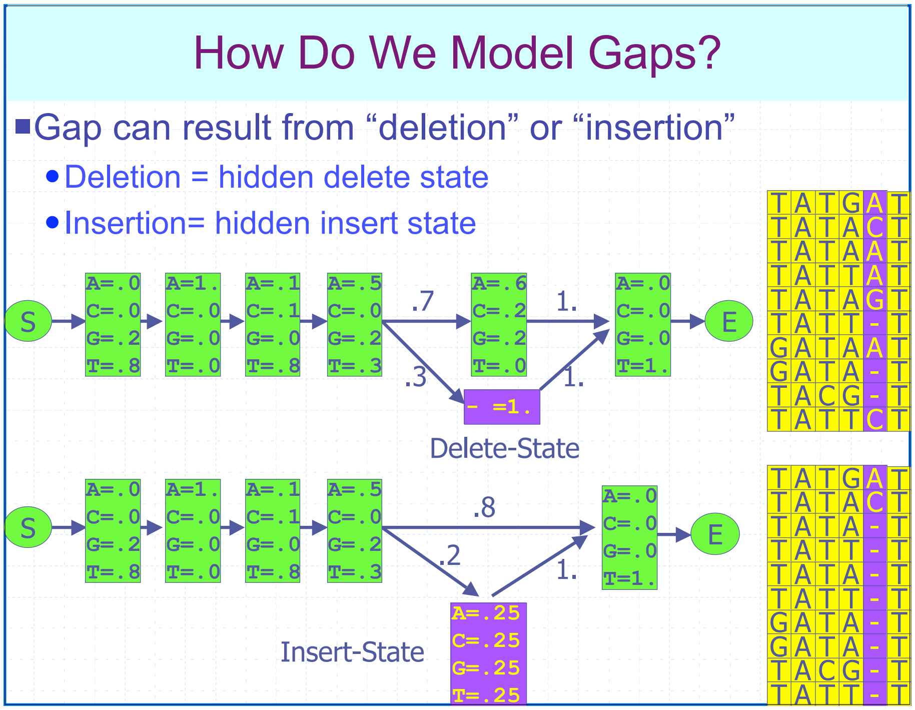
A	0	10	1	5	9	0
C	0	0	1	0	1	0
G	2	0	0	2	0	0
T	8	0	8	3	0	10

```
TATGAT
TATAAT
TATAAT
TAATAT
TATAAT
TATTAT
GATAAT
GATACT
TACGAT
TATTAT
```



```
tyr tRNA     TCTCAACGTAACACTTTACAGCGGCG••CGTCATTTGATATGATGC•GCCCCGCTTCCCGATAAGGG
rrn D1       GATCAAAAAAATACTTGTGCAAAAAA••TTGGGATCCCTATAATGCGCCTCCGTTGAGACGACAACG
rrn X1       ATGCATTTTTCCGCTTGTCTTCCTGA••GCCGACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn (DXE)2   CCTGAAATTCAGGGTTGACTCTGAAA••GAGGAAAGCGTAATATAC•GCCACCTCGCGACAGTGAGC
rrn E1       CTGCAATTTTTCTATTGCGGCCTGCG••GAGAACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn A1       TTTTAAATTTCCTCTTGTCAGGCCGG••AATAACTCCCTATAATGCGCCACCACTGACACGGAACAA
rrn A2       GCAAAAATAAATGCTTGACTCTGTAG••CGGGAAGGCGTATTATGC•ACACCCCGCGCCGCTGAGAA
λ PR         TAACACCGTGCGTGTTGACTATTTTA•CCTCTGGCGGTGATAATGG••TTGCATGTACTAAGGAGGT
λ PL         TATCTCTGGCGGTGTTGACATAAATA•CCACTGGCGGTGATACTGA••GCACATCAGCAGGACGCAC
T7 A3        GTGAAACAAAACGGTTGACAACATGA•AGTAAACACGGTACGATGT•ACCACATGAAACGACAGTGA
T7 A1        TATCAAAAAGAGTATTGACTTAAAGT•CTAACCTATAGGATACTTA•CAGCCATCGAGAGGGACACG
T7 A2        ACGAAAAACAGGTATTGACAACATGAAGTAACATGCAGTAAGATAC•AAATCGCTAGGTAACACTAG
fd VIII      GATACAAATCTCCGTTGTACTTTGTT••TCGCGCTTGGTATAATCG•CTGGGGGTCAAAGATGAGTG
                           −35                       −10          +1
```

