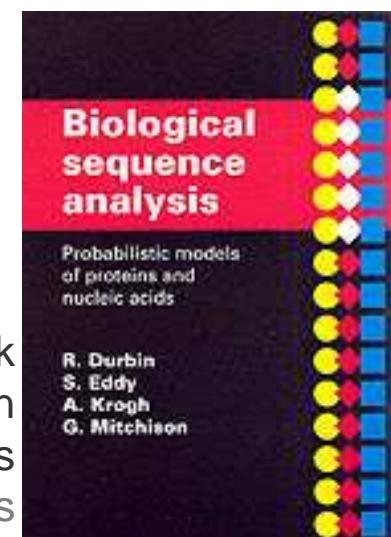
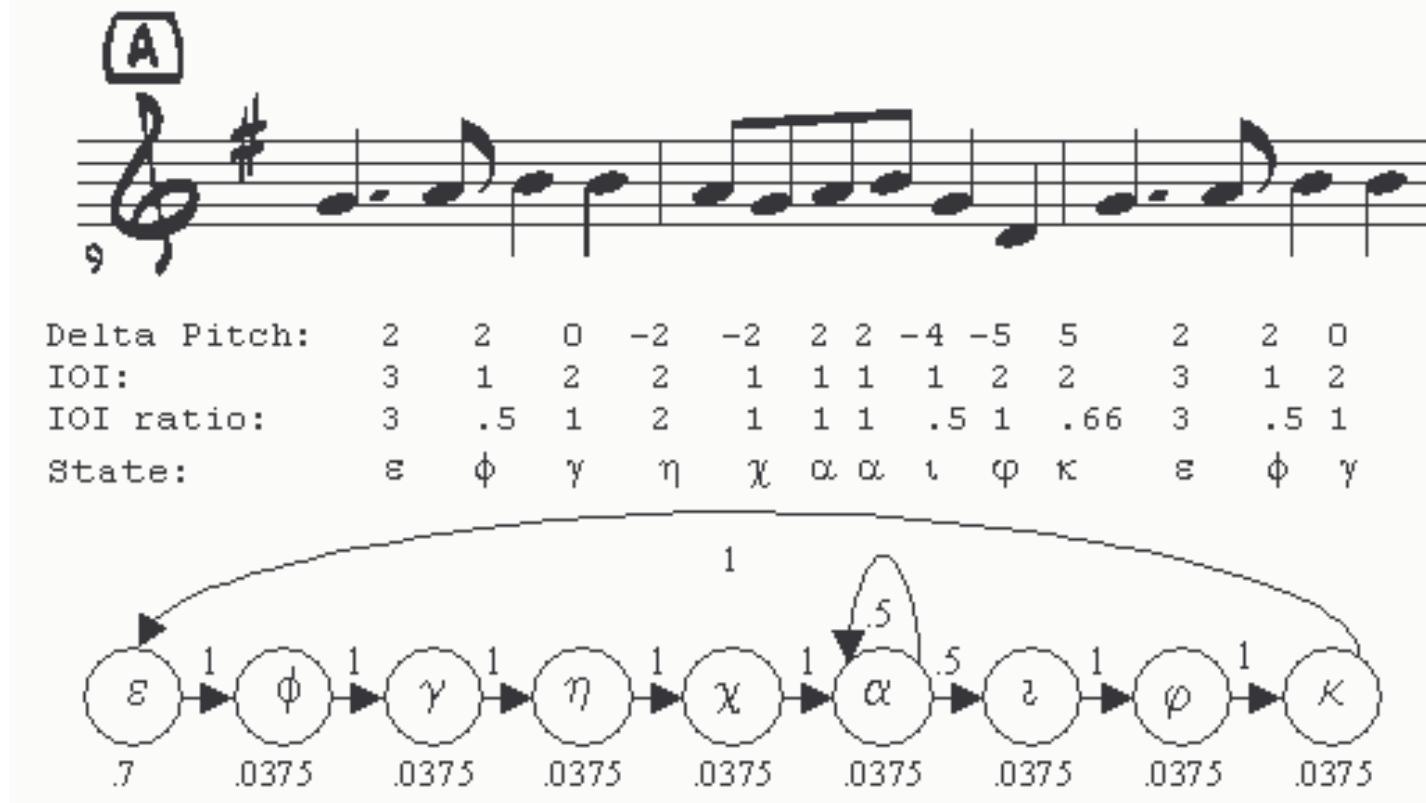


# Hidden Markov Models

based on chapters from the book  
Durbin, Eddy, Krogh and Mitchison  
Biological Sequence Analysis  
via Shamir's lecture notes

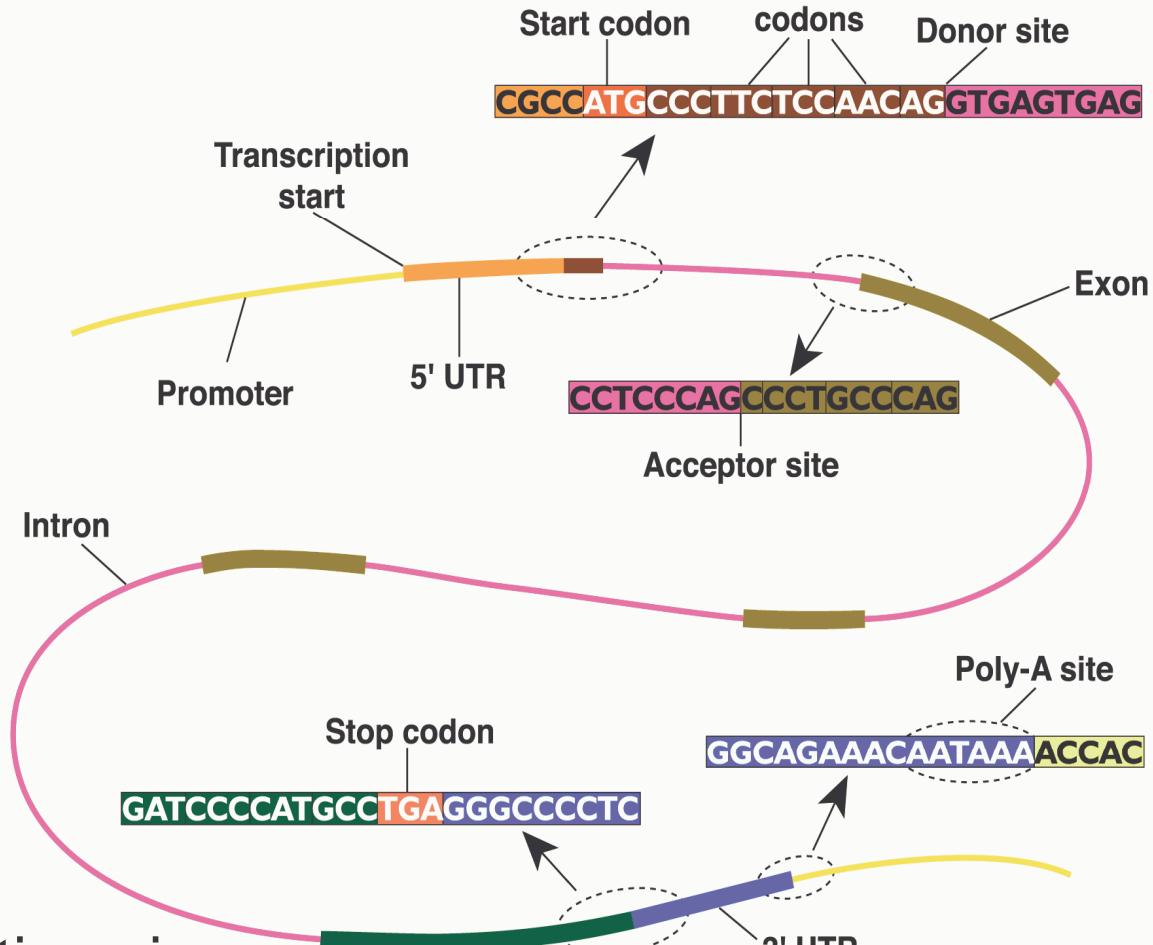


# music recognition



deal with variations in  
- actual sound  
- timing

# application: gene finding



deal with variations in

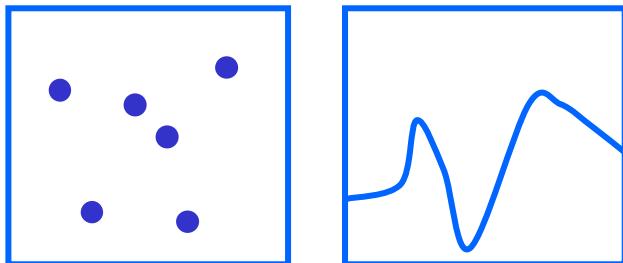
- actual sound → actual base
- timing → insertions/deletions

# basic questions

start with

- sequence of ‘observations’
- probabilistic model of our ‘domain’
- does this sequence belong to a certain family?  
Markov chains
- can we say something about the internal structure?  
(indirect observation)  
HMM: Hidden Markov Models

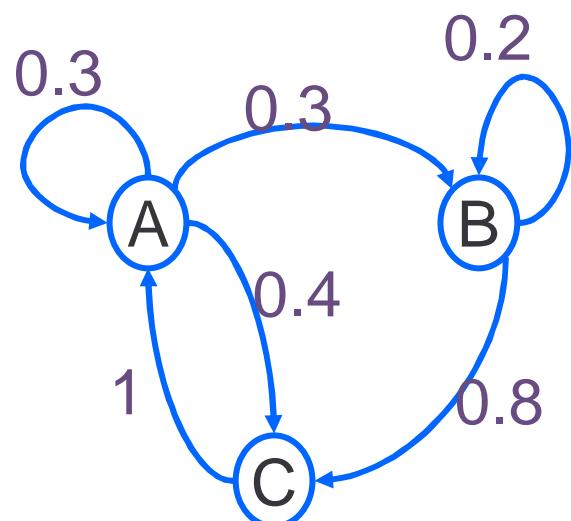
# introduction



discrete time  
discrete space

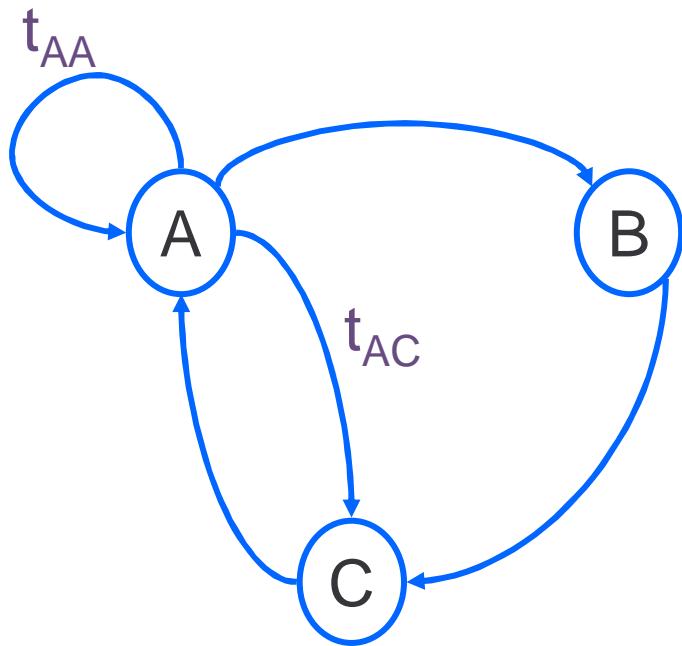
no state history:  
present state only

states, transitions



$P(X)$	
$P(X,Y)$	X and Y
$P(X Y)$	X, given Y

# Markov model



model  $M = (Q, P, T)$

- states  $Q$
- initial probabilities  $p_x$
- transition probabilities  $t_{xy}$   
matrix / graph

