Hidden Markov Models

Modified from:http://www.cs.iastate.edu/~cs544/Lectures/lectures.html

Nucleotide frequencies in the human genome

A	C	T	G
29.5	20.4	20.5	29.6

Written **CpG** to distinguish from a **C≡G** base pair)

CpG Islands

- **CpG** dinucleotides are rarer than would be expected from the independent probabilities of **C** and **G**.
  - Reason: When **CpG** occurs, C is typically chemically modified by methylation and there is a relatively high chance of **methyl-C** mutating into **T**
- High CpG frequency may be biologically significant; e.g., may signal promoter region ("start" of a gene).
- A **CpG island** is a region where **CpG** dinucleotides are much more abundant than elsewhere.

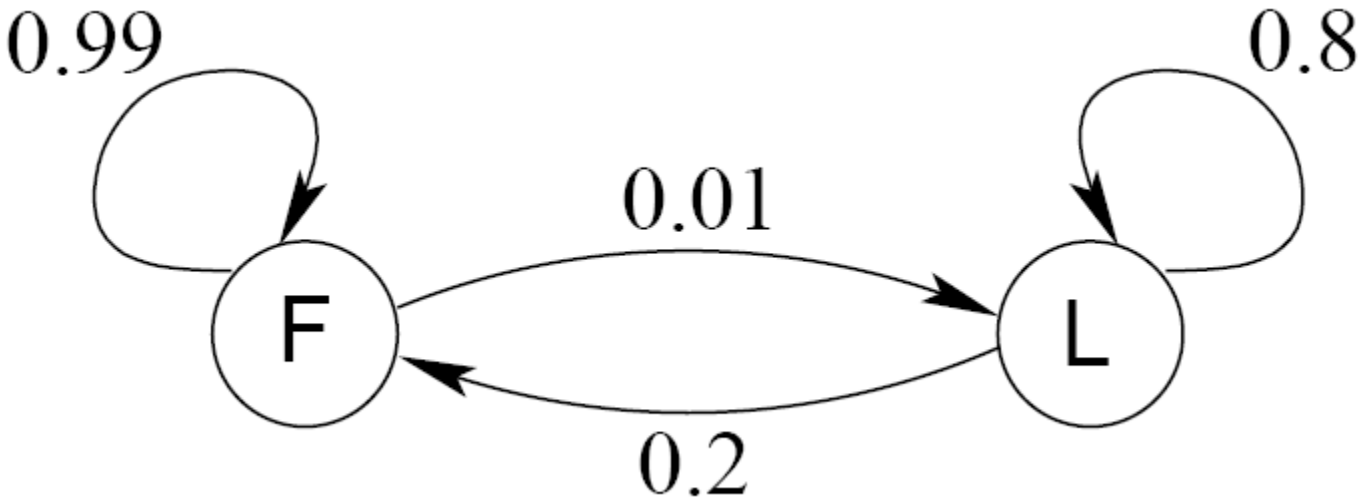
Hidden Markov Models

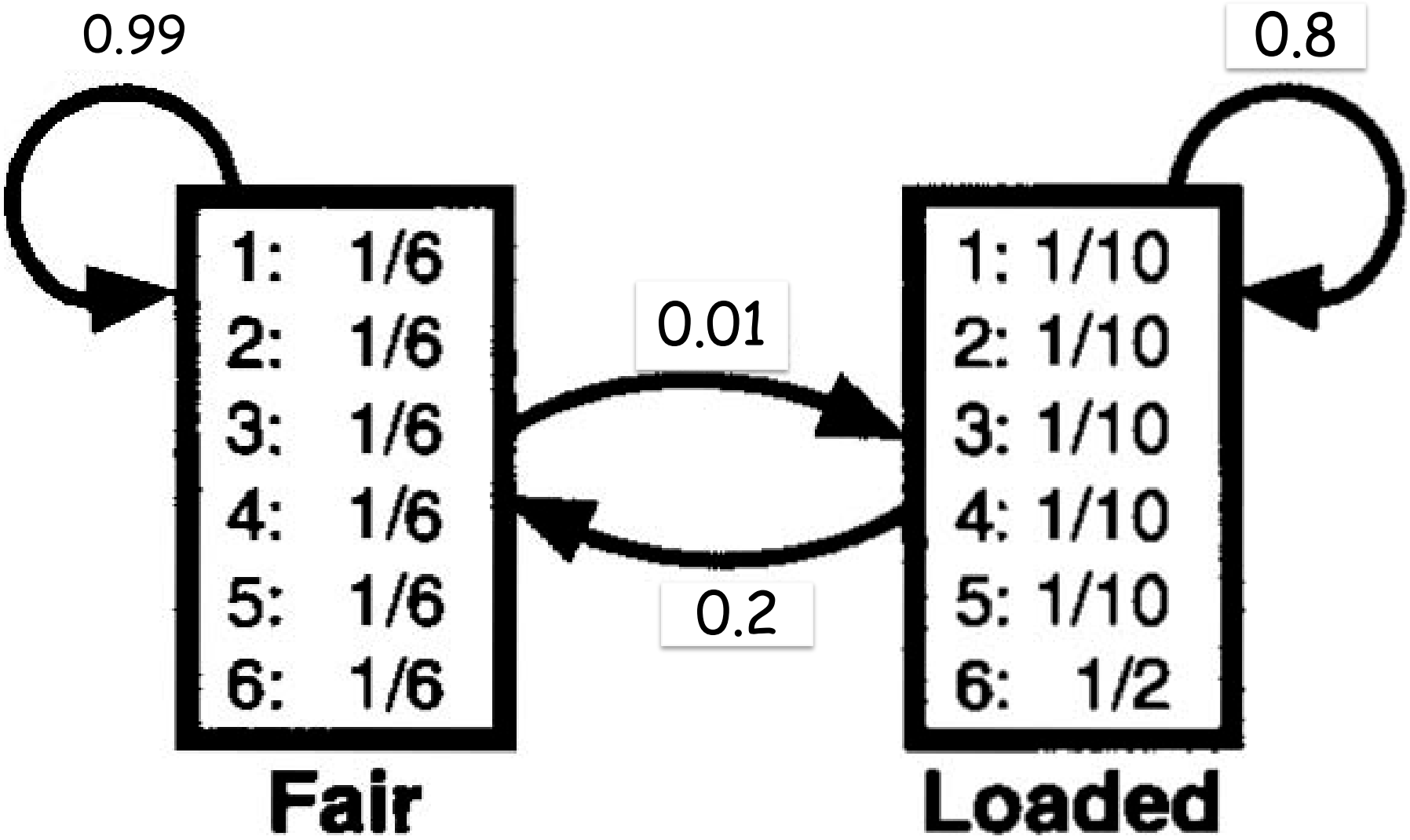
- Components:
  - Observed variables
    - Emitted symbols
  - Hidden variables
  - Relationships between them
    - Represented by a graph with transition probabilities
- **Goal**: Find the most likely explanation for the observed variables