10

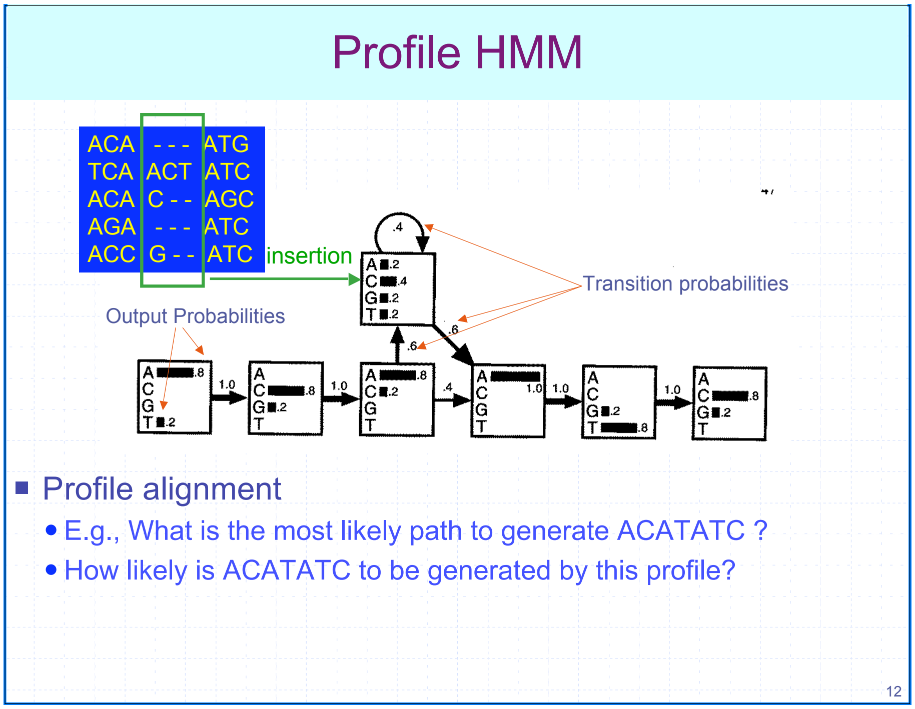


How Do We Model Gaps?

- Gap can result from “deletion” or “insertion”
  - Deletion = hidden delete state
  - Insertion= hidden insert state

