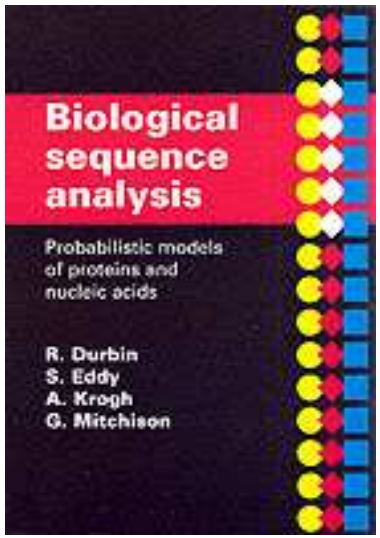
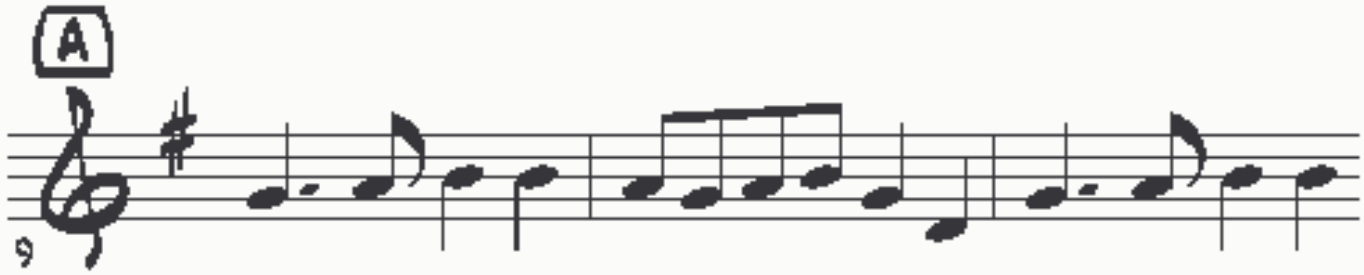
Hidden Markov Models

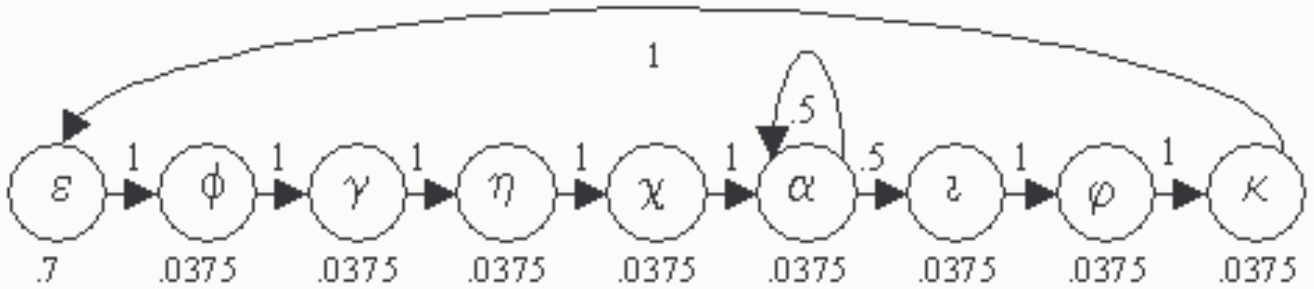
based on chapters from the book
Durbin, Eddy, Krogh and Mitchison
Biological Sequence Analysis
via Shamir's lecture notes



music recognition

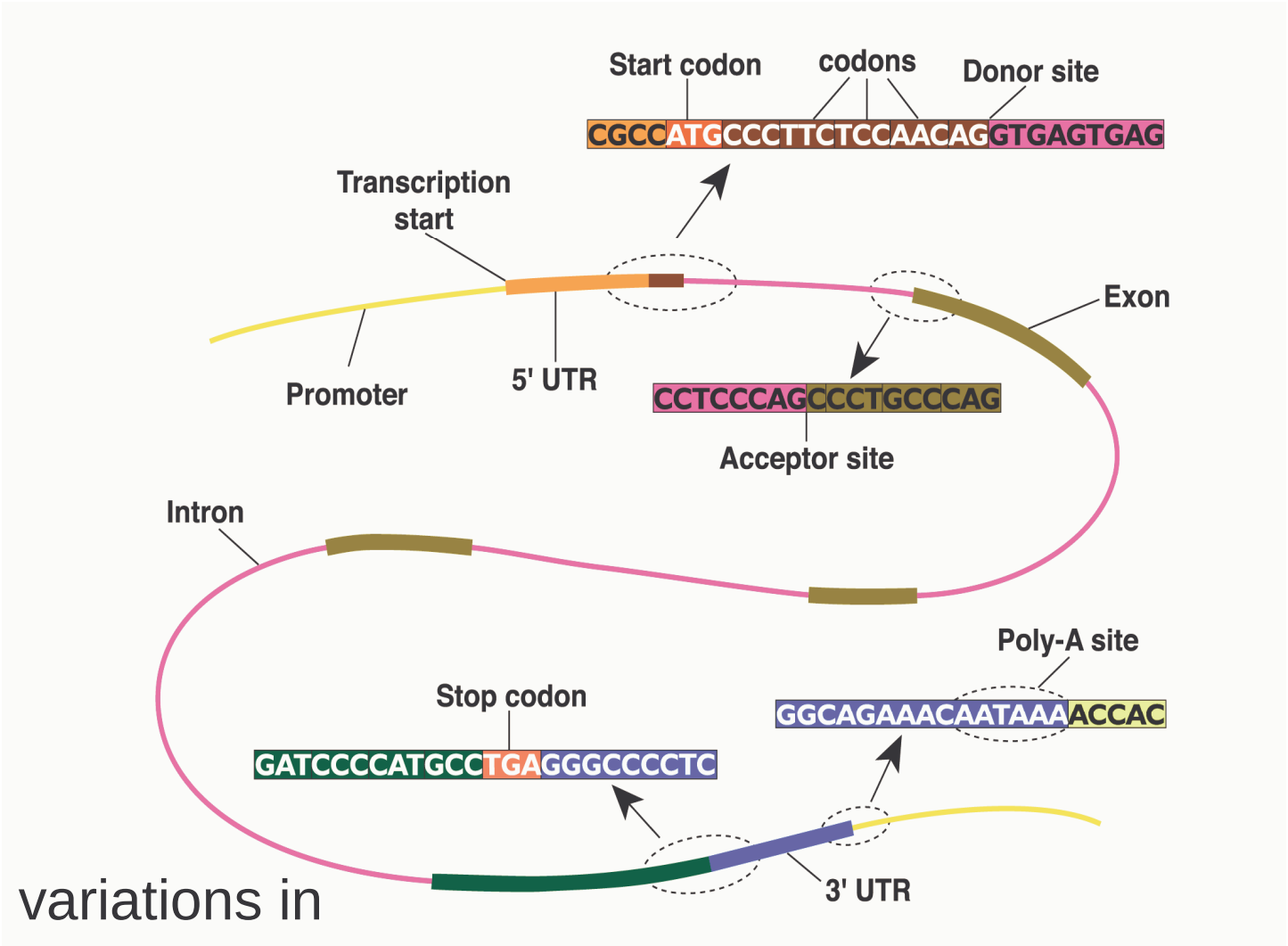


Delta Pitch:	2	2	0	-2	-2	2	2	-4	-5	5	2	2	0
IOI:	3	1	2	2	1	1	1	1	2	2	3	1	2
IOI ratio:	3	.5	1	2	1	1	1	.5	1	.66	3	.5	1
State:	ε	ϕ	γ	η	χ	α	α	ι	φ	κ	ε	ϕ	γ



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’

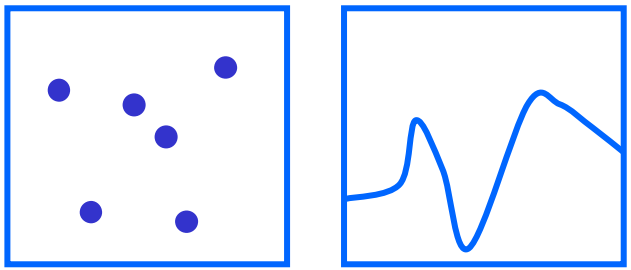
• does this sequence belong to a certain family?