The occasionally dishonest casino

- A casino uses a fair die most of the time, but occasionally switches to a loaded one
  - Fair die: Prob(1) = Prob(2) = . . . = Prob(6) = 1/6
  - Loaded die: Prob(1) = Prob(2) = . . . = Prob(5) = 1/10, Prob(6) = $\frac{1}{2}$
  - These are the ***emission*** probabilities
- ***Transition probabilities***
  - Prob(Fair $\rightarrow$ Loaded) = 0.01
  - Prob(Loaded $\rightarrow$ Fair) = 0.2
  - Transitions between states obey a Markov process

An HMM for the occasionally dishonest casino




0.99
0.8
1: 1/6
2: 1/6
3: 1/6
4: 1/6
5: 1/6
6: 1/6
0.01
0.2
1: 1/10
2: 1/10
3: 1/10
4: 1/10
5: 1/10
6: 1/2
Fair
Loaded

The occasionally dishonest casino

- Known:
  - The structure of the model
  - The transition probabilities
- Hidden: What the casino did
  - FFFFFLLLLLLLFFFF...
- Observable: The series of die tosses
  - 3415256664666153...
- What we must infer:
  - When was a fair die used?
  - When was a loaded one used?
    - The answer is a sequence FFFFFFFLLLLLLFFF...

Making the inference

- Model assigns a probability to each explanation of the observation:

  $P(326|FFL)$
  $= P(3|F) \cdot P(F \rightarrow F) \cdot P(2|F) \cdot P(F \rightarrow L) \cdot P(6|L)$
  $= 1/6 \cdot 0.99 \cdot 1/6 \cdot 0.01 \cdot \frac{1}{2}$

- **Maximum Likelihood:** Determine which explanation is most likely
  - Find the path *most likely* to have produced the observed sequence
- **Total probability:** Determine probability that observed sequence was produced by the HMM
  - Consider *all* paths that could have produced the observed sequence