T	A	T	G	A	T
T	A	T	A	C	T
T	A	T	A	A	T
T	A	T	T	A	T
T	A	T	A	G	T
T	A	T	T	-	T
G	A	T	A	A	T
G	A	T	A	-	T
T	A	C	G	-	T
T	A	T	T	C	T

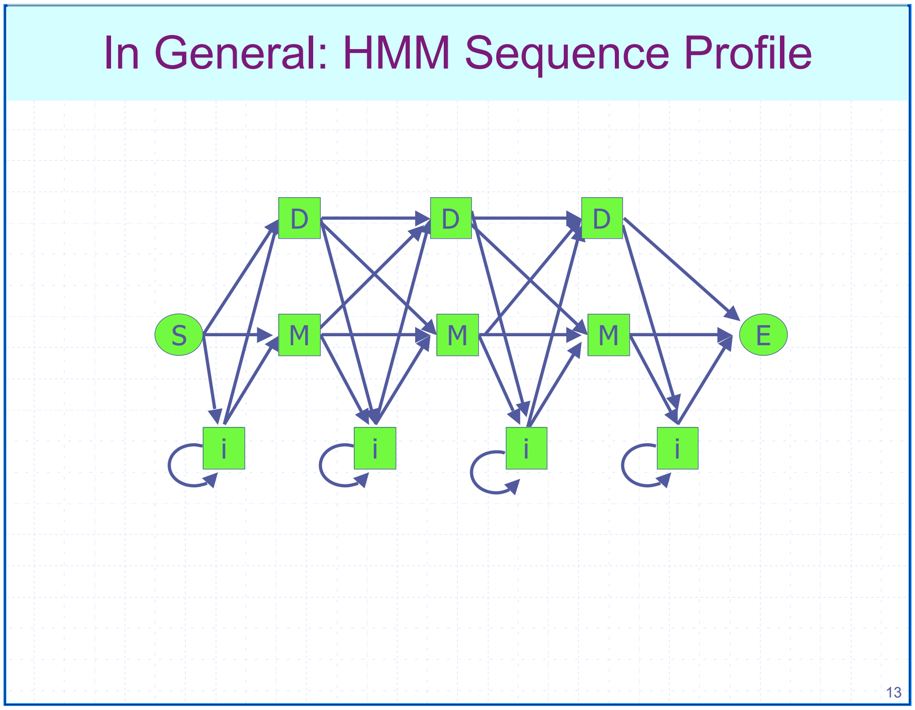
T	A	T	G	A	T
T	A	T	A	C	T
T	A	T	A	-	T
T	A	T	T	-	T
T	A	T	A	-	T
T	A	T	T	-	T
G	A	T	A	-	T
G	A	T	A	-	T
T	A	C	G	-	T
T	A	T	T	-	T



Profile HMM

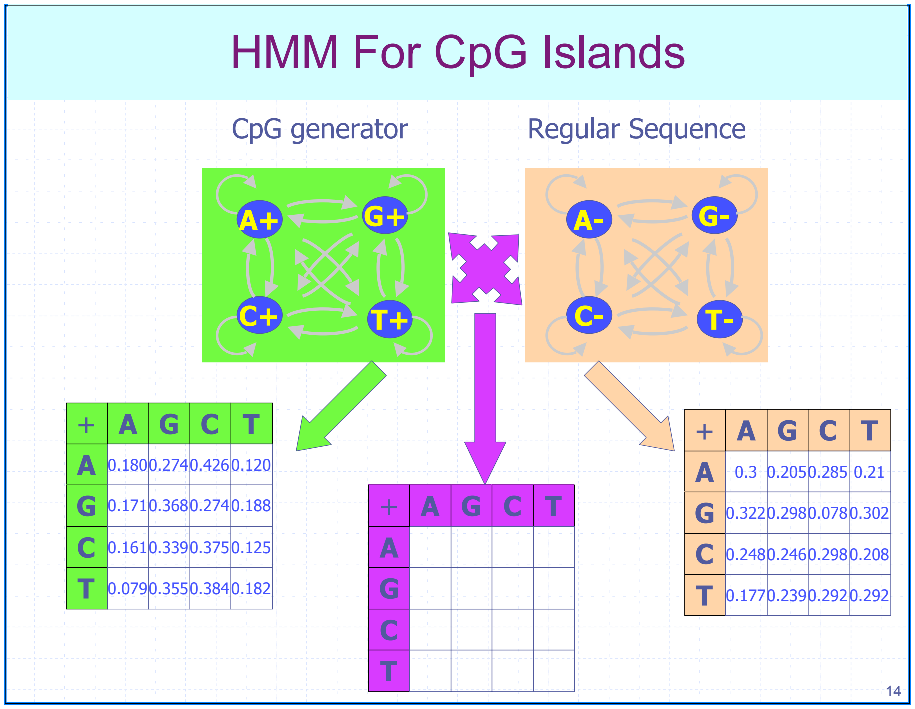
ACA	- - -	ATG
TCA	ACT	ATC
ACA	C - -	AGC
AGA	- - -	ATC
ACC	G - -	ATC

- Profile alignment
  - E.g., What is the most likely path to generate ACATATC ?
  - How likely is ACATATC to be generated by this profile?