first order: no history

observation  $X$

sequence of states

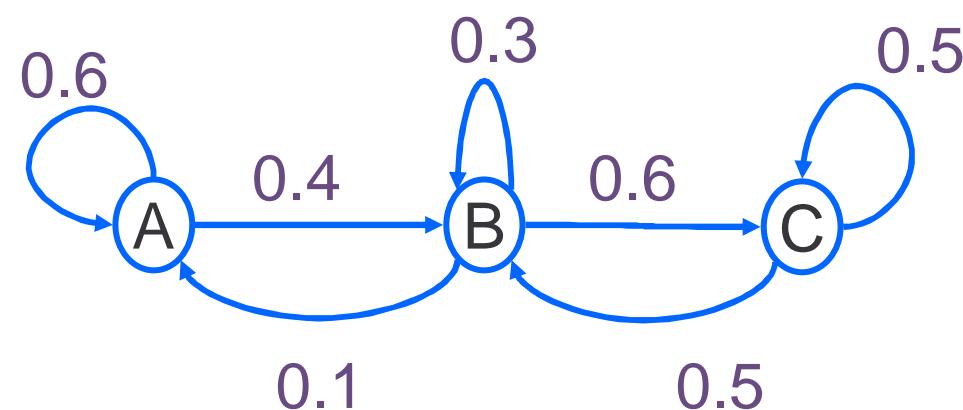
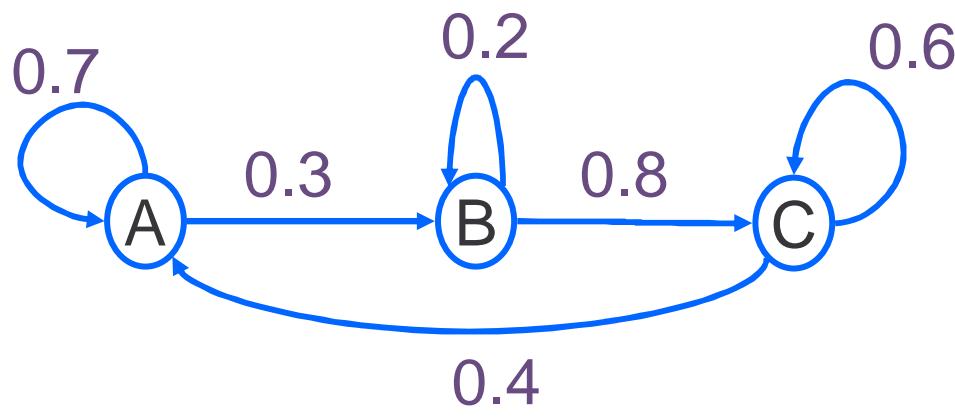
$$X = x_1 x_2 \dots x_n$$

probability

( observation given the model )

$$P(X|M) = p_{x_1} t_{x_1 x_2} t_{x_2 x_3} \dots t_{x_{n-1} x_n} = p_{x_1} \cdot \prod_{i=2}^n t_{x_{i-1} x_i}$$

## example



$$Q = \{ A, B, C \}$$

$$P = (1, 0, 0)$$

unique starting state A

$$T = \begin{pmatrix} .7 & .3 & 0 \\ 0 & .2 & .8 \\ .4 & 0 & .6 \end{pmatrix}$$

$$P( AABBBCCC | M ) =$$

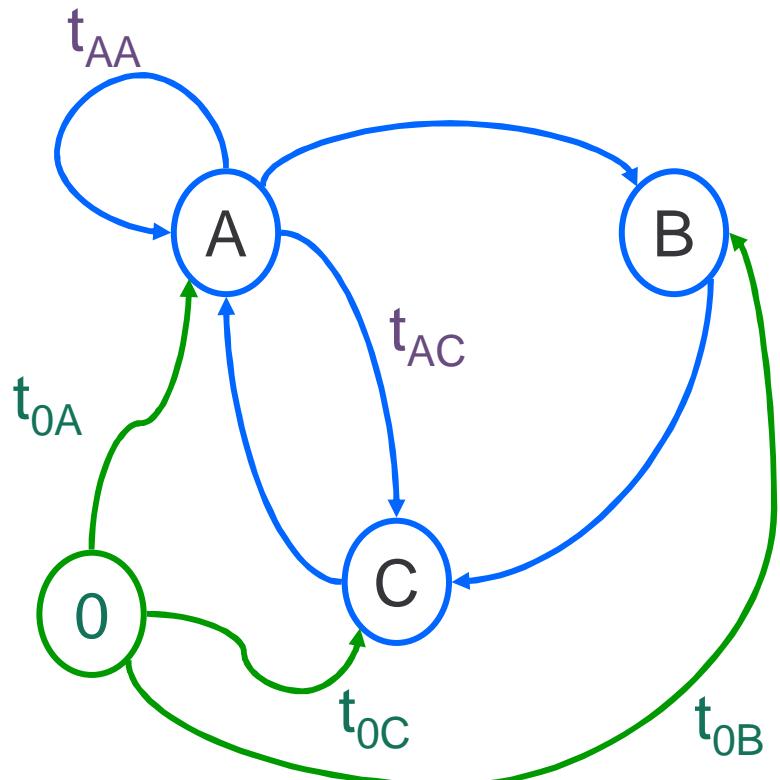
$$\rightarrow^1 A \xrightarrow{.7} A \xrightarrow{.3} B \xrightarrow{.2} B \xrightarrow{.8} C \xrightarrow{.6} C \xrightarrow{.6} C$$

$$1 \cdot 7 \cdot 3 \cdot 2 \cdot 8 \cdot 6 \cdot 6 \cdot 10^{-6} = 1.2 \cdot 10^{-2}$$

vs

$$1 \cdot 6 \cdot 4 \cdot 3 \cdot 6 \cdot 5 \cdot 5 \cdot 10^{-6} = 2.2 \cdot 10^{-2}$$

# Markov model



$$t_{0x} = p_x$$

- initial state  $x_0$  fixed  
~ initial probabilities
- final state [not in this picture]

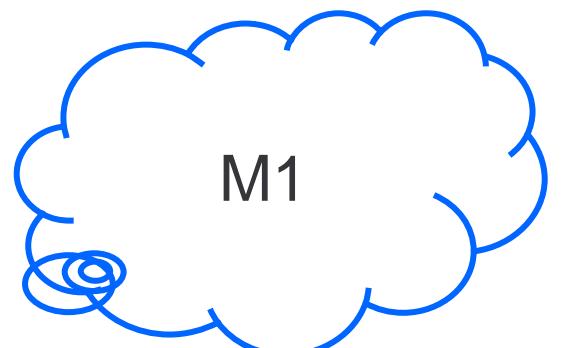
$$X = x_1 x_2 \dots x_n$$

$$P(X|M) = \prod_{i=1}^n t_{x_{i-1} x_i}$$

small values: underflow

$$\log P(X|M) = \sum_{i=1}^n \log t_{x_{i-1} x_i}$$

# comparing models



$$X = x_1 x_2 \dots x_n$$

$$P(X|M) = \prod_{i=1}^n t_{x_{i-1} x_i}$$

best explained by which model?

$$P(X | M1) \text{ vs. } P(X | M2)$$

$$P(M1 | X) \text{ vs. } P(M2 | X) !!$$

Bayes:  $P(A|B) = P(B|A) \cdot P(A) / P(B)$

$$\frac{P(M1|X)}{P(M2|X)} = \frac{P(X|M1) \cdot P(M1)}{P(X|M2) \cdot P(M2)}$$

motto

bases are not random

# application: CpG islands

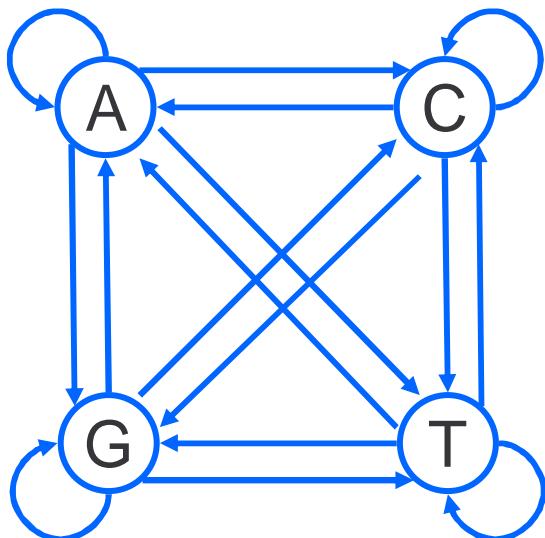
observed  
frequencies

island

island	+	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	

non island

non island	-	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	



consecutive CG pair      CG → TG  
mostly rare, although 'islands' occur  
signal (e.g.) promotor regions

# basic questions

- observation: DNA sequence
  - model: CpG islands / non-islands
- 
- does this sequence belong to a certain family?  
Markov chains  
is this a CpG island (or not)?
- 
- can we say something about the internal structure?  
HMM: Hidden Markov Models  
where are the CpG islands?