Notation

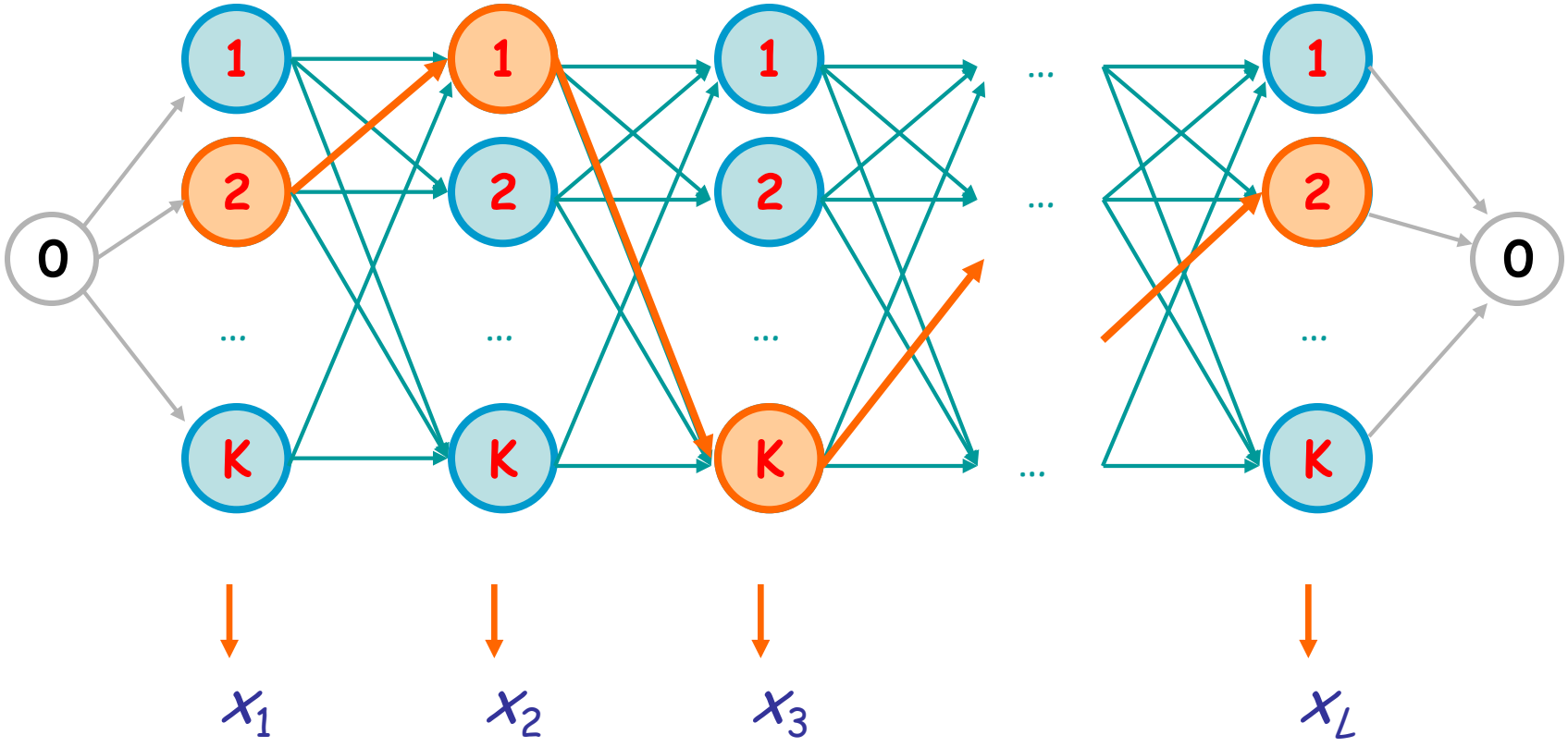
- x is the sequence of symbols emitted by model
  - x_i is the symbol emitted at time i
- A ***path***, π, is a sequence of states
  - The i-th state in π is π_i
- a_{kr} is the probability of making a transition from state k to state r:

$$a_{kr} = \Pr(\pi_i = r \mid \pi_{i-1} = k)$$

- $e_k(b)$ is the probability that symbol b is emitted when in state k

$$e_k(b) = \Pr(x_i = b \mid \pi_i = k)$$

A "parse" of a sequence



$$\Pr(x,\pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) \cdot a_{\pi_i \pi_{i+1}}$$

The occasionally dishonest casino

$$x = \langle x_1, x_2, x_3 \rangle = \langle 6,2,6 \rangle$$

$\pi^{(1)} = FFF$

$$\Pr(x, \pi^{(1)}) = a_{0F} e_F(6) a_{FF} e_F(2) a_{FF} e_F(6)$$
$$= 0.5 \times \frac{1}{6} \times 0.99 \times \frac{1}{6} \times 0.99 \times \frac{1}{6}$$
$$\approx 0.00227$$

$\pi^{(2)} = LLL$

$$\Pr(x, \pi^{(2)}) = a_{0L} e_L(6) a_{LL} e_L(2) a_{LL} e_L(6)$$
$$= 0.5 \times 0.5 \times 0.8 \times 0.1 \times 0.8 \times 0.5$$
$$= 0.008$$

$\pi^{(3)} = LFL$

$$\Pr(x, \pi^{(3)}) = a_{0L} e_L(6) a_{LF} e_F(2) a_{FL} e_L(6) a_{L0}$$
$$= 0.5 \times 0.5 \times 0.2 \times \frac{1}{6} \times 0.01 \times 0.5$$
$$\approx 0.0000417$$

The most probable path

The most likely path π^* satisfies

$$\pi^* = \arg\max_{\pi} \Pr(x, \pi)$$

To find π^*, consider all possible ways the last symbol of x could have been emitted

Let

$$v_k(i) = \text{Prob. of path } \langle \pi_1, \cdots, \pi_i \rangle \text{ most likely to emit } \langle x_1, \ldots, x_i \rangle \text{ such that } \pi_i = k$$