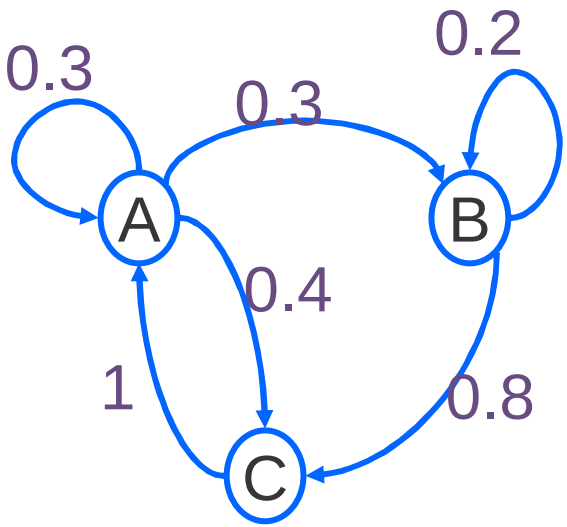
Markov chains

• can we say something about the internal structure? (indirect observation)

HMM: Hidden Markov Models

introduction





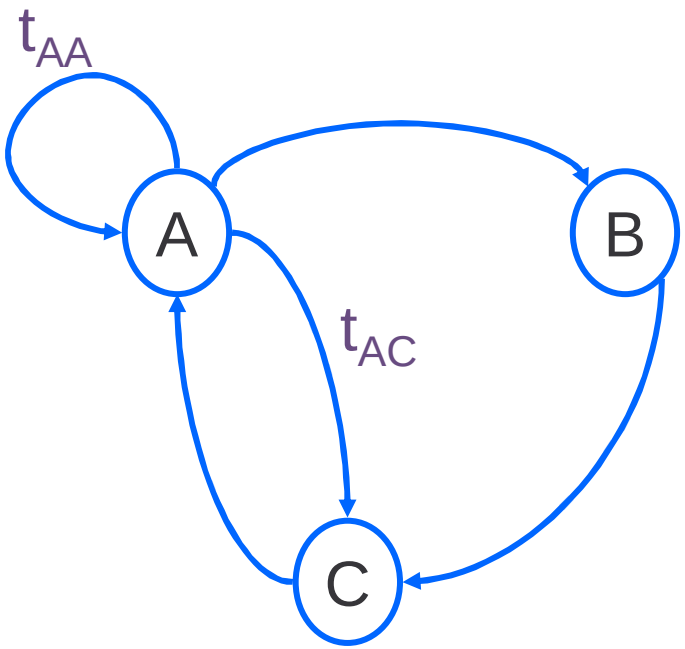
discrete time
discrete space

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

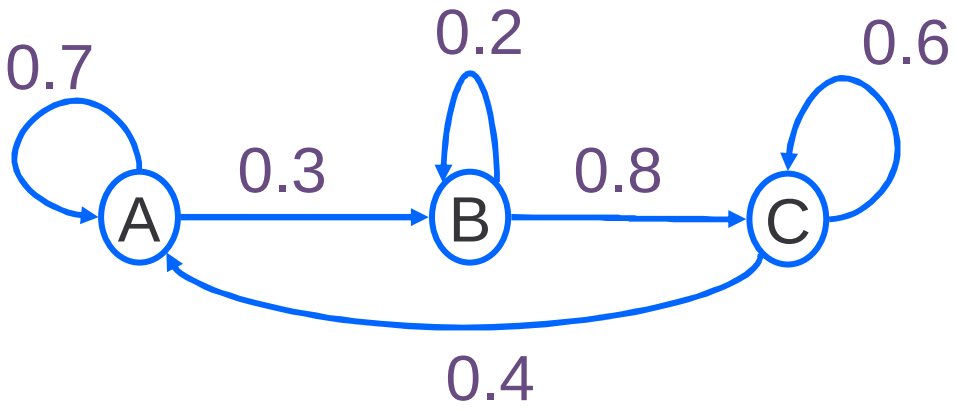
$$X = x_1 x_2 \dots x_n$$

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

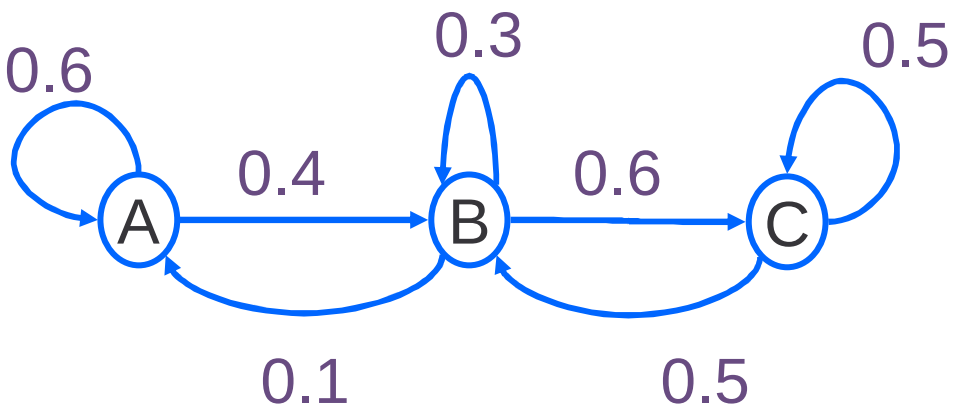


Q = { A, B, C }

P = (1, 0, 0)

unique starting state A

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



P(AABBCCC | M) =

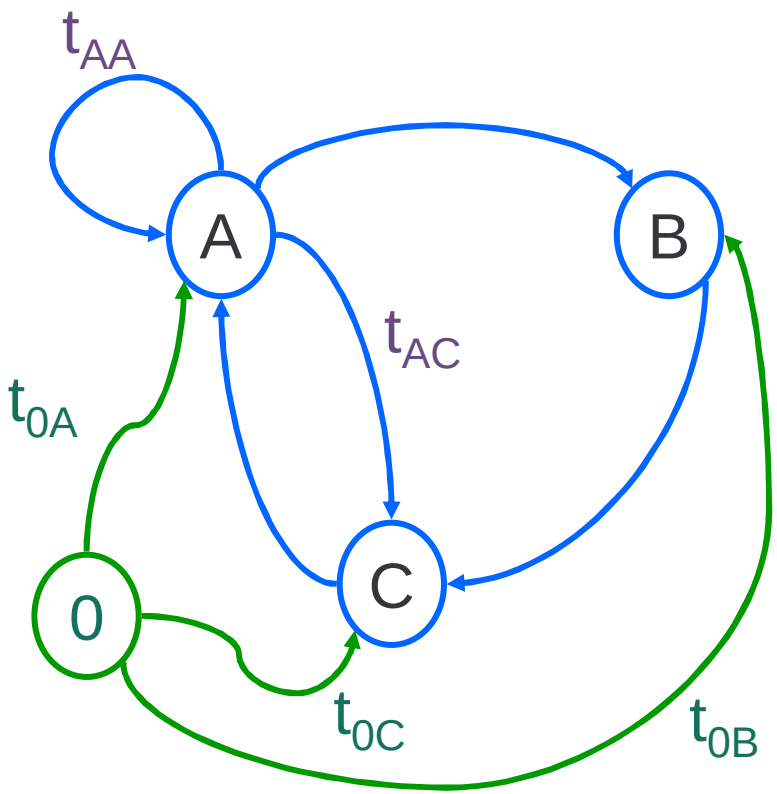
$$\xrightarrow{1} A \xrightarrow{.7} A \xrightarrow{.3} B \xrightarrow{.2} B \xrightarrow{.8} C \xrightarrow{.6} C \xrightarrow{.6} C$$