In General: HMM Sequence Profile

13



HMM For CpG Islands

CpG generator

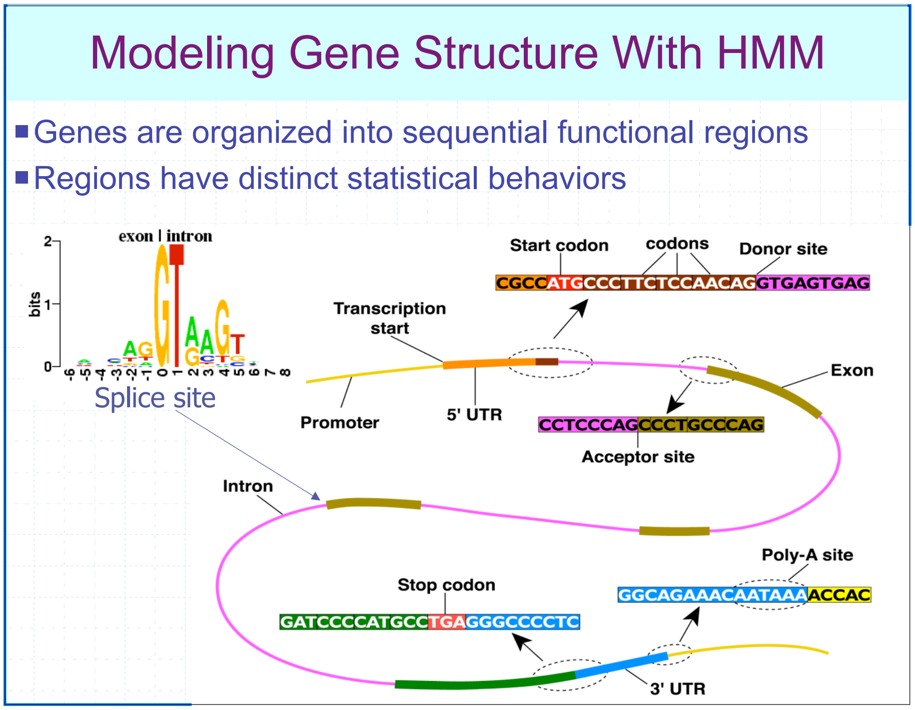
Regular Sequence

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

+	A	G	C	T
A				
G				
C				
T				

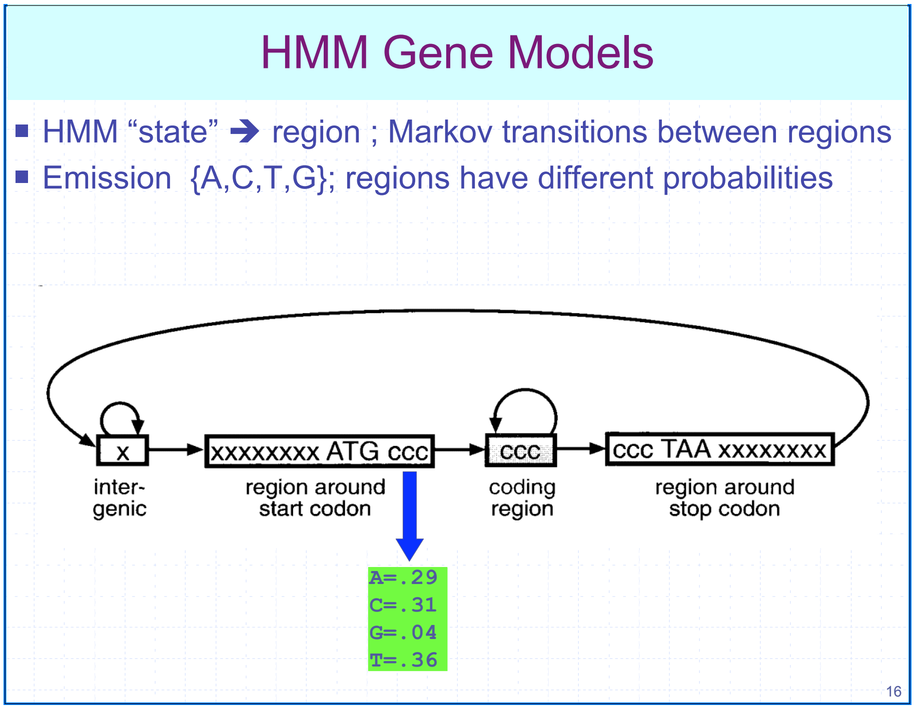
+	A	G	C	T
A	0.3	0.205	0.285	0.21
G	0.322	0.298	0.078	0.302
C	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