# application: CpG islands

+	A	C	G	T	-	A	C	G	T
A	0.180	0.274	0.426	0.120	A	0.300	0.205	0.285	0.210
C	0.171	0.368	0.274	0.188	C	0.322	0.298	0.078	0.302
G	0.161	0.339	0.375	0.125	G	0.248	0.246	0.298	0.208
T	0.079	0.355	0.384	0.182	T	0.177	0.239	0.292	0.292

## score

$$\frac{P(X \mid \text{island})}{P(X \mid \text{non})} = \frac{\prod_{i=1}^n t_{x_{i-1} x_i}^+}{\prod_{i=1}^n t_{x_{i-1} x_i}^-}$$

X = ACGT

$$\frac{0.274 \cdot 0.274 \cdot 0.125}{0.205 \cdot 0.078 \cdot 0.208} = 2.82$$

## application: CpG islands

	LLR	A	C	G	T
log( $t_{xy}^+ / t_{xy}^-$ )	A	-0.74	0.42	0.58	-0.80
'bits' ( $\log_2$ )	C	-0.91	0.30	1.81	-0.69
	G	-0.62	0.46	0.33	-0.73
	T	-1.17	0.57	0.39	-0.68

log-score

$$\log \frac{P(X \mid \text{island})}{P(X \mid \text{non})} = \log \frac{\prod_{i=1}^n t_{x_{i-1} x_i}^+}{\prod_{i=1}^n t_{x_{i-1} x_i}^-} = \sum_{i=1}^n \log \left( \frac{t_{x_{i-1} x_i}^+}{t_{x_{i-1} x_i}^-} \right)$$

X = ACGT

$$\log \frac{0.274 \cdot 0.274 \cdot 0.125}{0.205 \cdot 0.078 \cdot 0.208} = 0.42 + 1.81 - 0.73 = 1.50$$

# CpG Log-Likelihood Ratio

$\log(t_{xy}^+ / t_{xy}^-)$

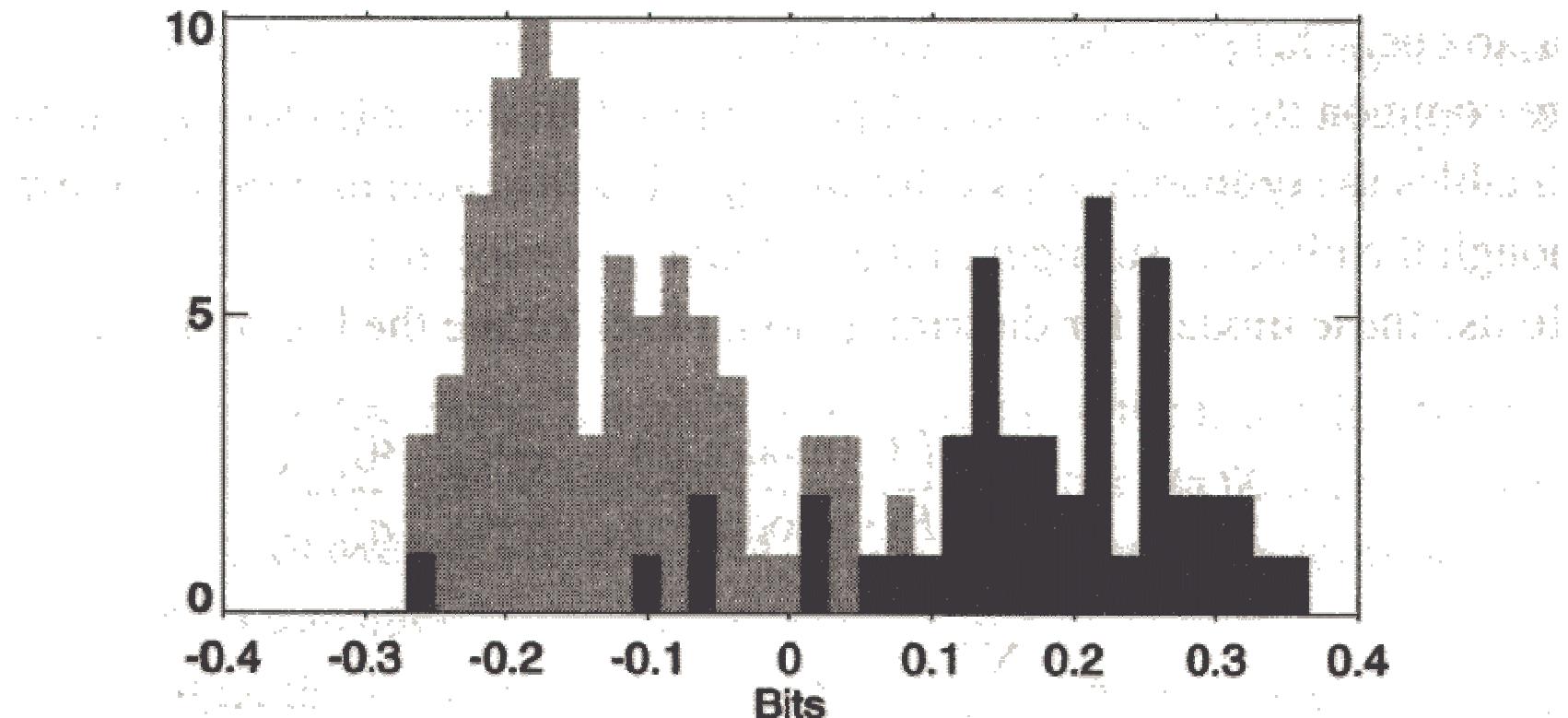
LLR	A	C	G	T
A	-0.74	0.42	0.58	-0.80
C	-0.91	0.30	1.81	-0.69
G	-0.62	0.46	0.33	-0.73
T	-1.17	0.57	0.39	-0.68

$$\text{LLR(ACGT)} = 0.42 + 1.81 - 0.73 = 1.50 \quad (0.37 \text{ per base})$$

- is a (short) sequence a CpG island ?  
compare with observed data (normalized for length)
- where (in long sequence) are CpG islands ?  
first approach: *sliding window*  
length of window?

# empirical data

- is a (short) sequence a CpG island ?  
compare with observed data (normalized for length)



**Figure 3.2** The histogram of the length-normalised scores for all the sequences. CpG islands are shown with dark grey and non-CpG with light grey.

- where (in long sequence) are CpG islands ?  
first approach: *sliding window*

## CpGplot

**EMBL-EBI European Bioinformatics Institute**

Get Nucleotide sequences for

EBI Home About EBI Research Services Toolbox Databases Downloads Submission SEQUENCE ANALYSIS

**EMBOSS CpGPlot/CpGReport/Isochore**

Detection of regions of genomic sequences that are rich in the CpG pattern is important because such regions are resistant to methylation and tend to be associated with genes which are frequently switched on. Regions rich in the CpG pattern are known as CpG islands. The function of the program [cpgplot](#) is to plot CpG rich areas, and [cpgreport](#) to report all CpG rich regions.

The nuclear genomes of vertebrates are mosaics of isochores, very long stretches of DNA that are homogeneous in base composition and are compositionally correlated with the coding sequences that they embed. Isochores can be partitioned in a small number of families that cover a range of GC levels. Program [isochores](#) plots GC content over a sequence.

Program	Window	Step	Obs/Exp	MinPC	Length	Reverse	Complement
cpgplot	100	1	0.6	50	50	no	no

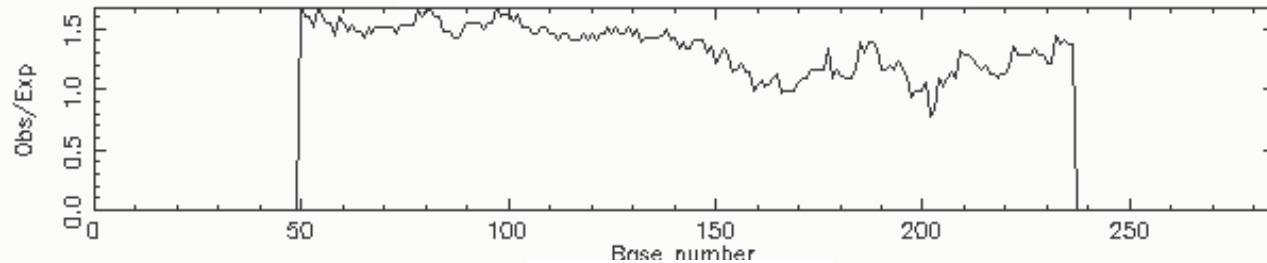
Enter or Paste a nucleic acid Sequence (at least 100bp) in any format:  Help

```
ACCGATACGATGAGAATGAGCAATGTAGTGAATCGTTTAGCTACT
CTCTATCGTAGCATTACTATGCAGTCAGTGATGCGCGCTAGCCGCG
TAGCTCGCGTCGCATCGCTGGCCGTAGCTGCGTACGATCTGCTGT
ACGCTGATCGGAGCGCTGCATCTCAACTGACTCATACTCATATGTC
TACATCATCATCATTGATGTCAGTCTAGCATACTATTATGACGAC
TGATCGATCTGACTGCTAGTAGACGTACCGAGCCAGGCATACGACA
TCAGTCGACT
```

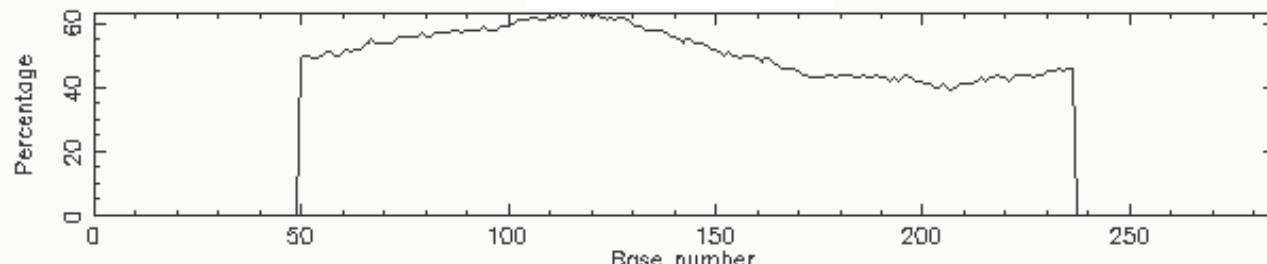
Upload a file:  Bläderen...