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

vs

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

Markov model



- 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

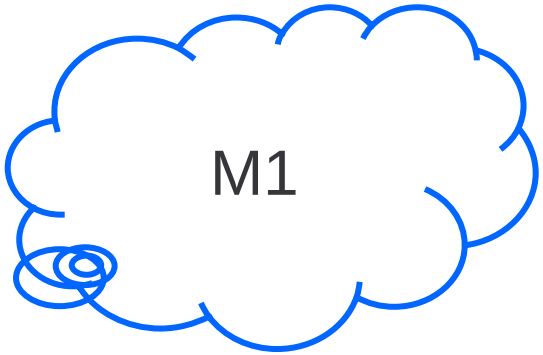
$$\log P(X|M) = \sum_{i=1}^{n} \log t_{x_{i-1}x_i}$$

$$t_{0x} = p_x$$

comparing models

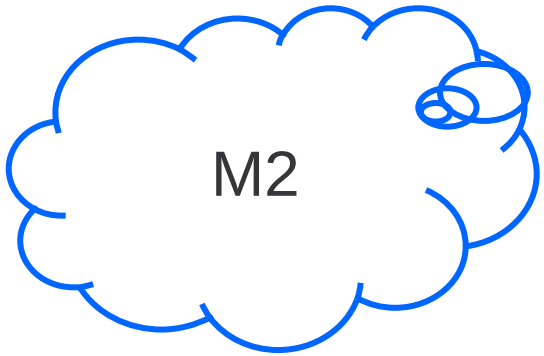
$$X = x_1 x_2 \ldots x_n$$

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



best explained by which model?

P(X | M1) vs. P(X | M2)



P(M1 | X) vs. P(M2 | X) !!

Bayes: P(A|B) = P(B|A)·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

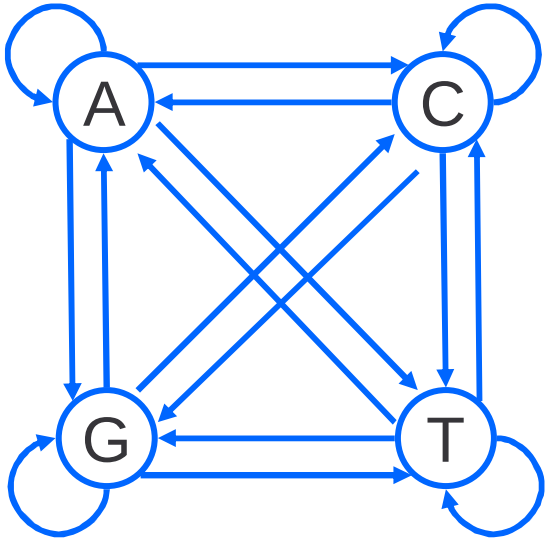
observed frequencies

island

+	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

-	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	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

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

non island

score

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

X = ACGT

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

application: CpG islands

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

'bits' ($\log_2$)

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

log-score

$$\log \frac{P(X|\text{ island})}{P(X|\text{ non})} = \log \frac{\prod_{i=1}^{n} t^+_{x_{i-1}x_i}}{\prod_{i=1}^{n} t^-_{x_{i-1}x_i}} = \sum_{i=1}^{n} \log(\frac{t^+_{x_{i-1}x_i}}{t^-_{x_{i-1}x_i}})$$

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

LLR(ACGT) = 0.42+1.81–0.73 = 1.50 (0.37 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)

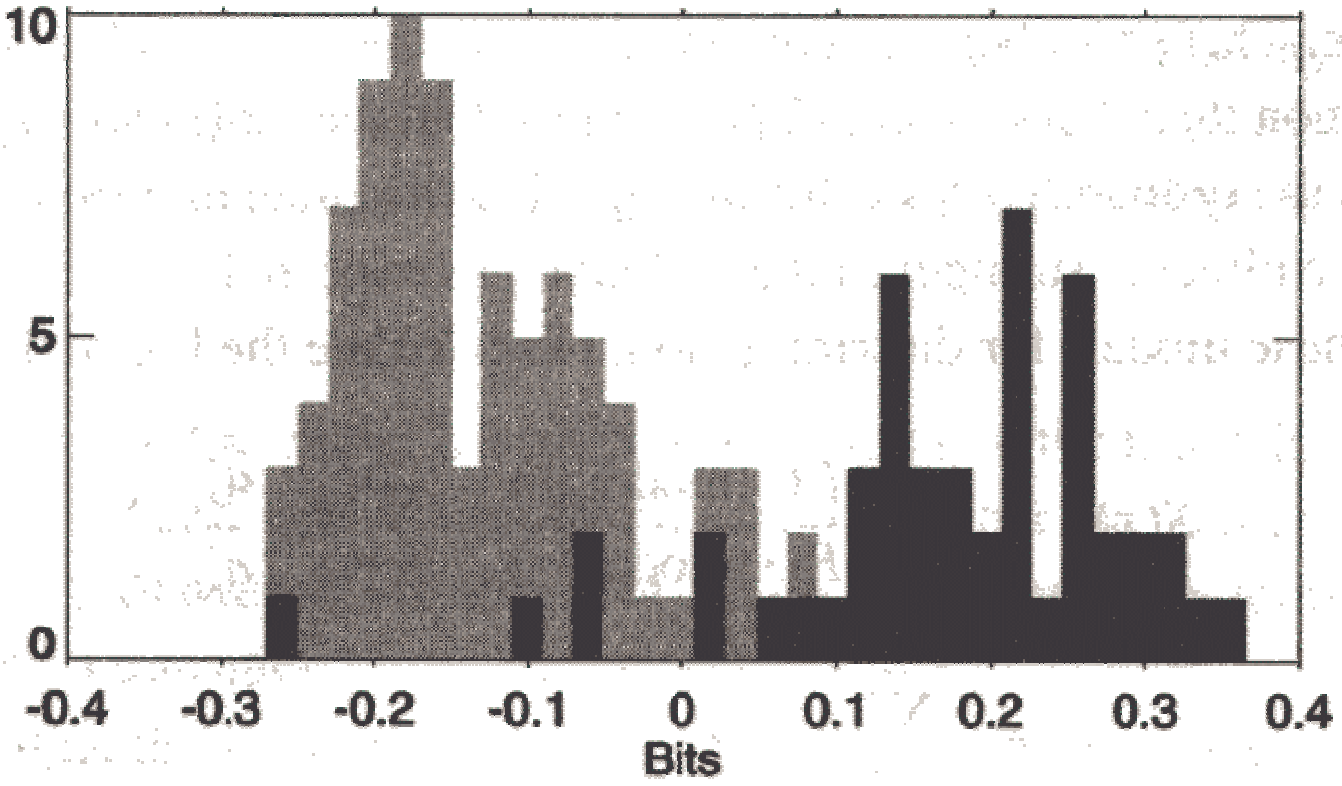
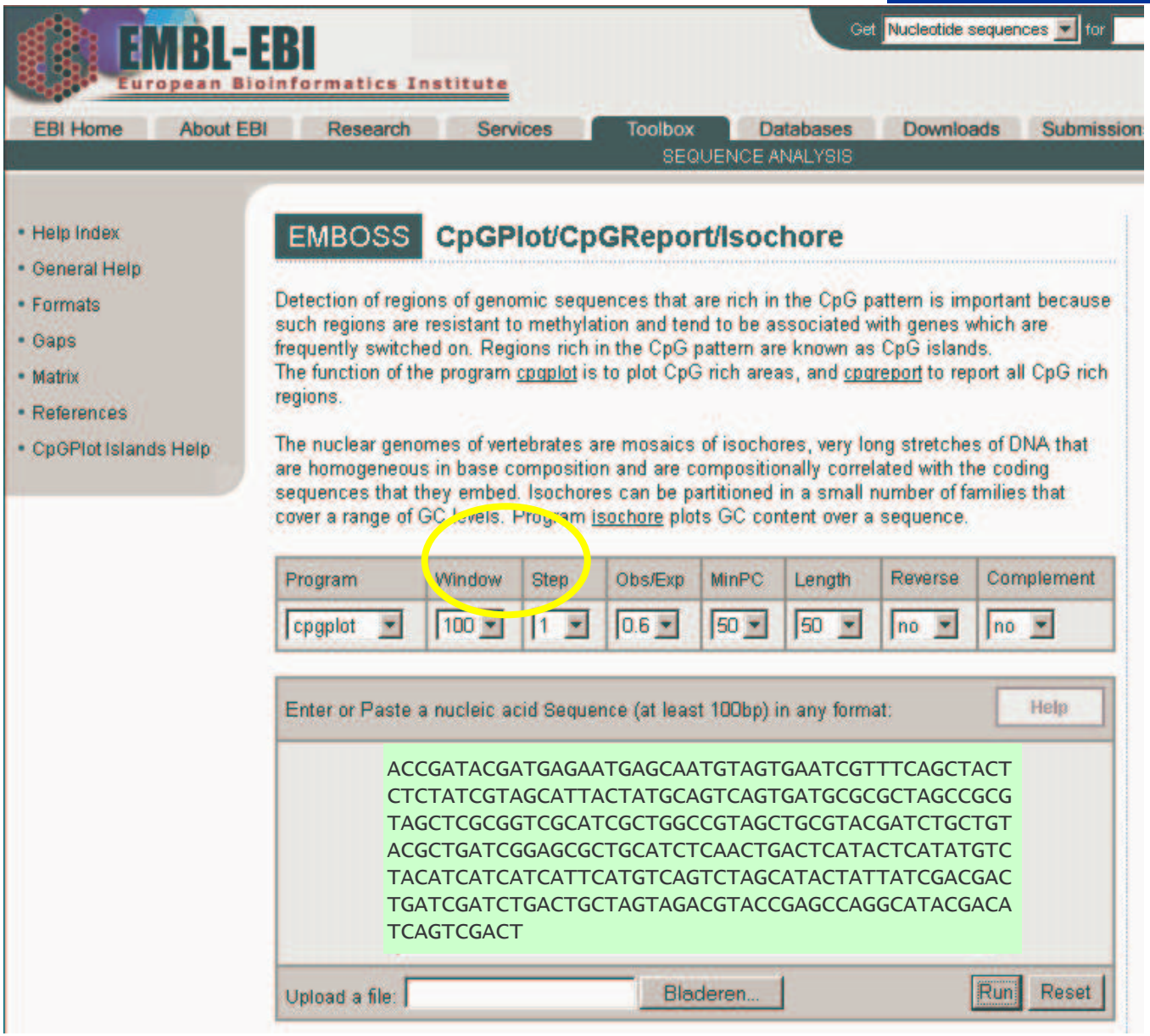