14



Modeling Gene Structure With HMM

- Genes are organized into sequential functional regions
- Regions have distinct statistical behaviors

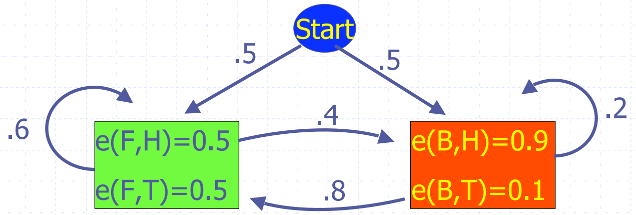


HMM Gene Models

- HMM "state" ➔ region ; Markov transitions between regions
- Emission {A,C,T,G}; regions have different probabilities

Computing Probabilities on HMM

- Path = a sequence of states
  - E.g., X=FFBBBF
  - Path probability: 0.5 $(0.6)^2$ $0.4(0.2)^3$ 0.8 =4.608 $*10^{-4}$
- Probability of a sequence emitted by a path: p(S|X)
  - E.g., p(HHHHHH|FFBBBF)=p(H|F)p(H|F)p(H|B)p(H|B)p(H|B)p(H|F) $=(0.5)^3(0.9)^3$ =0.09
- Note: usually one avoids multiplications and computes logarithms to minimize error propagation



17

The Three Computational Problems of HMM

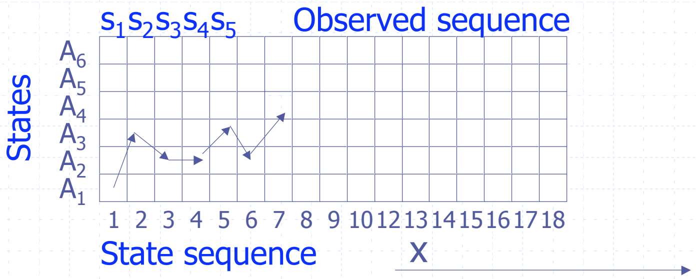
- Decoding: what is its most likely sequence of transitions & emissions that generated a given observed sequence?
- Likelihood: how likely is an observed sequence to have been generated by a given HMM?
- Learning: how should transition and emission probabilities be learned from observed sequences?

18

The Decoding Problem: Viterbi's Decoder

- Input: an observed sequence S
- Output: a hidden path X maximizing P(S|X)

- Key Idea (Viterbi): map to a dynamic programming problem
  - Describe the problem as optimizing a path over a grid
  - DP search: (a) compute "price" of forward paths (b) backtrack
  - Complexity: O(m^2n) (m=number of states, n= sequence size)