# CpGplot

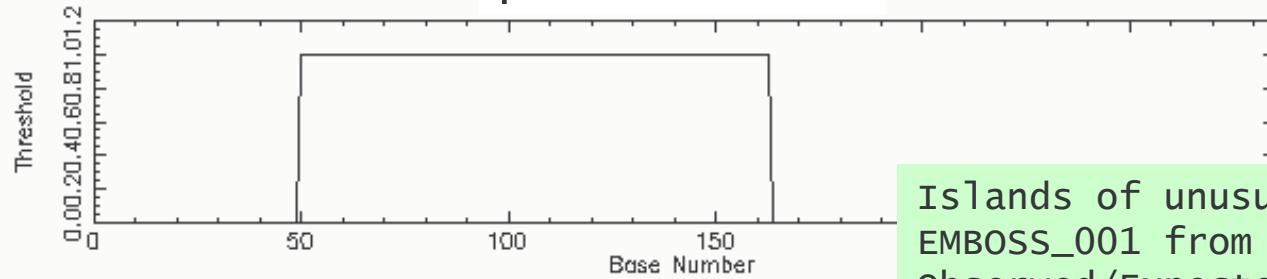
observed vs. expected



percentage



putative islands

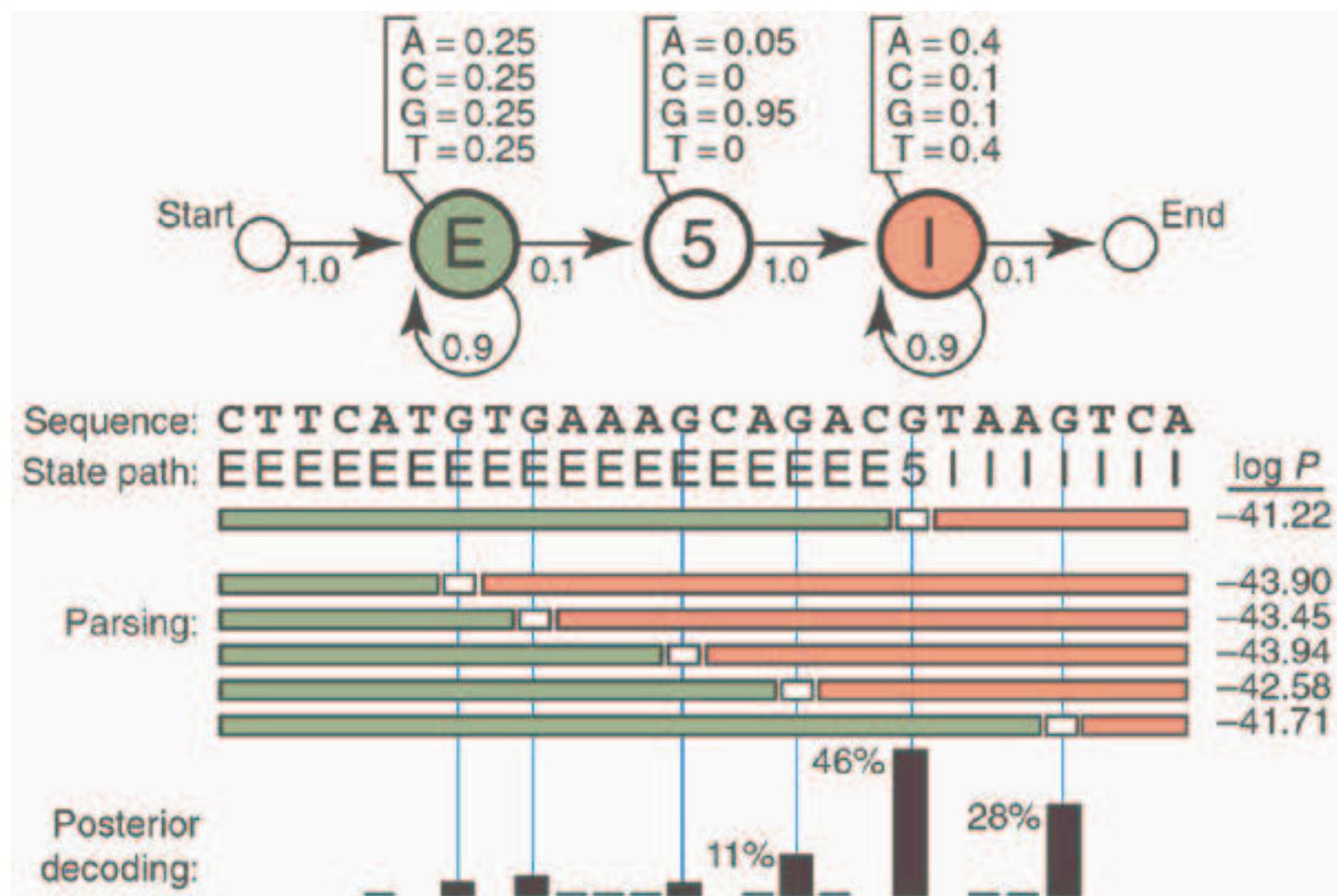


Islands of unusual CG composition  
EMBOSS\_001 from 1 to 286  
Observed/Expected ratio > 0.60  
Percent C + Percent G > 50.00  
Length > 50  
Length 114 (51..164)

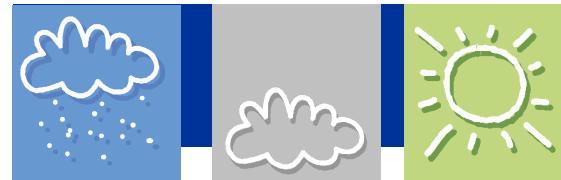
# hidden Markov model

- where (in long sequence) are CpG islands ?  
second approach: *hidden Markov model*

# Eddy (2004)



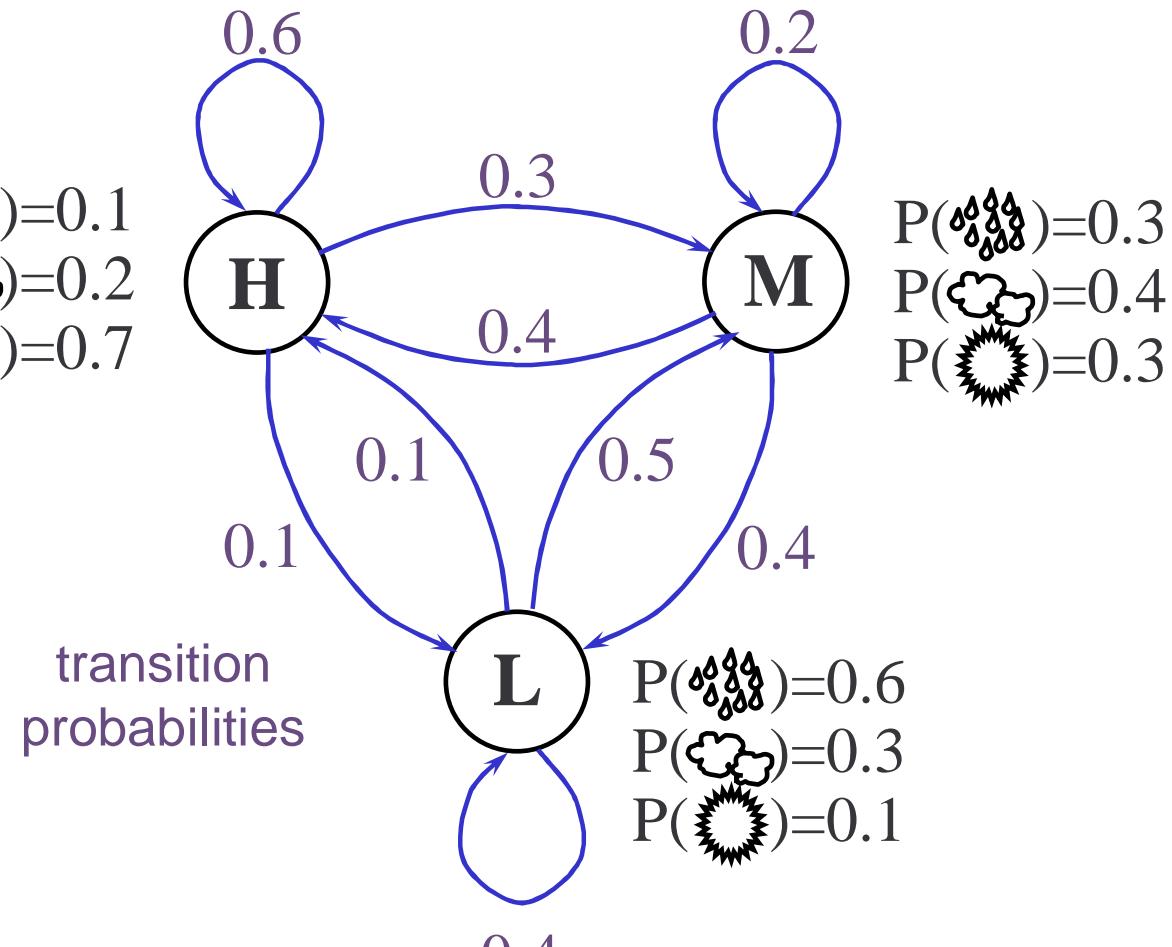
# What is a hidden Markov model? Sean R Eddy *Nature Biotechnology* 22, 1315 - 1316 (2004)



weather

emission  
probabilities

$$\begin{aligned}P(\text{rain}) &= 0.1 \\P(\text{clouds}) &= 0.2 \\P(\text{sun}) &= 0.7\end{aligned}$$

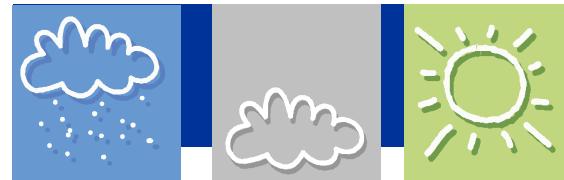


$$\begin{cases} p_H = 0.4 \\ p_M = 0.2 \\ p_L = 0.4 \end{cases}$$

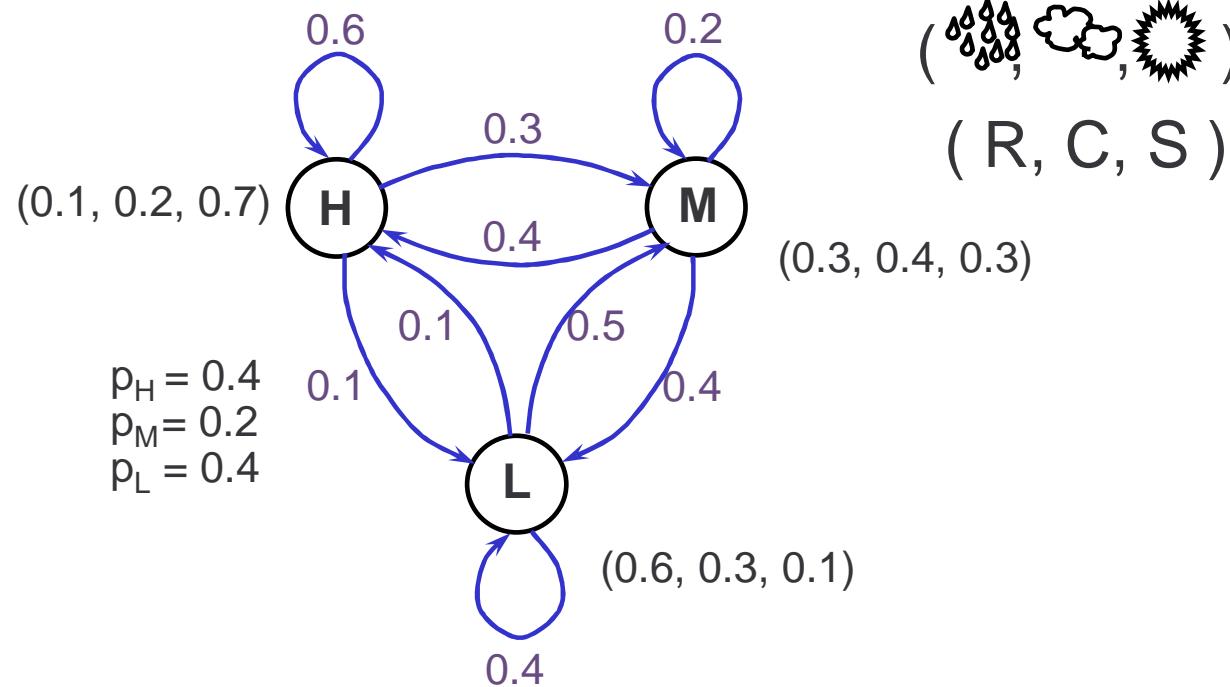
initial  
probabilities

$$\begin{aligned}P(\text{rain}) &= 0.3 \\P(\text{clouds}) &= 0.4 \\P(\text{sun}) &= 0.3\end{aligned}\quad \begin{aligned}P(\text{rain}) &= 0.6 \\P(\text{clouds}) &= 0.3 \\P(\text{sun}) &= 0.1\end{aligned}$$

observed weather vs. pressure



weather



$$P( \text{ RCCSS} | \text{HHHHH} ) = 1 \cdot 2 \cdot 2 \cdot 7 \cdot 7 = 196 \ (\times 10^{-5})$$