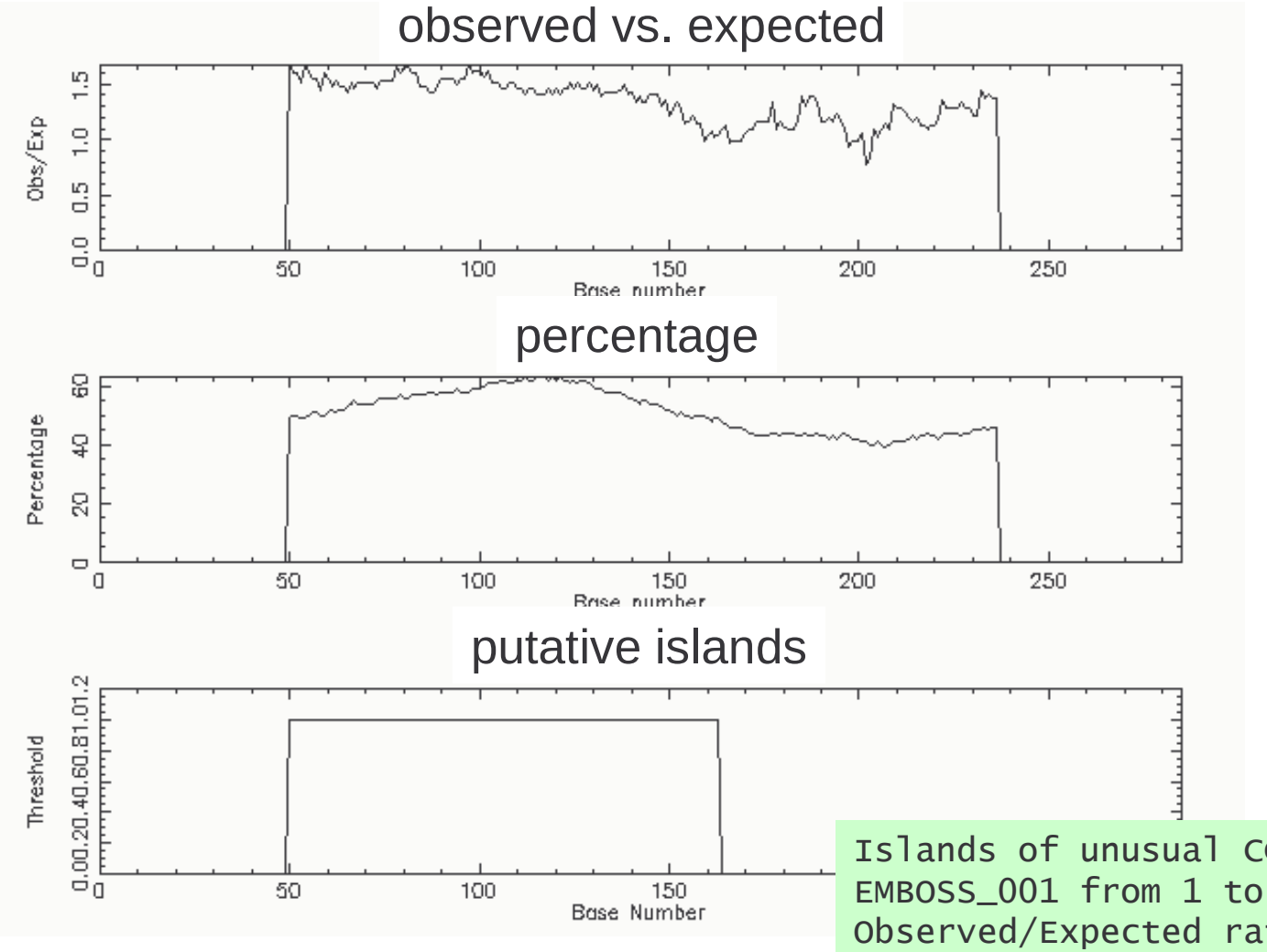
Figure 3.2 *The histogram of the length-normalised scores for all the sequences. CpG islands are shown with dark grey and non-CpG with light grey.*