Then

$$v_k(i) = e_k(x_i) \max_r \left(v_r(i-1) a_{rk} \right)$$

The Viterbi Algorithm

- Initialization ($i = 0$)

$$v_0(0) = 1, \quad v_k(0) = 0 \text{ for } k > 0$$

- Recursion ($i = 1, \ldots, L$): For each state k

$$v_k(i) = e_k(x_i) \max_r \left(v_r(i-1) a_{rk} \right)$$

- Termination:

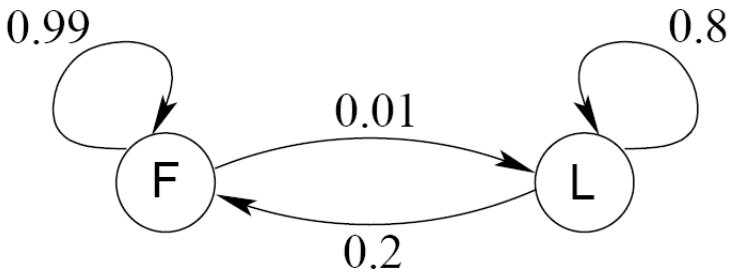
$$\Pr(x, \pi^*) = \max_k \left(v_k(L) a_{k0} \right)$$

To find π^*, use trace-back, as in dynamic programming

Viterbi: Example

π		ε	6	2	6
				x	
	B	1	0	0	0
	F	0	(1/6)×(1/2) = 1/12	(1/6)×max{(1/12)×0.99, (1/4)×0.2} = 0.01375	(1/6)×max{0.01375×0.99, 0.02×0.2} = 0.00226875
	L	0	(1/2)×(1/2) = 1/4	(1/10)×max{(1/12)×0.01, (1/4)×0.8} = 0.02	(1/2)×max{0.01375×0.01, 0.02×0.8} = 0.08

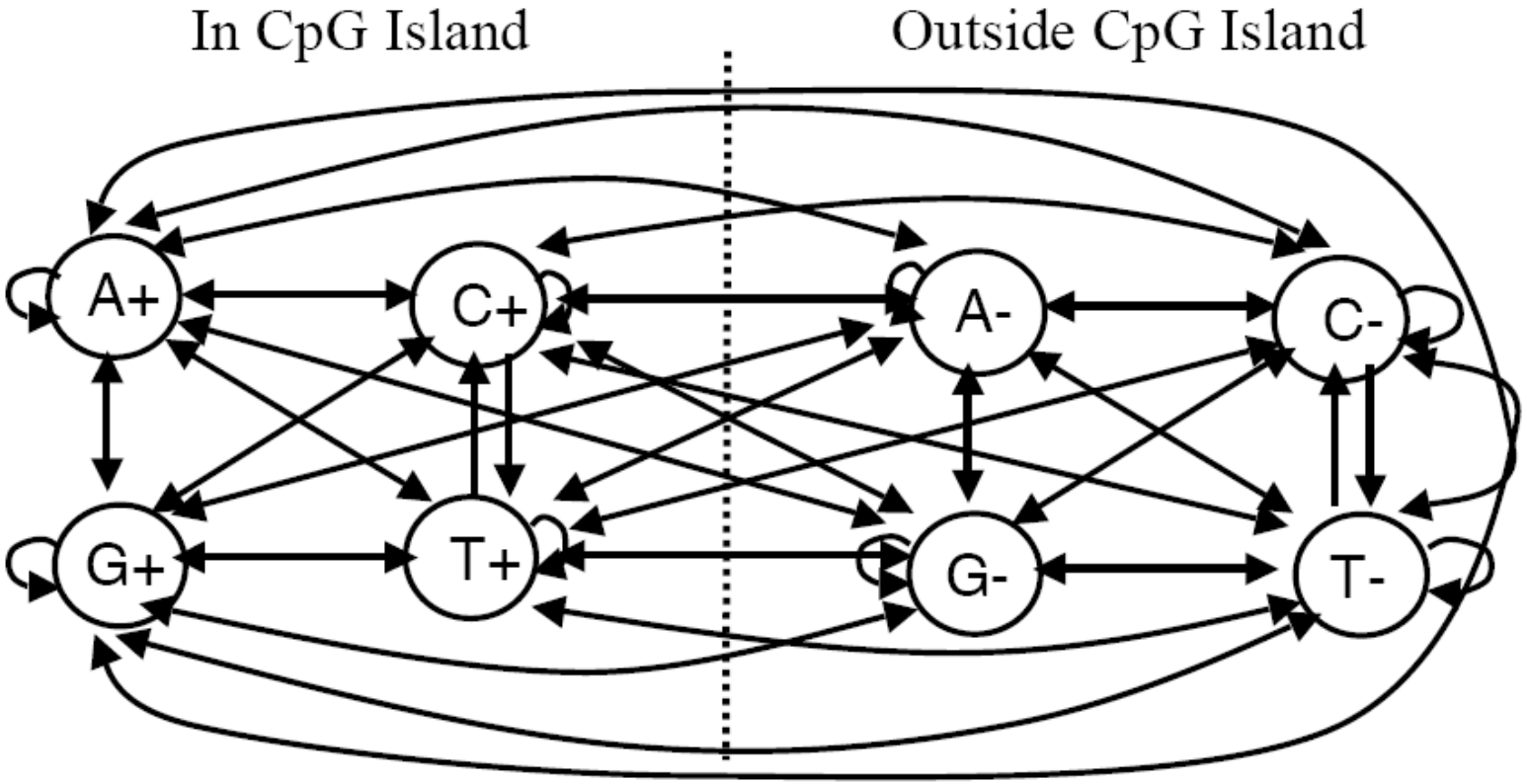
$$v_k(i) = e_k(x_i)\max_r\left(v_r(i-1)a_{rk}\right)$$