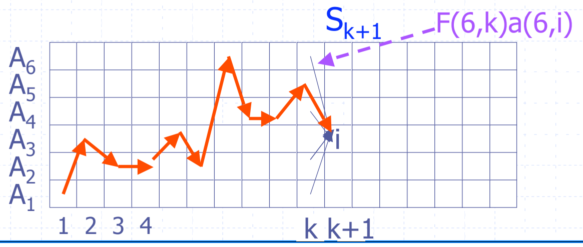
Viterbi's Decoder

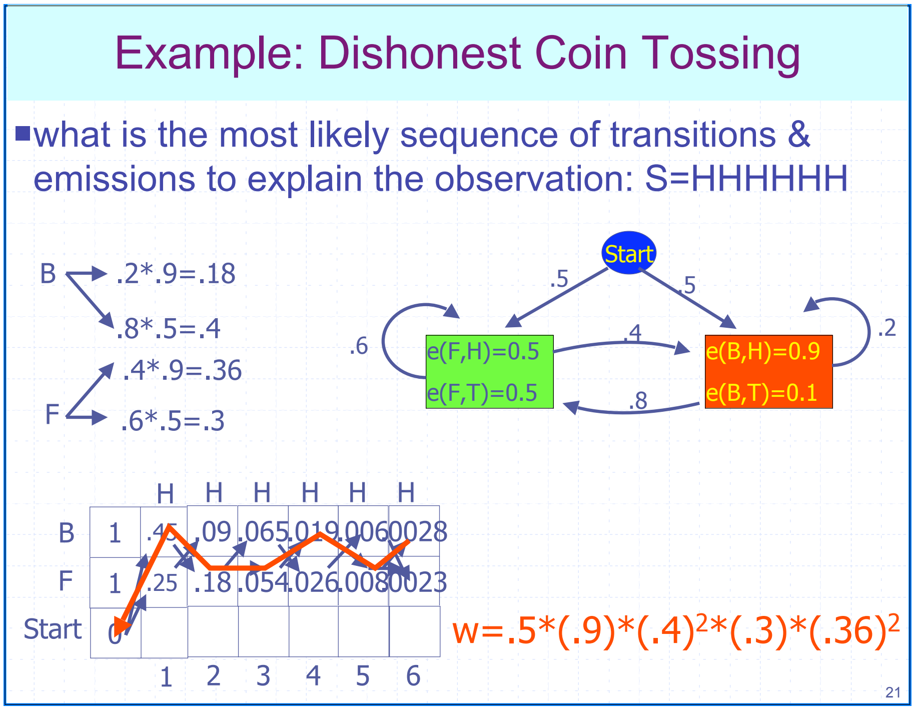
- F(i,k) = probability of the most likely path to state i generating $S_1 \ldots S_k$
- Forward recursion:

$$F(i,k+1)=e(i,S_{k+1})*\max_j\{F(j,k)a(i,j)\}$$

Best path to i

- Backtracking: start with highest F(i,n) and backtrack
- Initialization: F(0,0)=1, F(i,0)=0





Example: Dishonest Coin Tossing

- what is the most likely sequence of transitions & emissions to explain the observation: S=HHHHHH