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

CpGplot



CpGplot

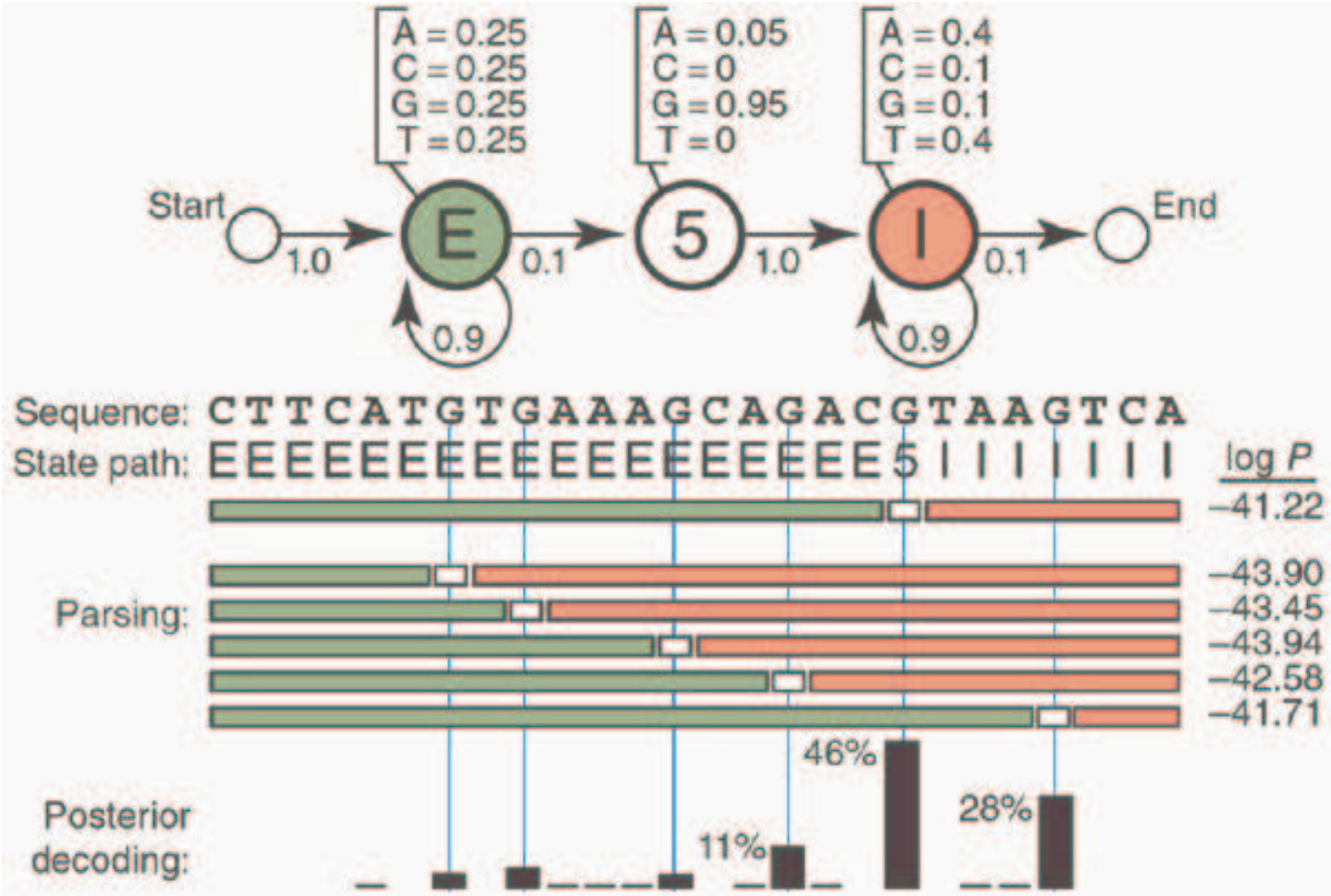



```
Islands of unusual CG composition
EMBOSS_001 from 1 to 286
Observed/Expected ratio > 0.60
Percent C + Percent G > 50.00
Length > 50
Length 114 (51..164)
```

hidden Markov model

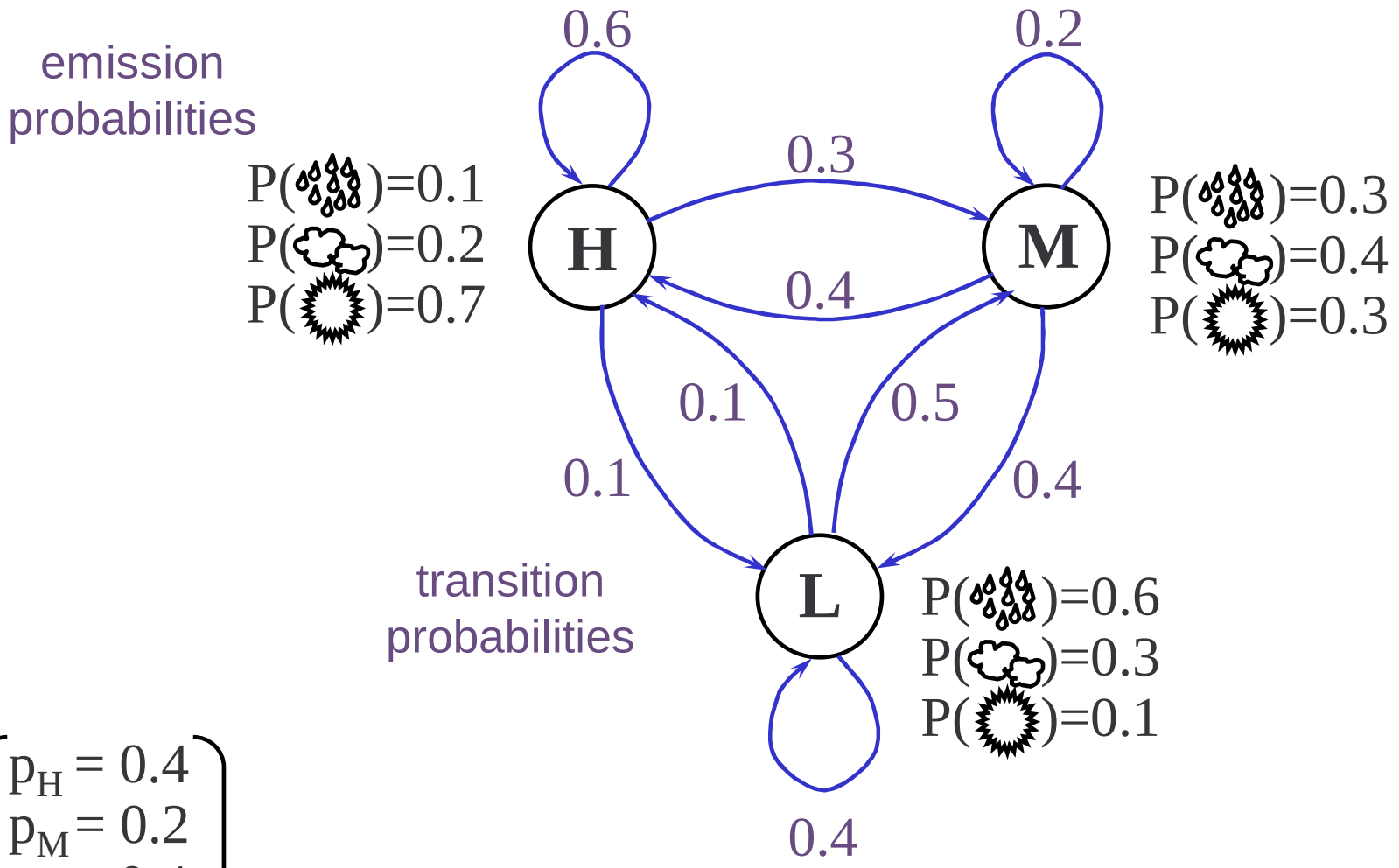
- where (in long sequence) are CpG islands ?
  second approach: *hidden Markov model*

Eddy (2004)



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

weather

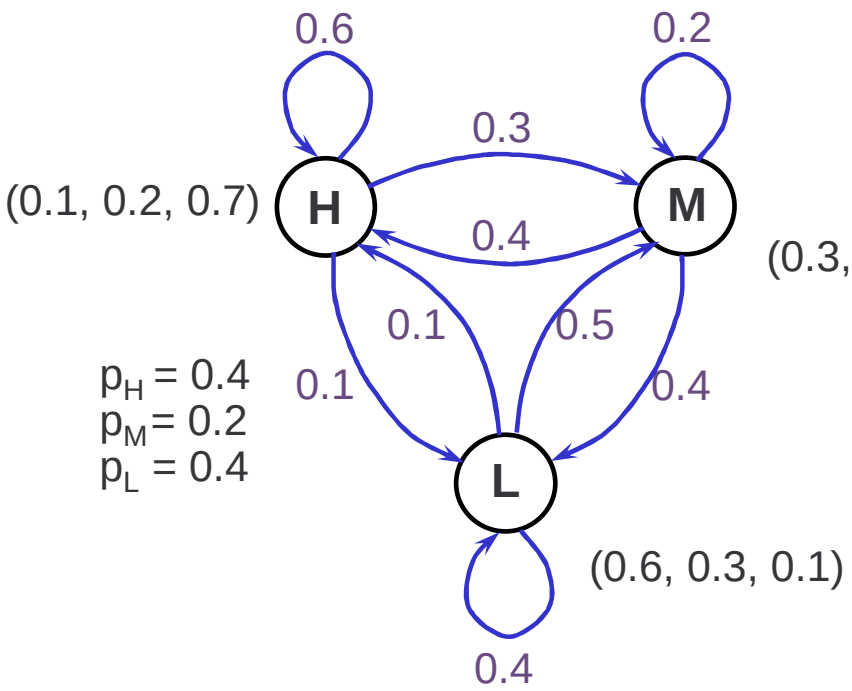


$$\begin{pmatrix} p_H = 0.4 \\ p_M = 0.2 \\ p_L = 0.4 \end{pmatrix}$$

initial probabilities

observed weather vs. pressure

weather



(, ,)