Viterbi gets it right more often than not

Rolls	31511624644664424532113163116415213362514454363165662656666
Die	FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLL
Viterbi	FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLL
Rolls	65116645313265124563666463163666316232645523526666662515163 1
Die	LLLLLLFFFFFFFFFFFLLLLLLLLLLLLLLLLFFFLLLLLLLLLLLLLFFFFFFFFF
Viterbi	LLLLLLFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF
Rolls	22255544166656656356432436413151346514635341112641462625335 6
Die	FFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL
Viterbi	FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL
Rolls	36616366646623253441366166116325256246225526525226643535333 6
Die	LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi	LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Rolls	23312162536441443233516324363366556246666263266661235524524 2
Die	FFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFF
Viterbi	FFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFF

An HMM for CpG islands



Emission probabilities are 0 or 1. E.g. $e_{G-}(G) = 1$, $e_{G-}(T) = 0$

See Durbin et al., *Biological Sequence Analysis*,. Cambridge 1998

Total probabilty

Many different paths can result in observation x.

The probability that our model will emit x is

$$\Pr(x) = \sum_{\pi} \Pr(x, \pi)$$

Total Probability

If HMM models a family of objects, we want total probability to peak at members of the family. (Training)

Total probability

Pr(x) can be computed in the same way as probability of most likely path.

Let

$$f_k(i) = \text{Prob. of observing } \langle x_1, \ldots, x_i \rangle$$
$$\text{assuming that } \pi_i = k$$

Then

$$f_k(i) = e_k(x_i) \sum_r f_r(i-1) a_{rk}$$

and

$$\Pr(x) = \sum_k f_k(L) a_{k0}$$

The Forward Algorithm

- Initialization ($i = 0$)

$$f_0(0) = 1, \quad f_k(0) = 0 \text{ for } k > 0$$

- Recursion ($i = 1, \ldots, L$): For each state k

$$f_k(i) = e_k(x_i) \sum_r f_r(i-1) a_{rk}$$

- Termination:

$$\Pr(x) = \sum_k f_k(L) a_{k0}$$

The Backward Algorithm

- Initialization ($i = L$)

$$b_k(L) = a_{k0} \text{ for all } k$$

- Recursion ($i = L-1, \ldots, 1$): For each state k

$$b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)$$

- Termination:

$$Pr(x) = \sum_l a_{0l} e_l(x_1) b_l(1)$$

Posterior Decoding

- How likely is it that my observation comes from a certain state?

  $P(x_i \text{ is emitted by state } k \mid \text{whole observation})$

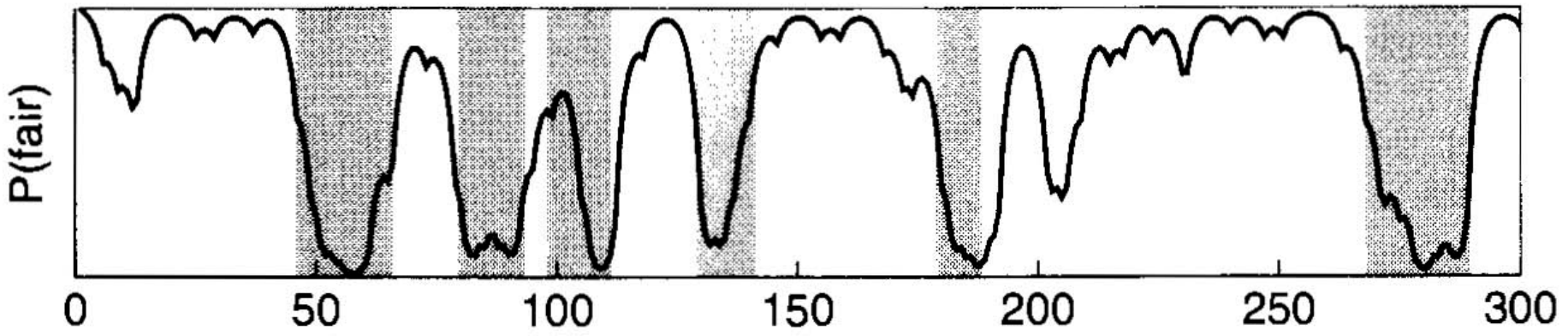
- Like the Forward matrix, one can compute a Backward matrix
- Multiply Forward and Backward entries