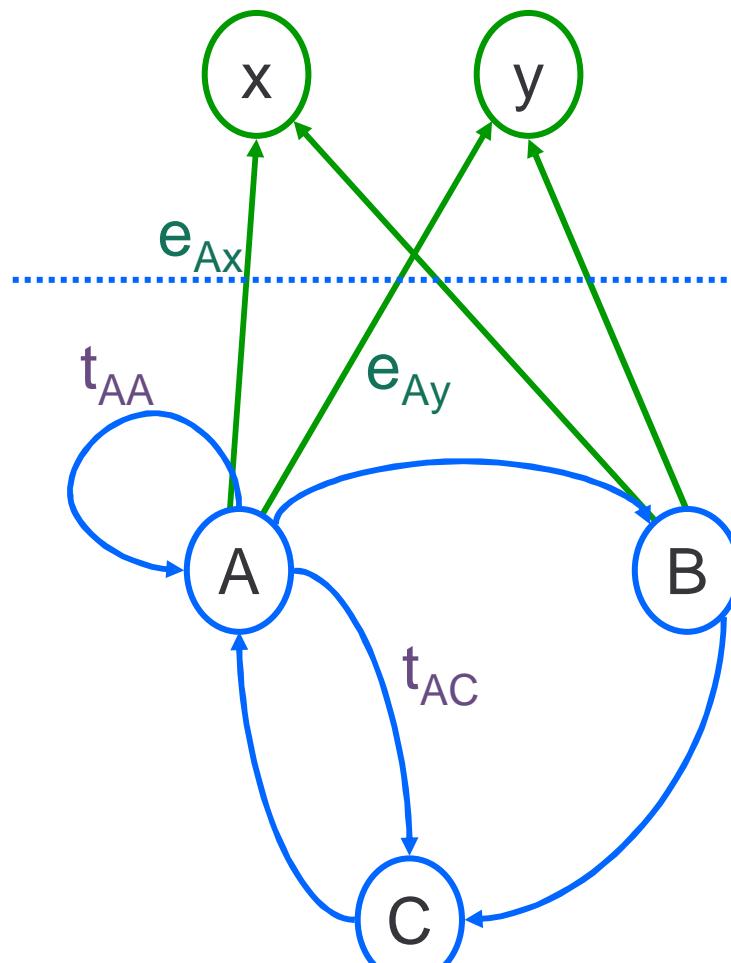
$$P( \text{ RCCSS} | \text{MMMMM} ) = 3 \cdot 4 \cdot 4 \cdot 3 \cdot 3 = 432 \ (\times 10^{-5})$$

$$P( \text{ RCCSS}, \text{HHHHH} ) = 4 \cdot 1 \cdot 6 \cdot 2 \cdot 6 \cdot 2 \cdot 6 \cdot 7 \cdot 6 \cdot 7 = 1016 \ (\times 10^{-7})$$

$$P( \text{ RCCSS}, \text{MMMMM} ) = 2 \cdot 3 \cdot 2 \cdot 4 \cdot 2 \cdot 4 \cdot 2 \cdot 3 \cdot 2 \cdot 3 = 14 \ (\times 10^{-7})$$

# hidden Markov model

what we see



underlying process

**model**  $M = (\Sigma, Q, T)$

- states  $Q$
- transition probabilities  $t_{pq}$ ,  $p, q \in Q$

**observation**  $X = x_1 x_2 \dots x_n \in \Sigma^*$

observe states *indirectly* ‘hidden’

- emission probabilities

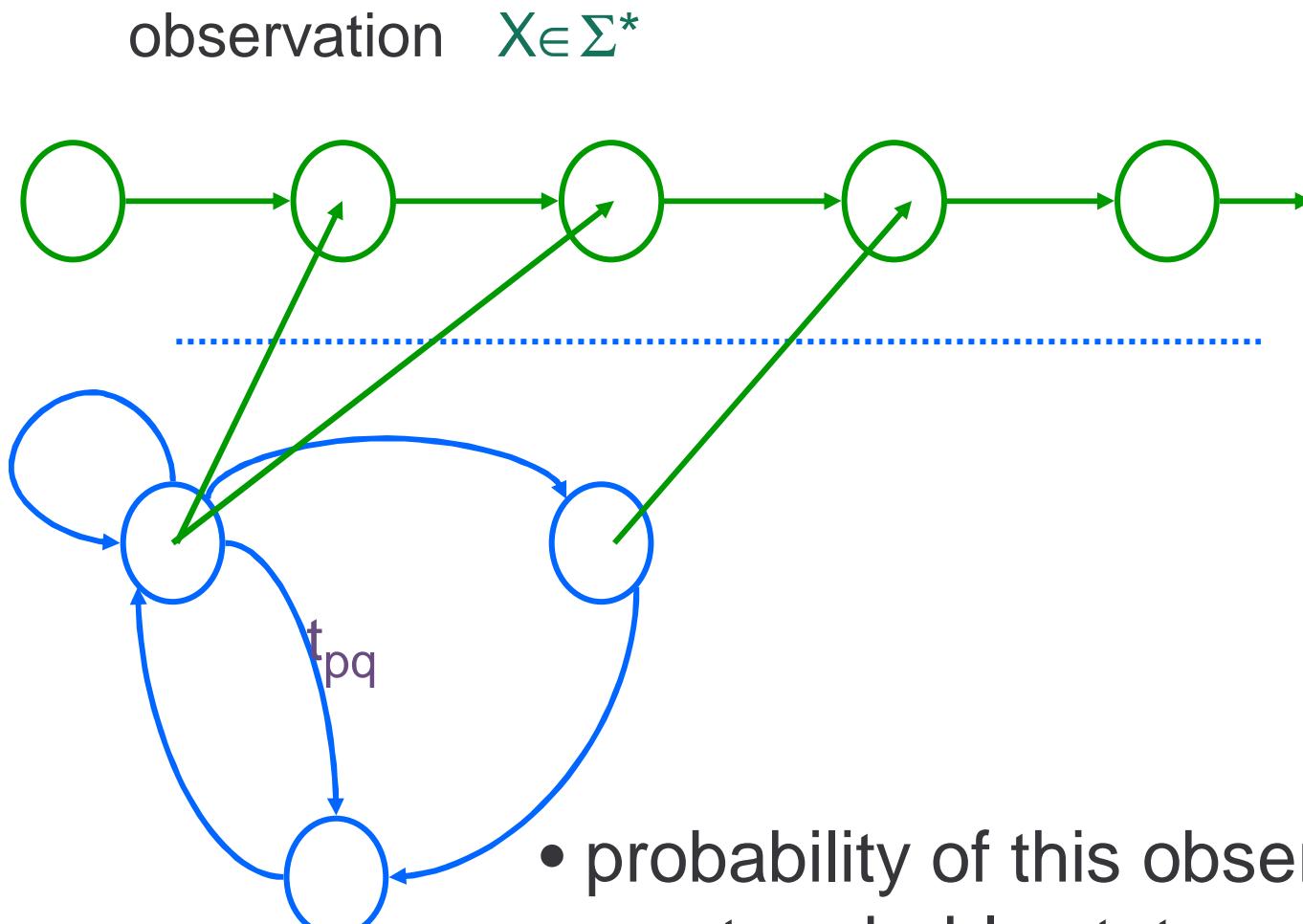
$$e_{px}, p \in Q, x \in \Sigma \quad e_p(x)$$

**probability**

observation given the model

? there may be *many* state seq's

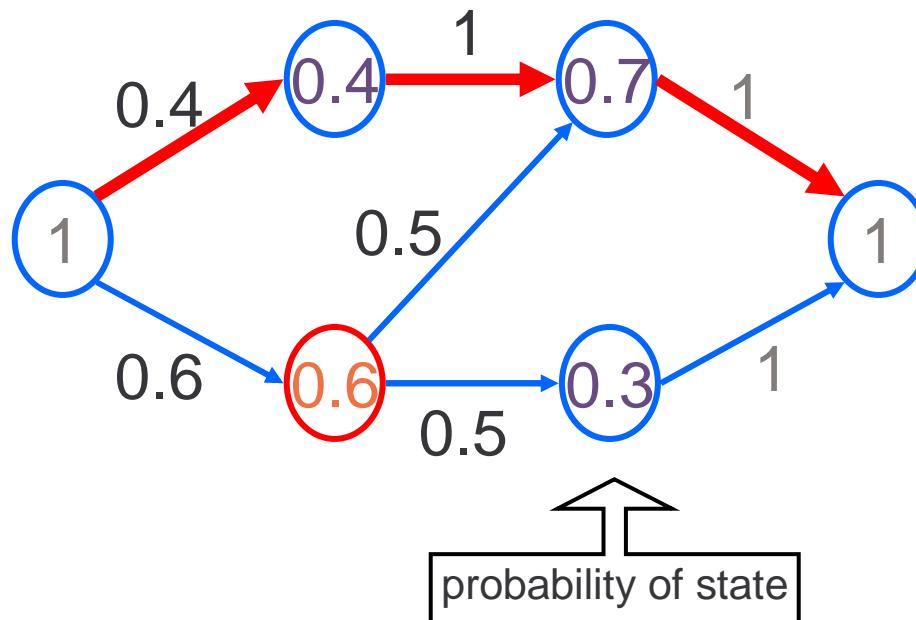
# HMM main questions



- probability of this observation?
- most probable state sequence?
- how to find the model? *training*