(R, C, S)

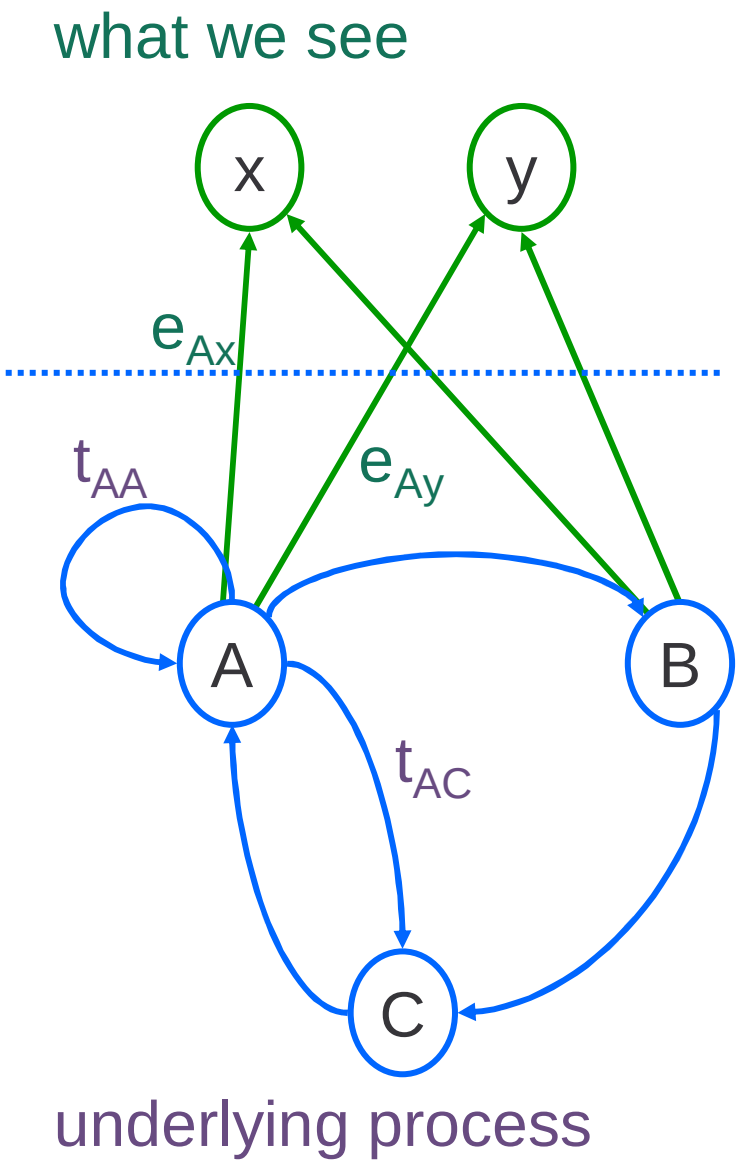
$P(\text{ RCCSS} \mid \text{HHHHH}) = 1 \cdot 2 \cdot 2 \cdot 7 \cdot 7 = 196 \quad (x10^{5})$

$P(\text{ RCCSS} \mid \text{MMMMM}) = 3 \cdot 4 \cdot 4 \cdot 3 \cdot 3 = 432 \quad (x10^{5})$

$P(\text{ RCCSS, HHHHH}) = 4 \cdot 1 \cdot 6 \cdot 2 \cdot 6 \cdot 2 \cdot 6 \cdot 7 \cdot 6 \cdot 7 = 1016 \quad (x10^{-7})$

$P(\text{ RCCSS, MMMMM}) = 2 \cdot 3 \cdot 2 \cdot 4 \cdot 2 \cdot 4 \cdot 2 \cdot 3 \cdot 2 \cdot 3 = 14 \quad (x10^{-7})$

hidden Markov model



model M = (Σ,Q,T)

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

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

observe states *indirectly* 'hidden'