$$P(\pi_i = k \mid x) = \frac{f_k(i) \cdot b_k(i)}{P(x)}$$

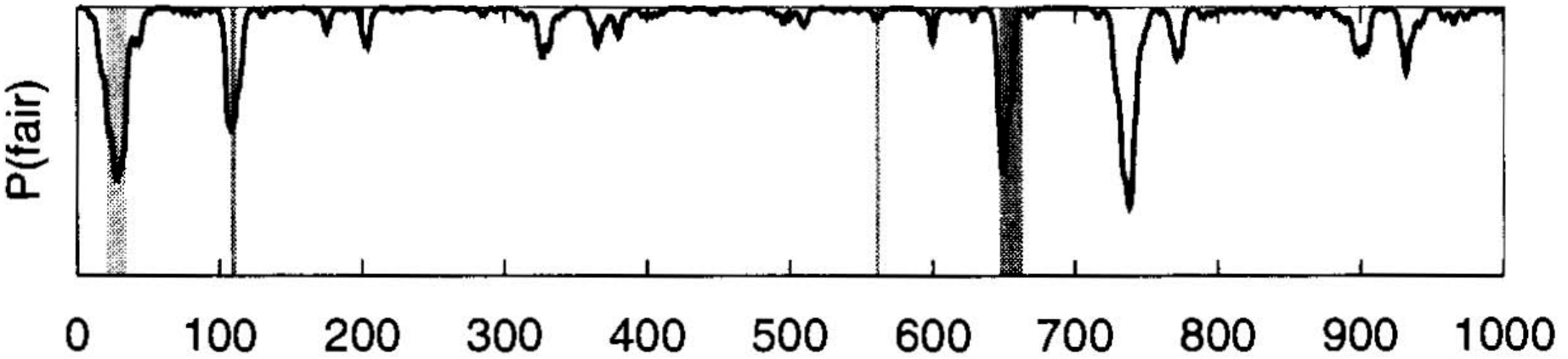
  - P(x) is the total probability computed by, e.g., forward algorithm

Posterior Decoding

With prob 0.05 for switching to the loaded die:



With prob 0.01 for switching to the loaded die:



Estimating the probabilities ("training")

- **Baum-Welch algorithm**
  - Start with initial guess at transition probabilities
  - Refine guess to improve the total probability of the training data in each step
    - May get stuck at local optimum
  - Special case of expectation-maximization (EM) algorithm
- **Viterbi training**
  - Derive probable paths for training data using Viterbi algorithm
  - Re-estimate transition probabilities based on Viterbi path
  - Iterate until paths stop changing

Profile HMMs

- Model a family of sequences
- Derived from a multiple alignment of the family
- Transition and emission probabilities are position-specific
- Set parameters of model so that total probability peaks at members of family
- Sequences can be tested for membership in family using Viterbi algorithm to match against profile

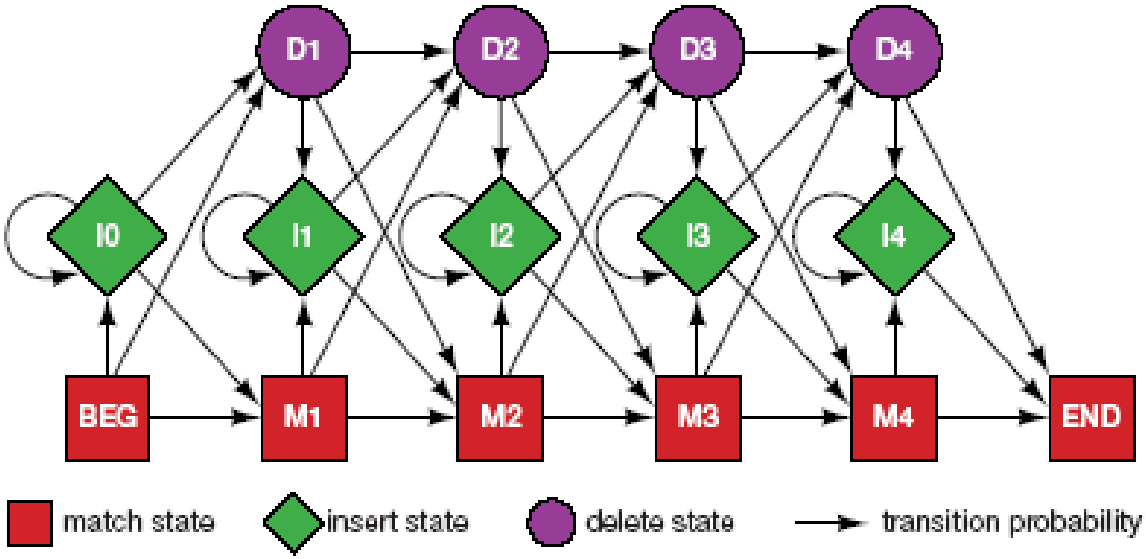
Profile HMMs

A. Sequence alignment

N	•	F	L	S
N	•	F	L	S
N	K	Y	L	T
Q	•	W	–	T

RED POSITION REPRESENTS ALIGNMENT IN COLUMN
GREEN POSITION REPRESENTS INSERT IN COLUMN
PURPLE POSITION REPRESENTS DELETE IN COLUMN