B .2*.9=.18

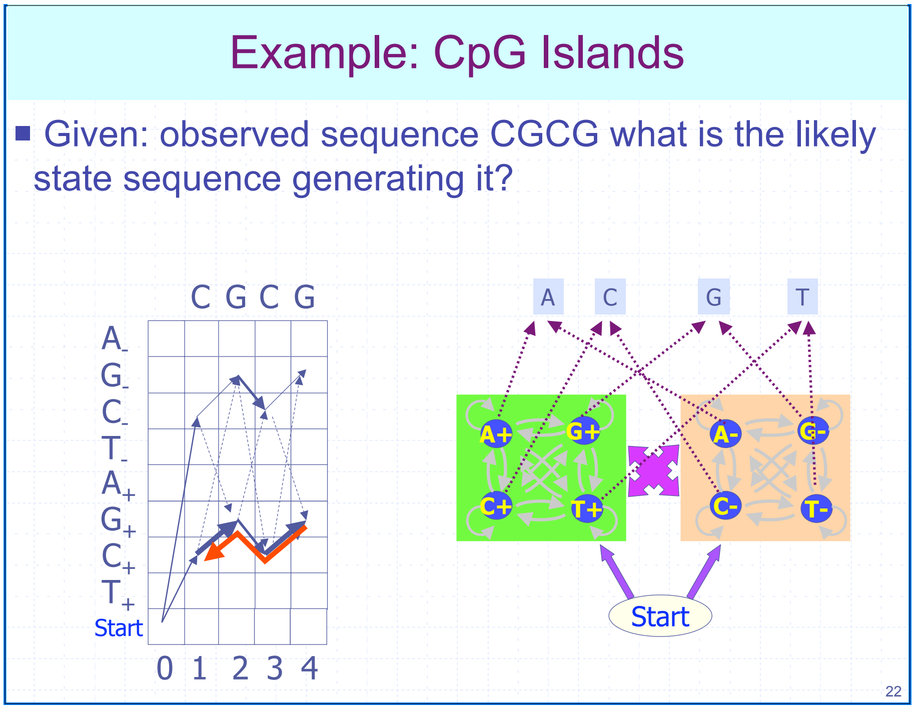
.8*.5=.4

.4*.9=.36

F .6*.5=.3

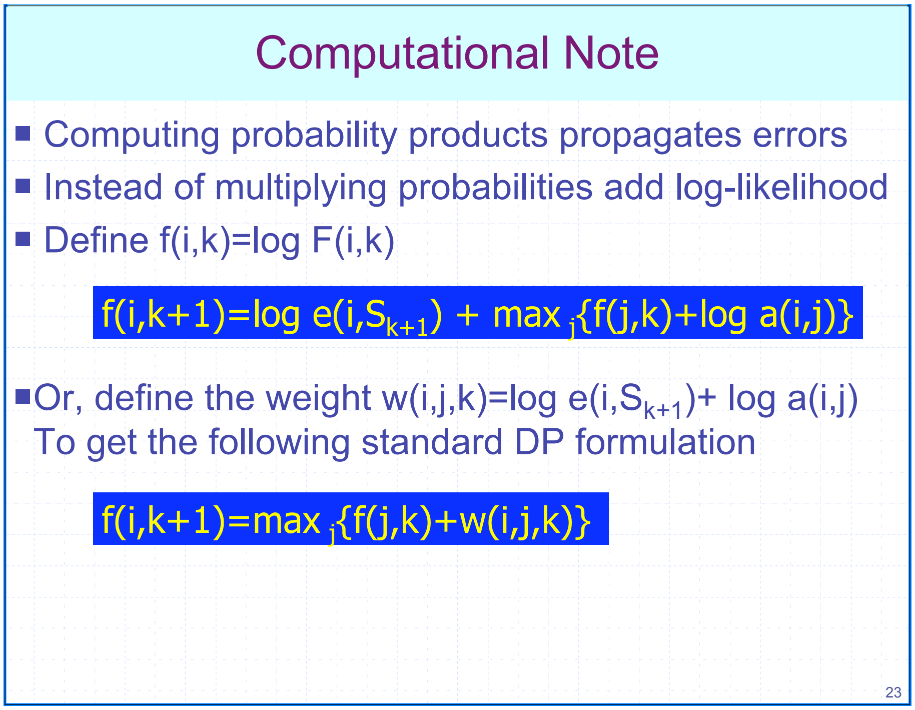
		H	H	H	H	H	H
B	1	.45	.09	.065	.019	.006	0028
F	1	.25	.18	.054	.026	.008	0023
Start	0						
		1	2	3	4	5	6

$w=.5*(.9)*(.4)^2*(.3)*(.36)^2$



Example: CpG Islands

- Given: observed sequence CGCG what is the likely state sequence generating it?



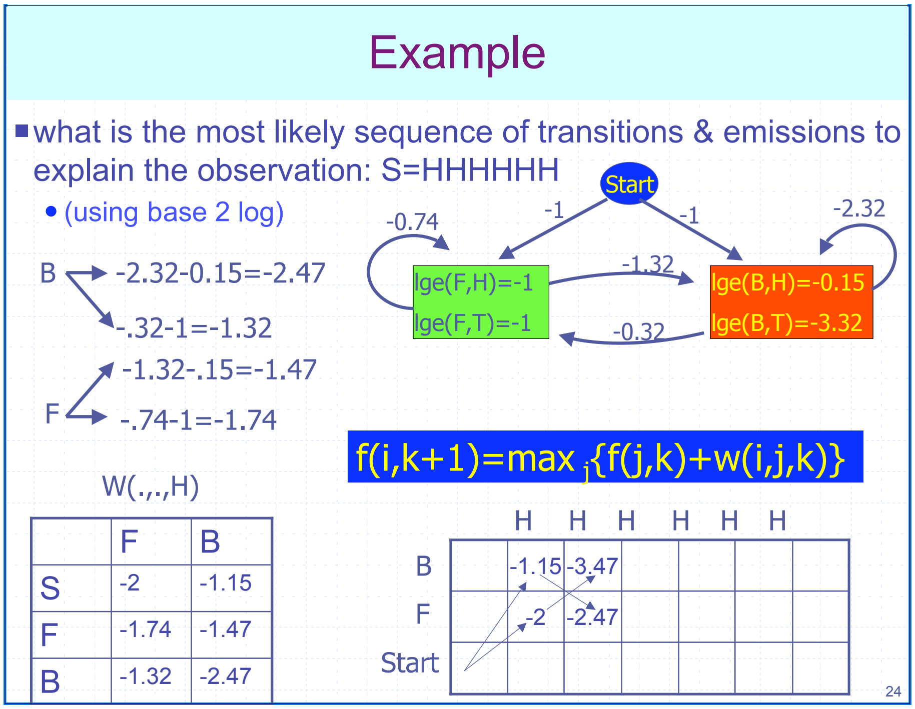
Computational Note

- Computing probability products propagates errors
- Instead of multiplying probabilities add log-likelihood
- Define f(i,k)=log F(i,k)