- emission probabilities

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

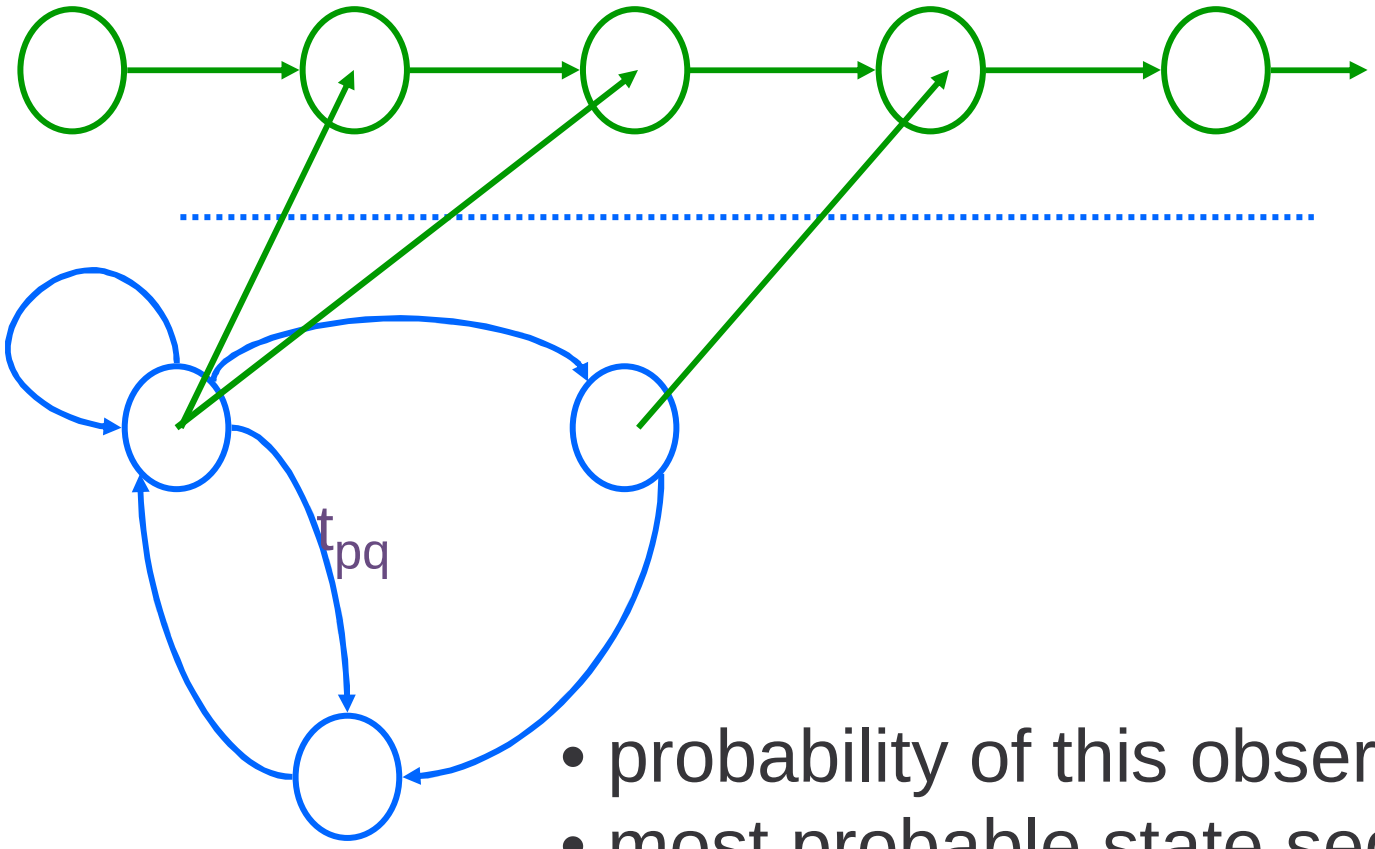
probability

observation given the model

? there may be *many* state seq's

HMM main questions

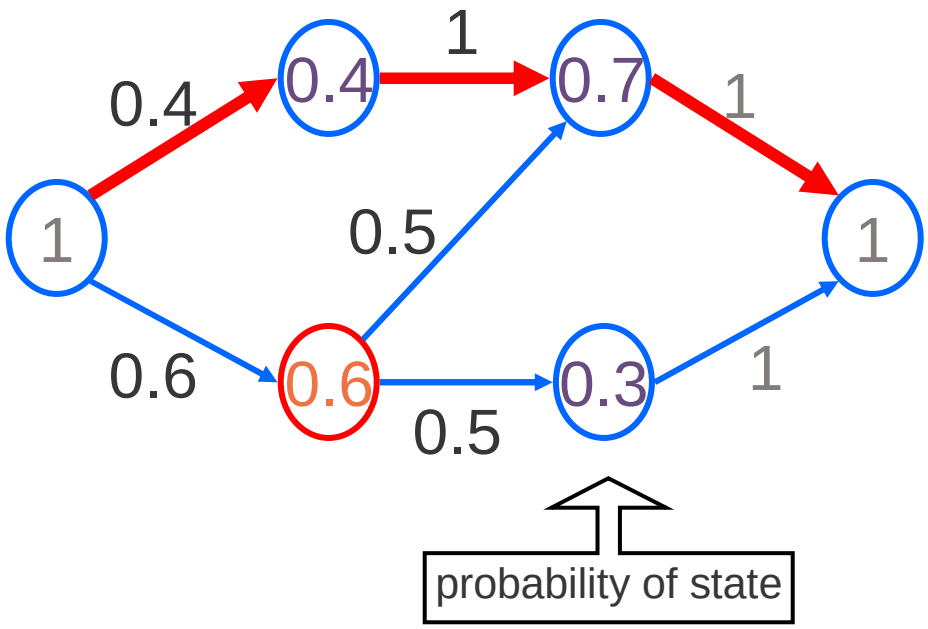
observation $X \in \Sigma^*$



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

probability … !

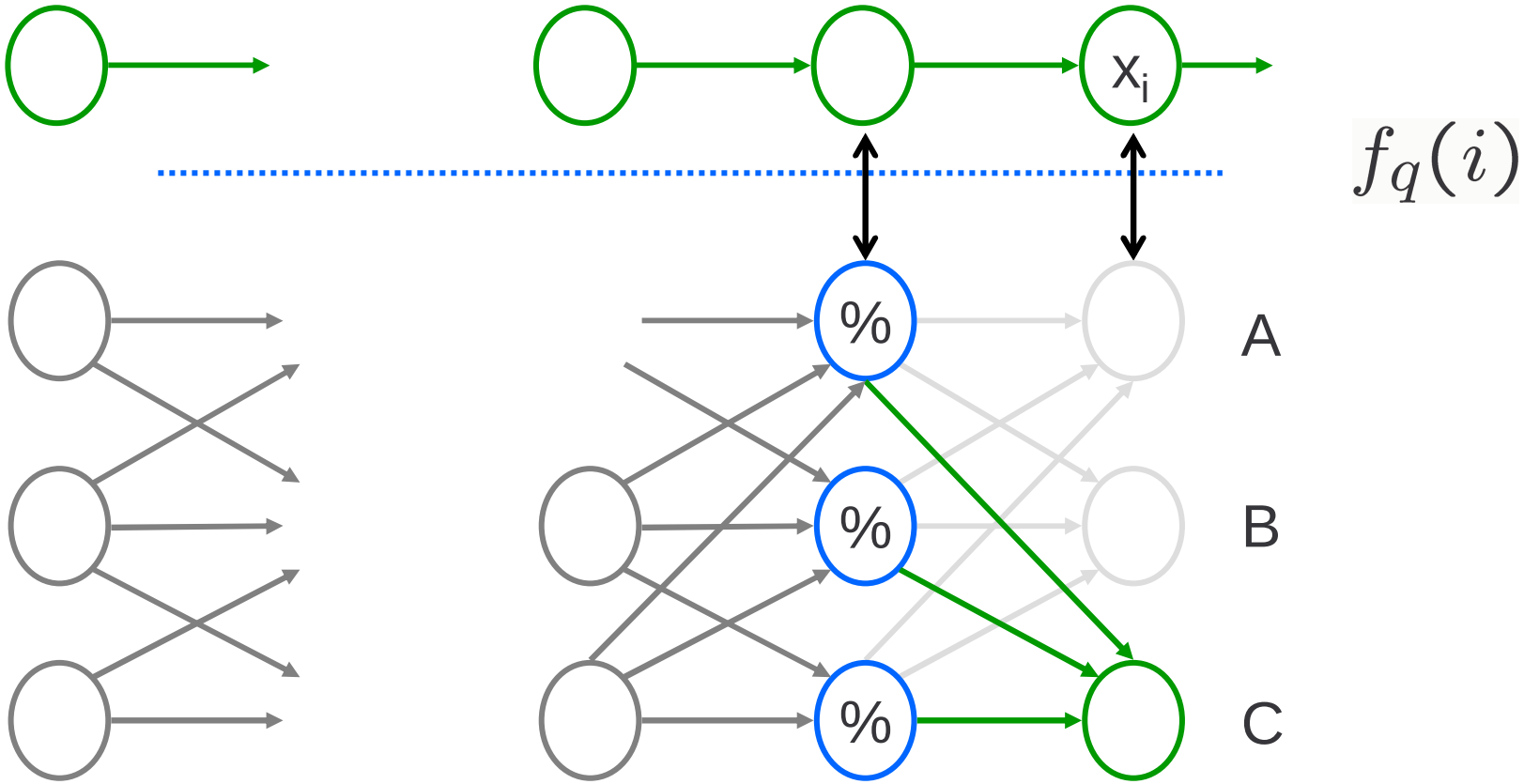
most probable state vs. optimal path



* 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

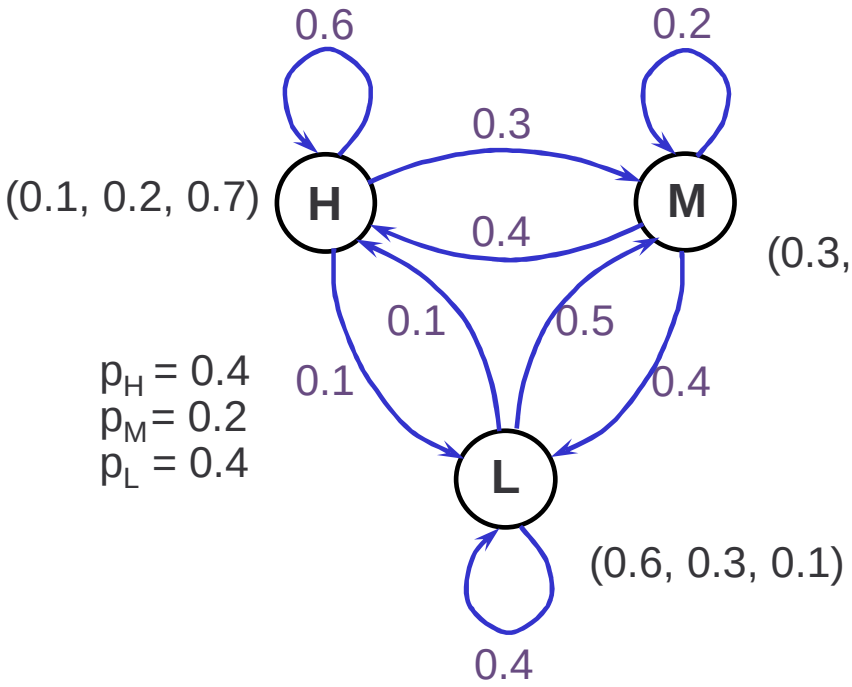
$$f_q(i) = P(x_1 \ldots x_i, \pi_i = q)$$