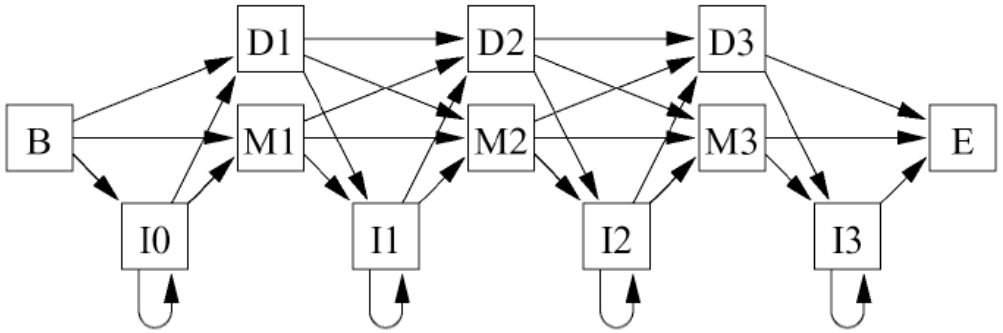
B. Hidden Markov model for sequence alignment



Profile HMMs: Example

An alignment of proteins from the HMM:

-	E	G	-	K	-
-	E	A	-	K	-
P	D	-	-	K	L
-	E	G	I	W	-



The states giving this alignment:

B → M1 → M2 → M3 → E

B → M1 → M2 → M3 → E

B → I0 → M1 → D2 → M3 → I3 → E

B → M1 → M2 → I2 → M3 → E

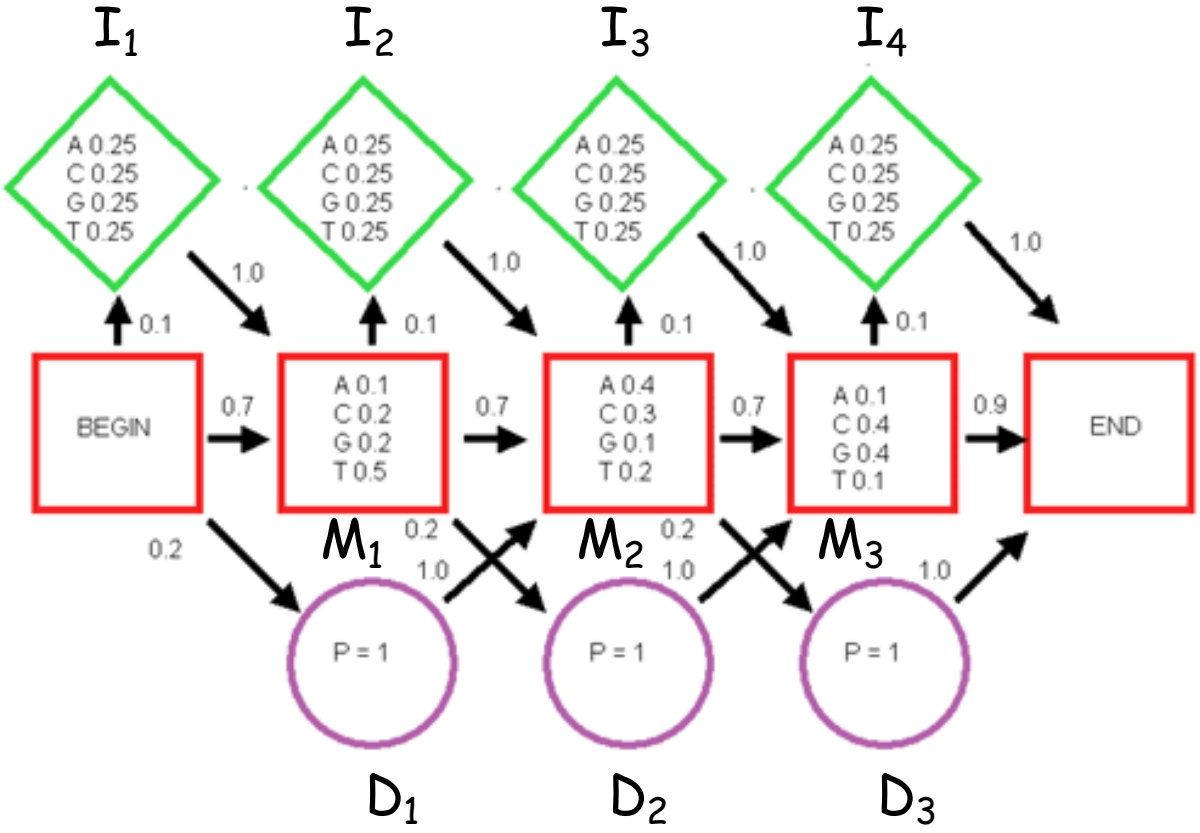
Note: These sequences could lead to other paths.