# probability ... !

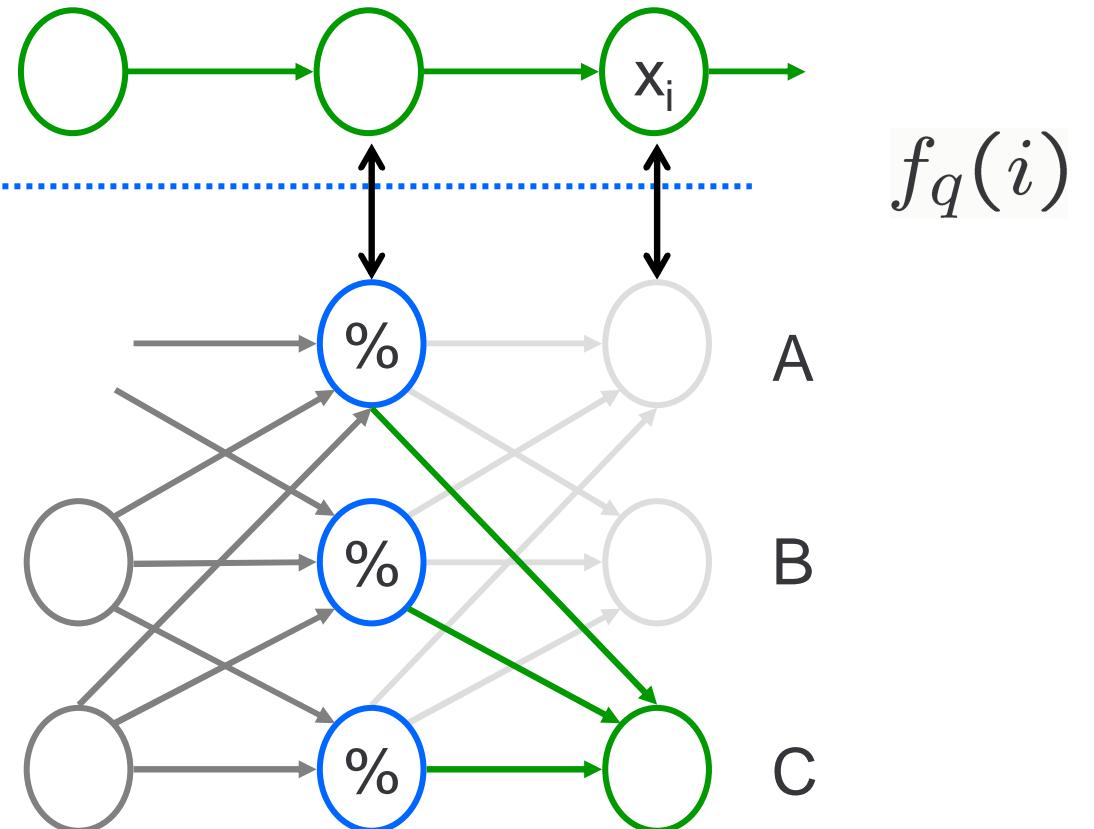
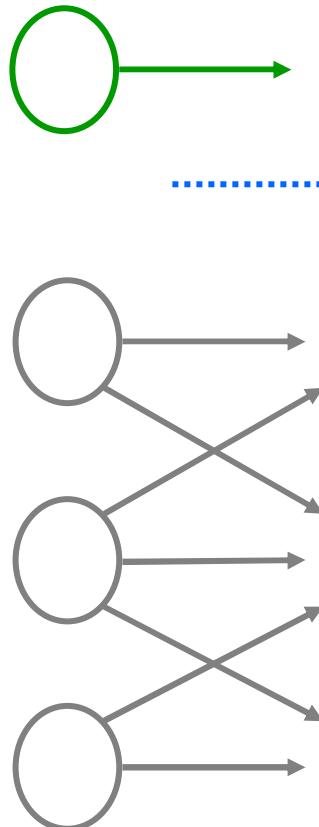
most probable state vs. optimal path



- \* most probable state (over all state sequences)  
**posterior decoding**  
forward & backward probabilities
- \* most probable path (= single state sequence)  
**Viterbi**

# probability of observation

*dynamic programming*: probability ending in state



$$f_q(i) = \sum_{p \in Q} f_p(i-1) t_{pq} e_q(x_i)$$

## probability of observation

probability ending in state

$$f_q(i) = P(x_1 \dots x_i, \pi_i = q)$$

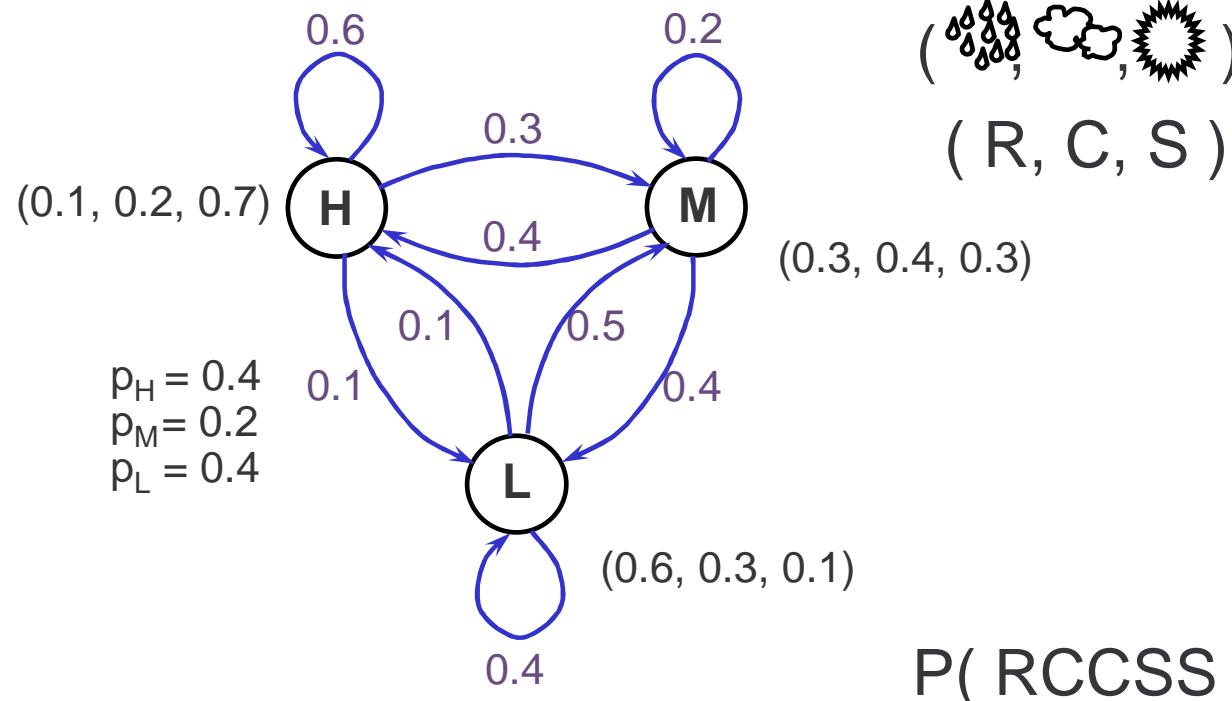
$$f_q(i) = \sum_{p \in Q} f_p(i-1) t_{pq} e_q(x_i)$$

'forward' probability

$$P(X) = \sum_{p \in Q} f_p(n) t_{p*}$$



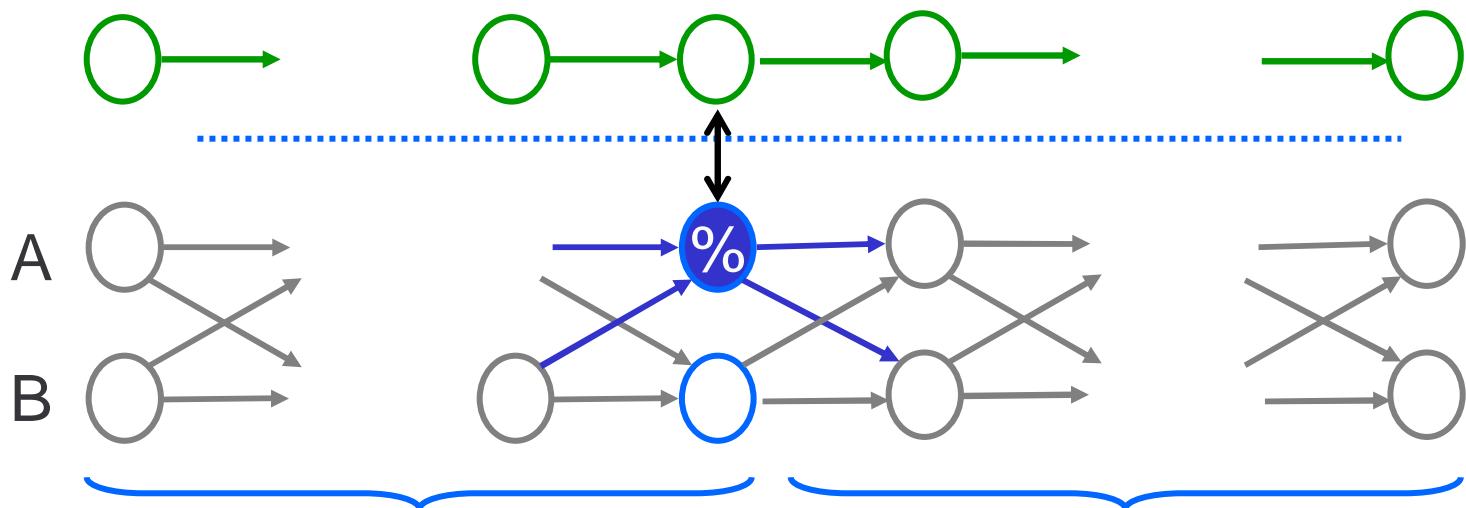
weather



		1:R	2:C	
H	0	$4 \cdot 1 = 4$	$(4 \cdot 6 + 6 \cdot 4 + 24 \cdot 1) \cdot 2 = 144$	$(\times 10^{-4})$
M	0	$2 \cdot 3 = 6$	$(4 \cdot 3 + 6 \cdot 2 + 24 \cdot 5) \cdot 4 = 576$	$(\times 10^{-4})$
L	0	$4 \cdot 6 = 24$	$(4 \cdot 1 + 6 \cdot 4 + 24 \cdot 4) \cdot 3 = 372$	$(\times 10^{-4})$
0	1			