$$f(i,k+1)=\log e(i,S_{k+1}) + \max_j\{f(j,k)+\log a(i,j)\}$$

- Or, define the weight $w(i,j,k)=\log e(i,S_{k+1})+ \log a(i,j)$
To get the following standard DP formulation

$$f(i,k+1)=\max_j\{f(j,k)+w(i,j,k)\}$$



Example

- what is the most likely sequence of transitions & emissions to explain the observation: S=HHHHHH
  - (using base 2 log)

B → -2.32-0.15=-2.47
→ -.32-1=-1.32

F → -1.32-.15=-1.47
→ -.74-1=-1.74

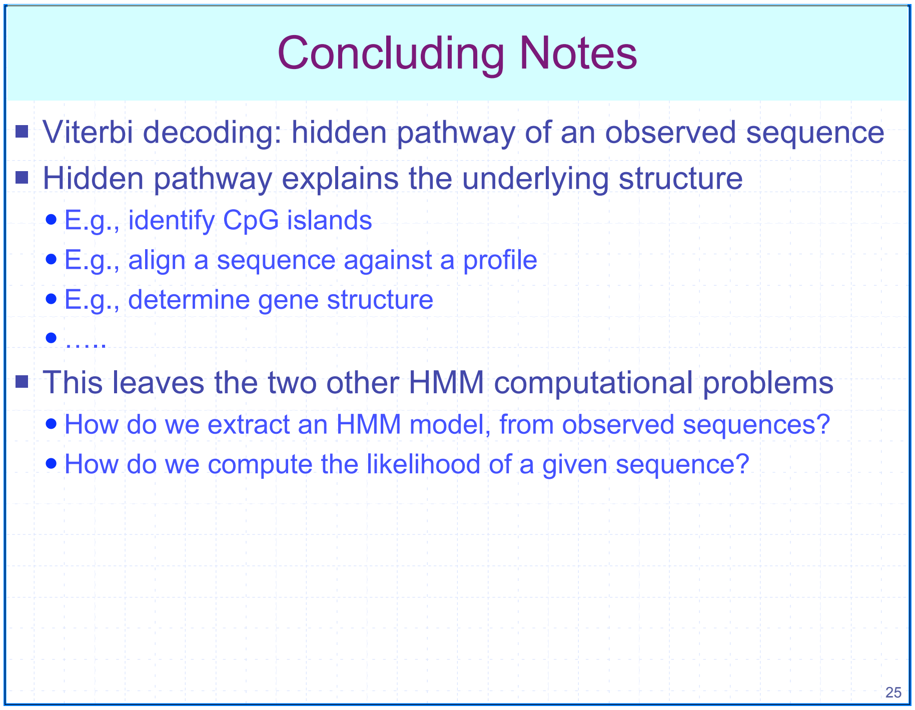
Start
-1
-1
-0.74
-2.32
lge(F,H)=-1
lge(F,T)=-1
-1.32
-0.32
lge(B,H)=-0.15
lge(B,T)=-3.32

$$f(i,k+1)=\max_j\{f(j,k)+w(i,j,k)\}$$

W(.,.,H)

	F	B
S	-2	-1.15
F	-1.74	-1.47
B	-1.32	-2.47

		H	H	H	H	H	H
B		-1.15	-3.47				
F		-2	-2.47				
Start							



Concluding Notes

- Viterbi decoding: hidden pathway of an observed sequence
- Hidden pathway explains the underlying structure
  - E.g., identify CpG islands
  - E.g., align a sequence against a profile
  - E.g., determine gene structure
  - …..
- This leaves the two other HMM computational problems
  - How do we extract an HMM model, from observed sequences?
  - How do we compute the likelihood of a given sequence?