Source: http://www.csit.fsu.edu/~swofford/bioinformatics_spring05/

Pfam

- "A comprehensive collection of protein domains and families, with a range of well-established uses including genome annotation."
- Each family is represented by two multiple sequence alignments and two profile-Hidden Markov Models (profile-HMMs).
- A. Bateman et al. *Nucleic Acids Research* (2004) Database Issue 32:D138-D141

Lab 5

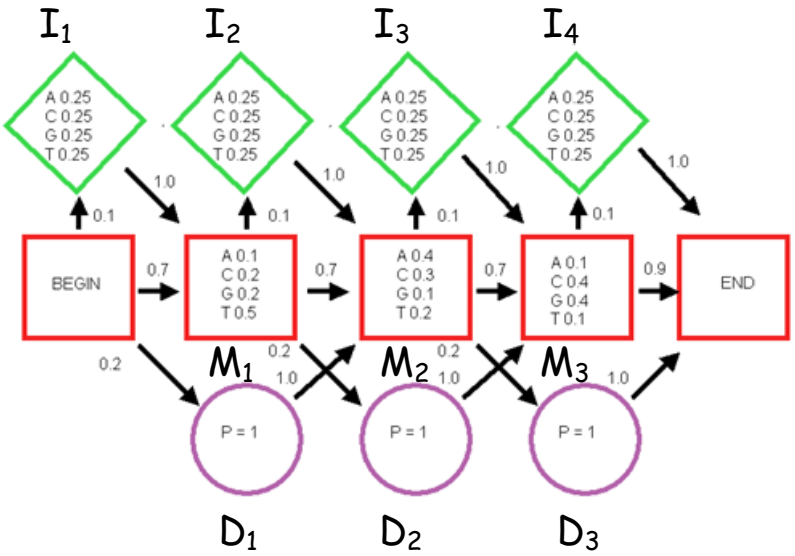

I1
I2
I3
I4
A 0.25
C 0.25
G 0.25
T 0.25
1.0
0.1
BEGIN
0.7
A 0.1
C 0.2
G 0.2
T 0.5
A 0.4
C 0.3
G 0.1
T 0.2
A 0.1
C 0.4
G 0.4
T 0.1
0.9
END
0.2
M1
M2
M3
P = 1
D1
D2
D3

Some recurrences

$$v_{M_1}(i) = e_{M_1}(x_i) \cdot \max \begin{cases} a_{BM_1} \cdot v_B(i-1) \\ a_{I_1M_1} \cdot v_{I_1}(i-1) \end{cases}$$

$$v_{I_1}(i) = e_{I_1}(x_i) \cdot a_{BI_1} \cdot v_B(i-1)$$

$$v_{D_1}(i) = e_{D_1}(-) \cdot a_{BD_1} \cdot v_B(i)$$

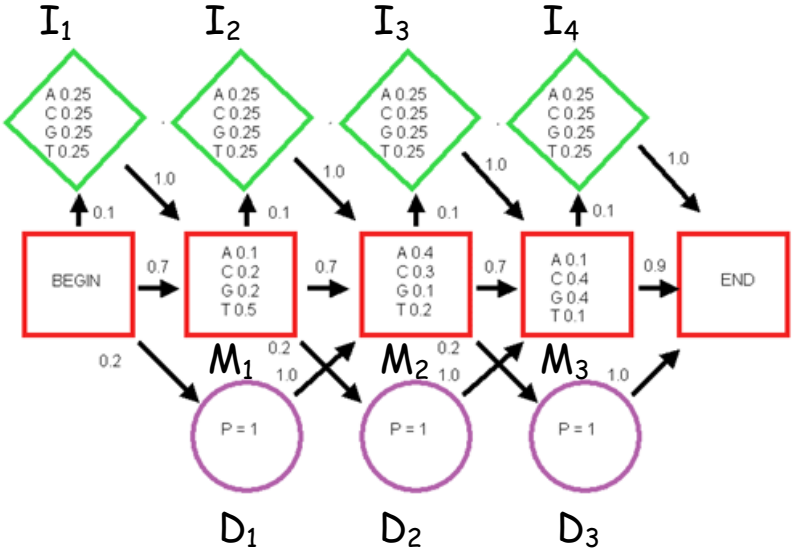


More recurrences

$$v_{M_2}(i) = e_{M_2}(x_i) \cdot \max \begin{cases} a_{I_2 M_2} \cdot v_{I_2}(i-1) \\ a_{M_1 M_2} \cdot v_{M_1}(i-1) \\ a_{D_1 M_2} \cdot v_{D_1}(i-1) \end{cases}$$

$$v_{I_2}(i) = e_{I_2}(x_i) \cdot a_{M_1 I_2} \cdot v_{M_1}(i-1)$$

$$v_{D_2}(i) = e_{D_2}(-) \cdot a_{M_1 D_2} \cdot v_{M_1}(i)$$



	ε	T	A	G	ε
Begin	1	0	0	0	0
M_1	0	0.35			
M_2	0	0.04			
M_3	0	0			
I_1	0	0.025			
I_2	0	0			
I_3	0	0			
I_4	0	0			
D_1	0.2	0			
D_2	0	0.07			
D_3	0	0			
End	0	0			