# posterior decoding

$P(\pi_i = q \mid X)$  i-th state equals q



$$f_q(i) = P(x_1 \dots x_i, \pi_i = q)$$

$$b_q(i) = P(x_{i+1} \dots x_n \mid \pi_i = q)$$

forward

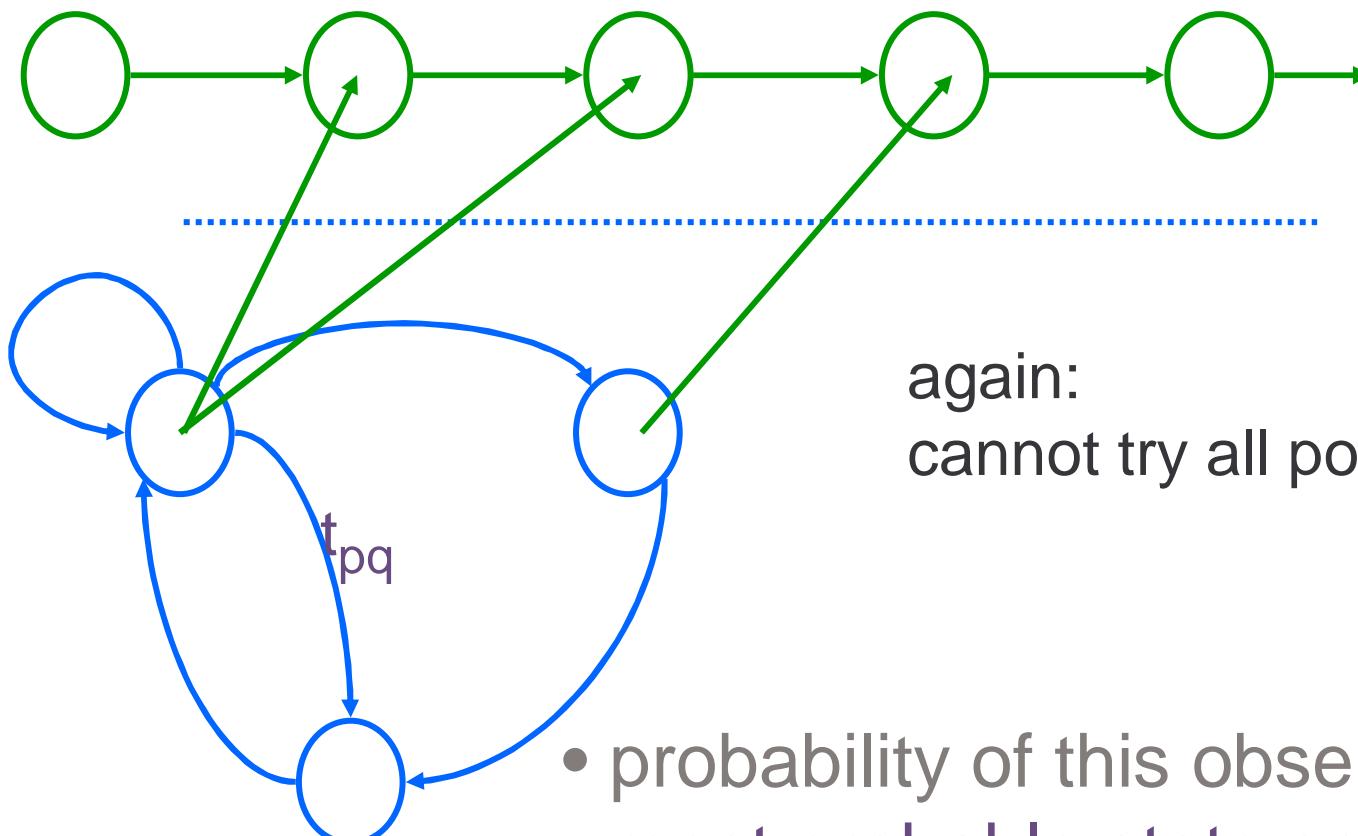
backward

$$P(X, \pi_i = q) = f_q(i)b_q(i)$$

$$P(\pi_i = q \mid X) = \frac{f_q(i)b_q(i)}{P(X)}$$

# HMM main questions

observation  $X \in \Sigma^*$   $\Rightarrow$  most probable state sequence



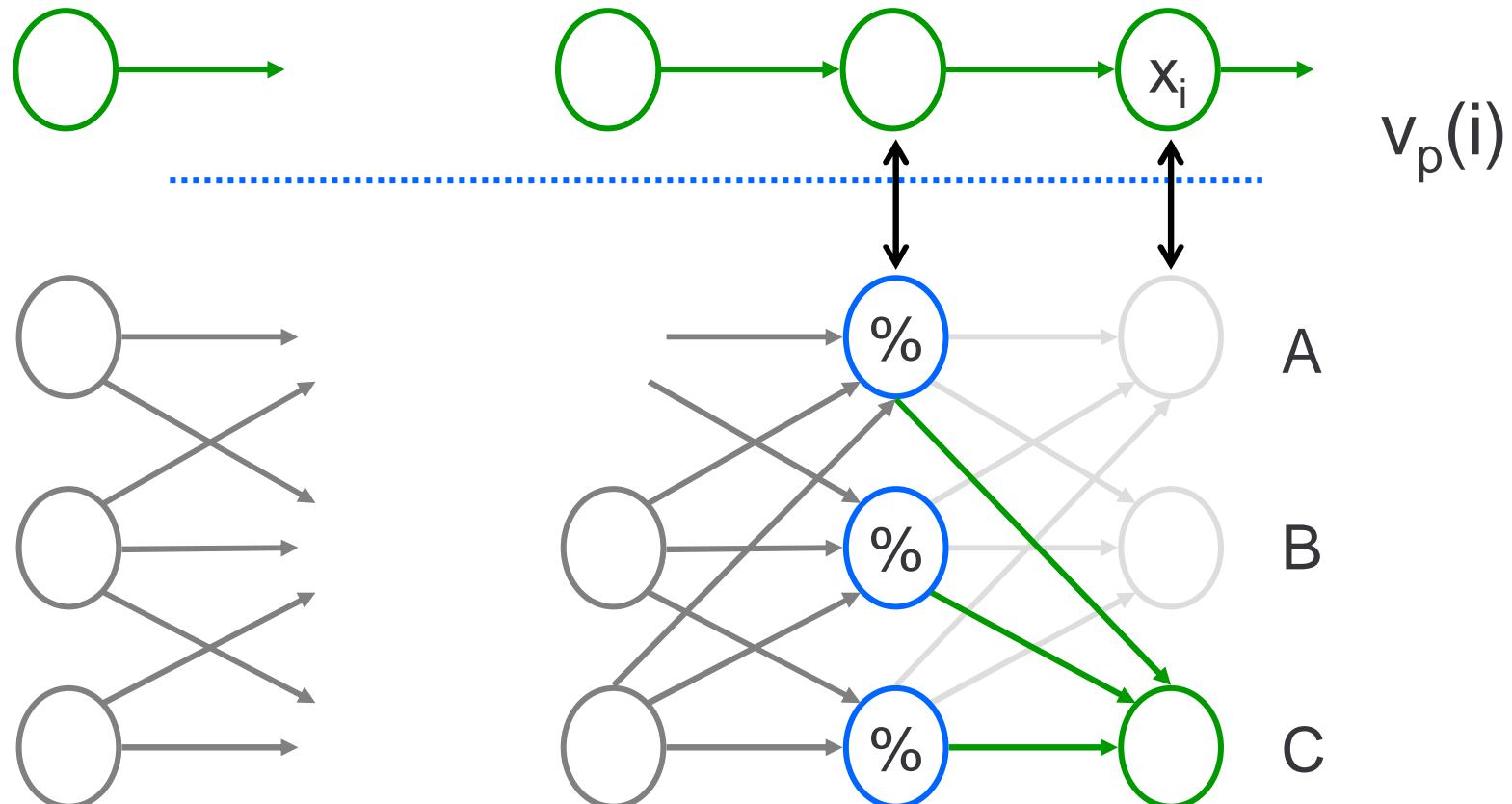
again:  
cannot try all possibilities  
Viterbi

- probability of this observation?
- most probable state sequence?
- how to find the model? *training*

# Viterbi algorithm

most probable state sequence for observation

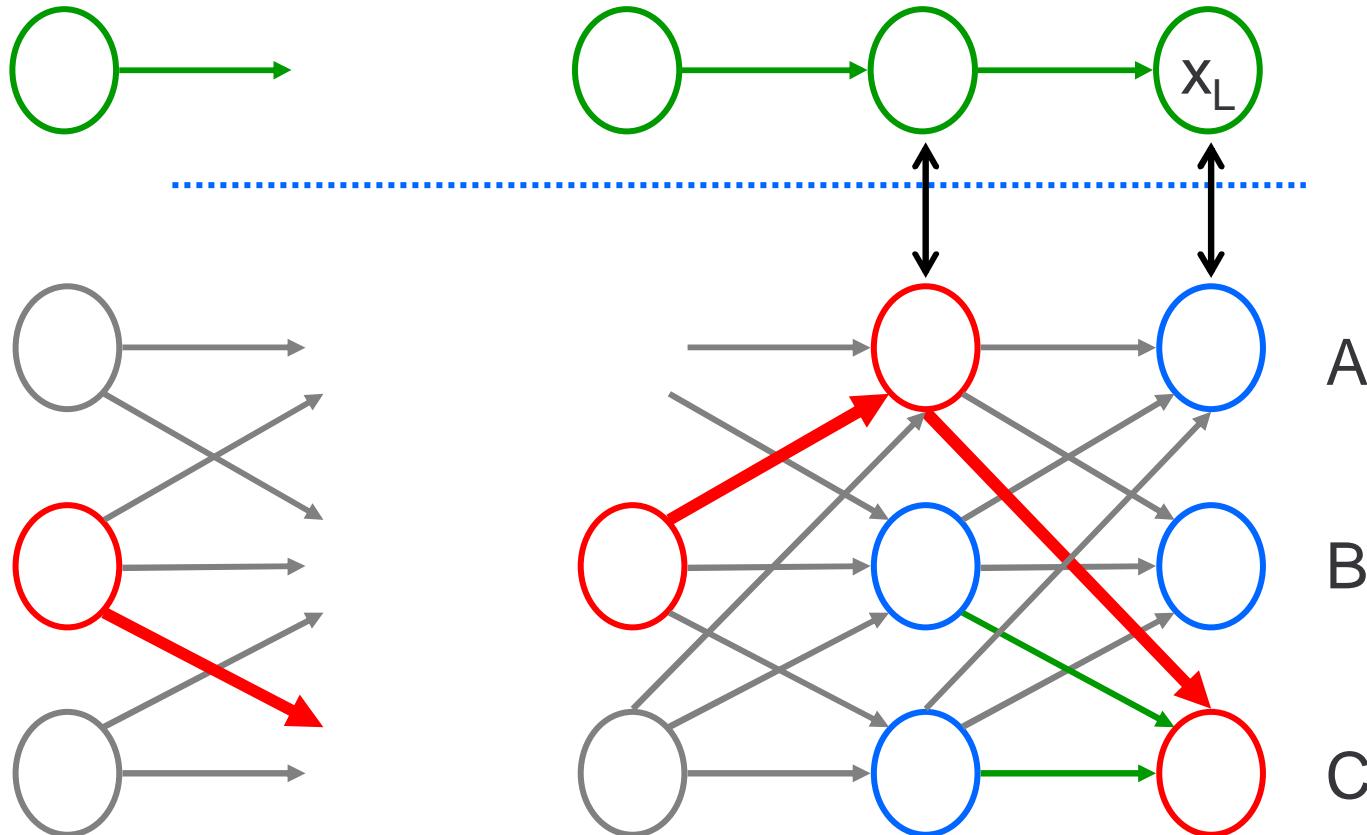
(1) *dynamic programming*: probability ending in state



$$v_q(i) = \max_{p \in Q} v_p(i-1) t_{pq} e_q(x_i)$$

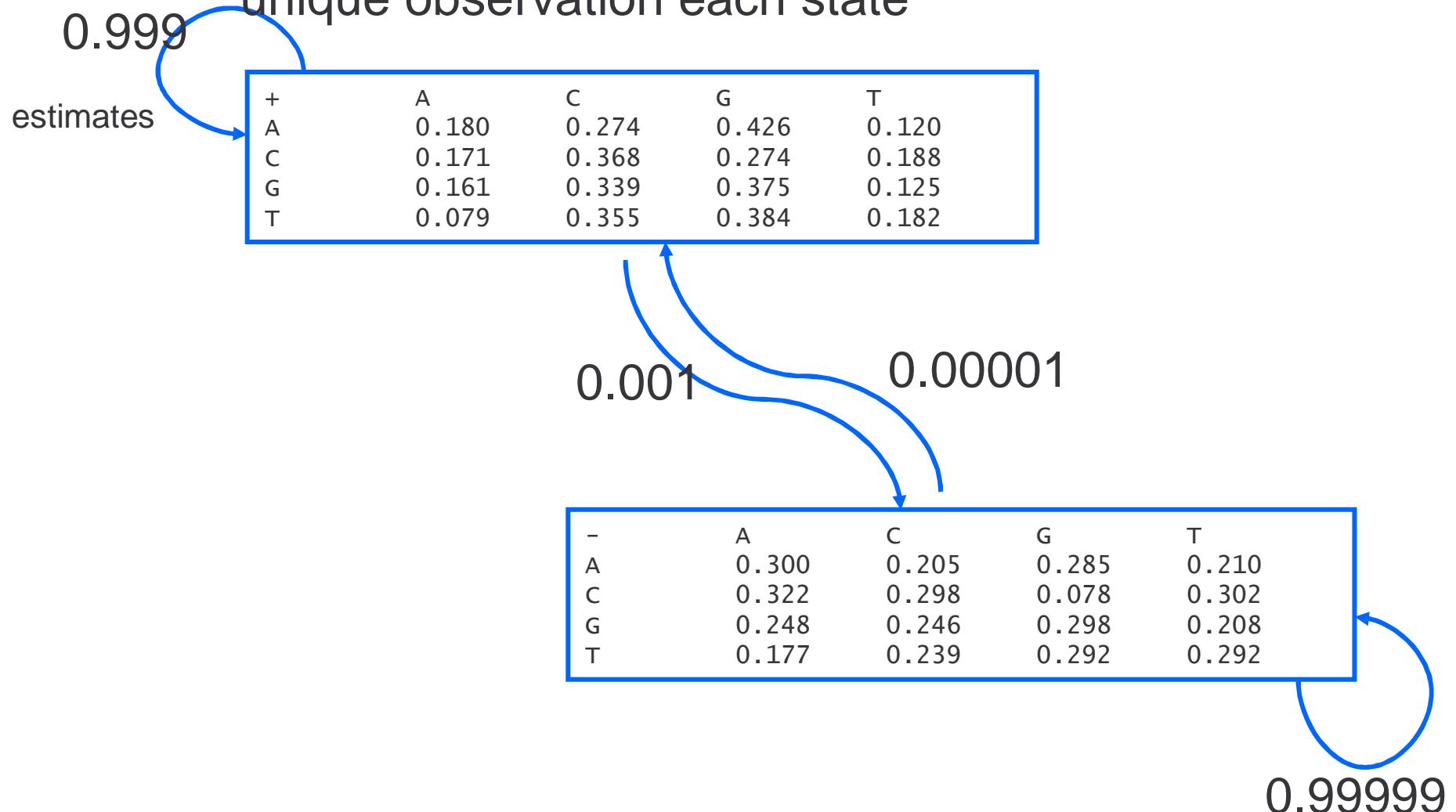
# Viterbi algorithm

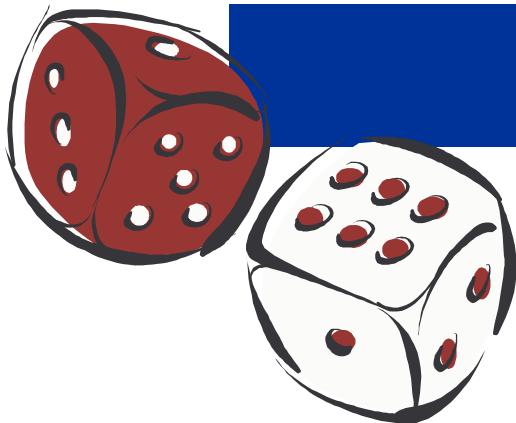
(2) *traceback*: most probable state sequence  
start with final maximum



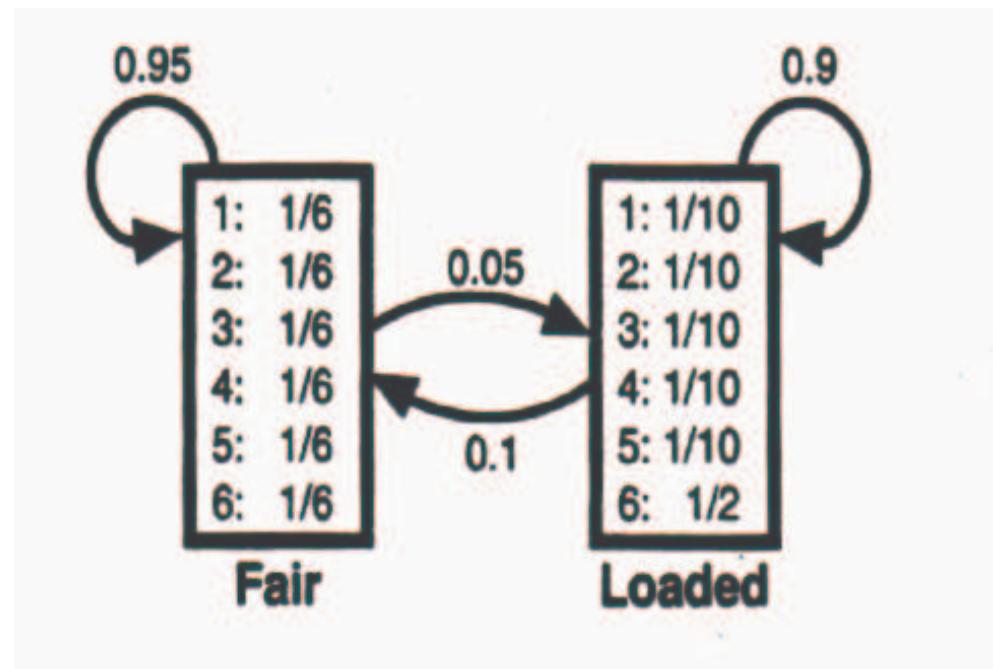
# CpG islands ctd.

8 states A<sup>+</sup> vs A<sup>-</sup>  
unique observation each state





## dishonest casino dealer



# dishonest casino dealer

Rolls	315116246446644245321131631164152133625144543631656626566666
Die	FFFLLLLLLLLLLLLLLL
Viterbi	FFFLLLLLLLLLLL
Rolls	65116645313265124563666463163666316232645523526666625151631
Die	LLLLLFFFFFFFFFLLLLLLLLLLLFFFLLLLLLLLLLFFFLLLLLLFFFF
Viterbi	LLLLLFFFFFFFFFLLLLLLLLLLLFFFLLLLLLFFFLLLLLLFFFL
Rolls	222555441666566563564324364131513465146353411126414626253356
Die	FFFFFFFLLLLLLLFFFLFFFFFFFLLLLLFFFL
Viterbi	FFFFFFFLLLLLFFFLFFFFFFFLLLLLFFFLFFFFFFFLLLLLFFFL
Rolls	366163666466232534413661661163252562462255265252266435353336
Die	LLLLLLLFFFFFFFFFLLLLLFFFLFFFFFFFLLLLLFFFL
Viterbi	LLLLLLLFFFLFFFFFFFLLLLLFFFLFFFFFFFLLLLLFFFL
Rolls	233121625364414432335163243633665562466662632666612355245242
Die	FFFFFFFLLLLLFFFLFFFFFFFLLLLLFFFL
Viterbi	FFFFFFFLLLLLFFFLFFFFFFFLLLLLFFFL

# dishonest casino dealer

## Observation

36616366646623253441366166116325256246225526525226643535336

# viterbi

LLLLLLLLLLLL FFF

## Forward

FFLLLLLLLLLFFF  
FFFFFFFFFFLFL  
LLLFLFFF  
FFFFFFFFFFF  
FFFFFFFFFFLFFF  
FFFFFFFFFF

## Posterior (total)

LLLLLLLLLLLLLFFFFFFFFFLLLLLLFFFFFFFFFLLLLLFFFFFFFFF

# Parameter estimation

training sequences  $X^{(i)}$

optimize score  $\prod_{i=1}^n P(X^{(i)} | \Theta)$

*state sequences known*

- count transitions  $pq$   $A_{pq}$
- count emissions  $b$  in  $p$   $E_p(b)$

divide by

- total transitions in  $p$
- emissions in  $q$

Laplace correction

# Baum-Welch

*state sequences unknown*

## Baum-Welch training

based on model

expected number of transitions, emissions

build new (better) model & iterate

$$P(\pi_i = p, \pi_{i+1} = q \mid X, \Theta) =$$

$$\frac{f_p(i) \cdot t_{pq} \cdot e_q(x_{i+1}) \cdot b_q(i+1)}{P(X)}$$

$A_{pq}$  sum over all training sequences X  
sum over all positions i

$E_p(b)$  sum over all training sequences X  
sum over all positions i with  $x_i = b$

# Baum-Welch training

concerns:

- guaranteed to converge target score, not  $\Theta$
- unstable solutions !
- local maximum

tips:

- repeat for several initial  $\Theta$
- start with meaningful  $\Theta$

**Viterbi training** (an alternative)

determine optimal paths

recompute as if paths known

- score may decrease!