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

‘forward’ probability

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

weather



(, ,)

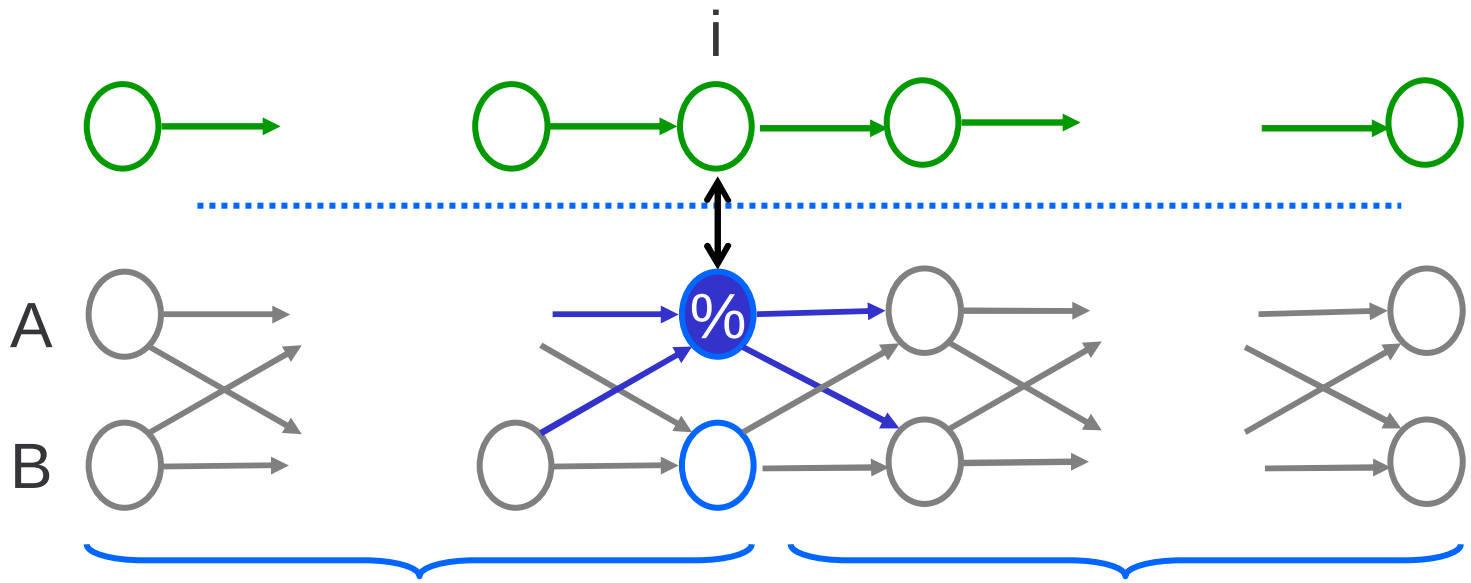
(R, C, S)

P(RCCSS) = P(RC…)

	1:R		2:C	
H	0	4·1 = 4	(4·6+6·4+24·1)·2 = 144	(x10^{-4})
M	0	2·3 = 6	(4·3+6·2+24·5)·4 = 576	(x10^{-4})
L	0	4·6 = 24	(4·1+6·4+24·4)·3 = 372	(x10^{-4})
0	1			

posterior decoding

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



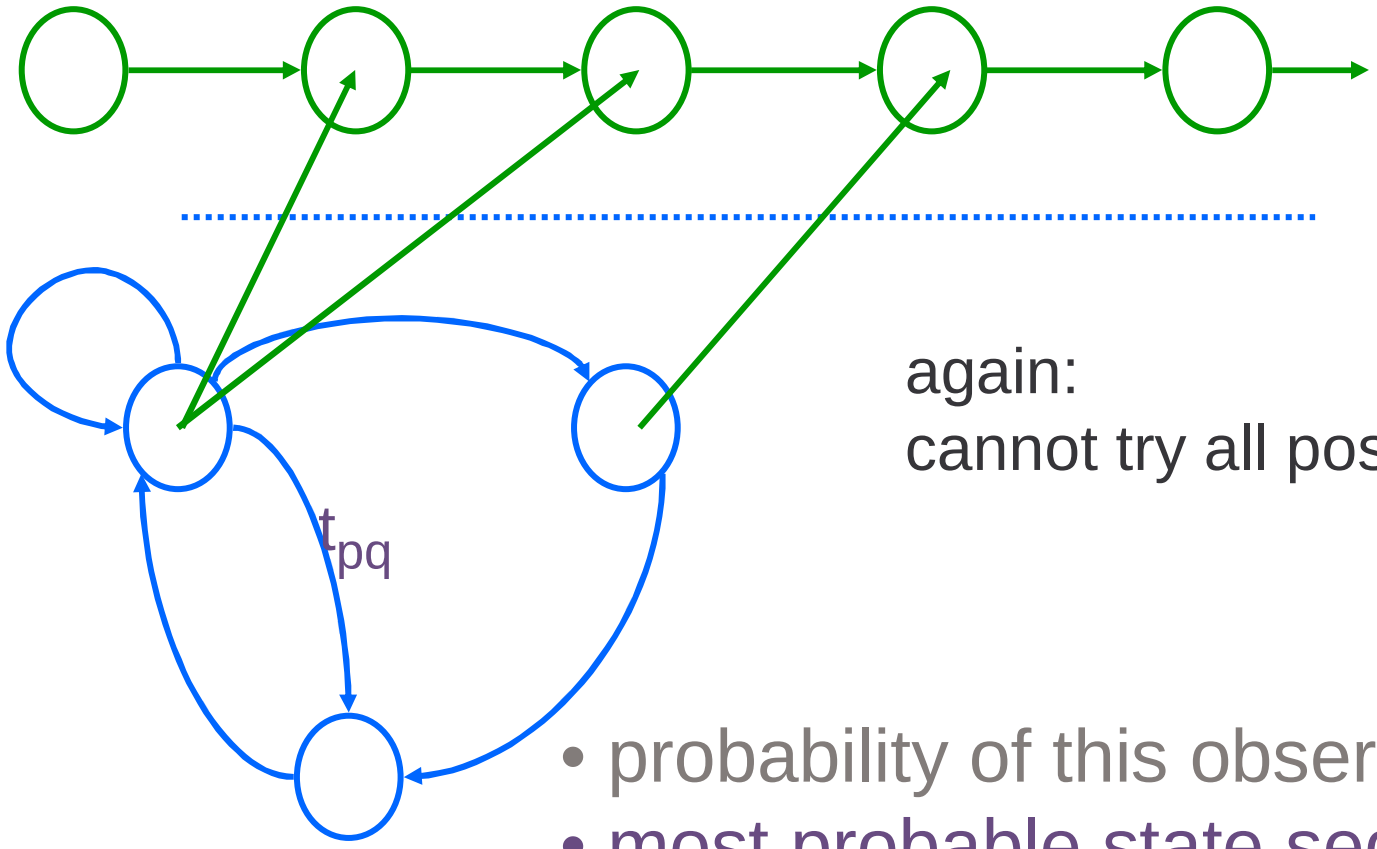
$f_q(i) = P(x_1 \dots x_i, \pi_i = q)$ $b_q(i) = P(x_{i+1} \dots x_n \mid \pi_i = q)$

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



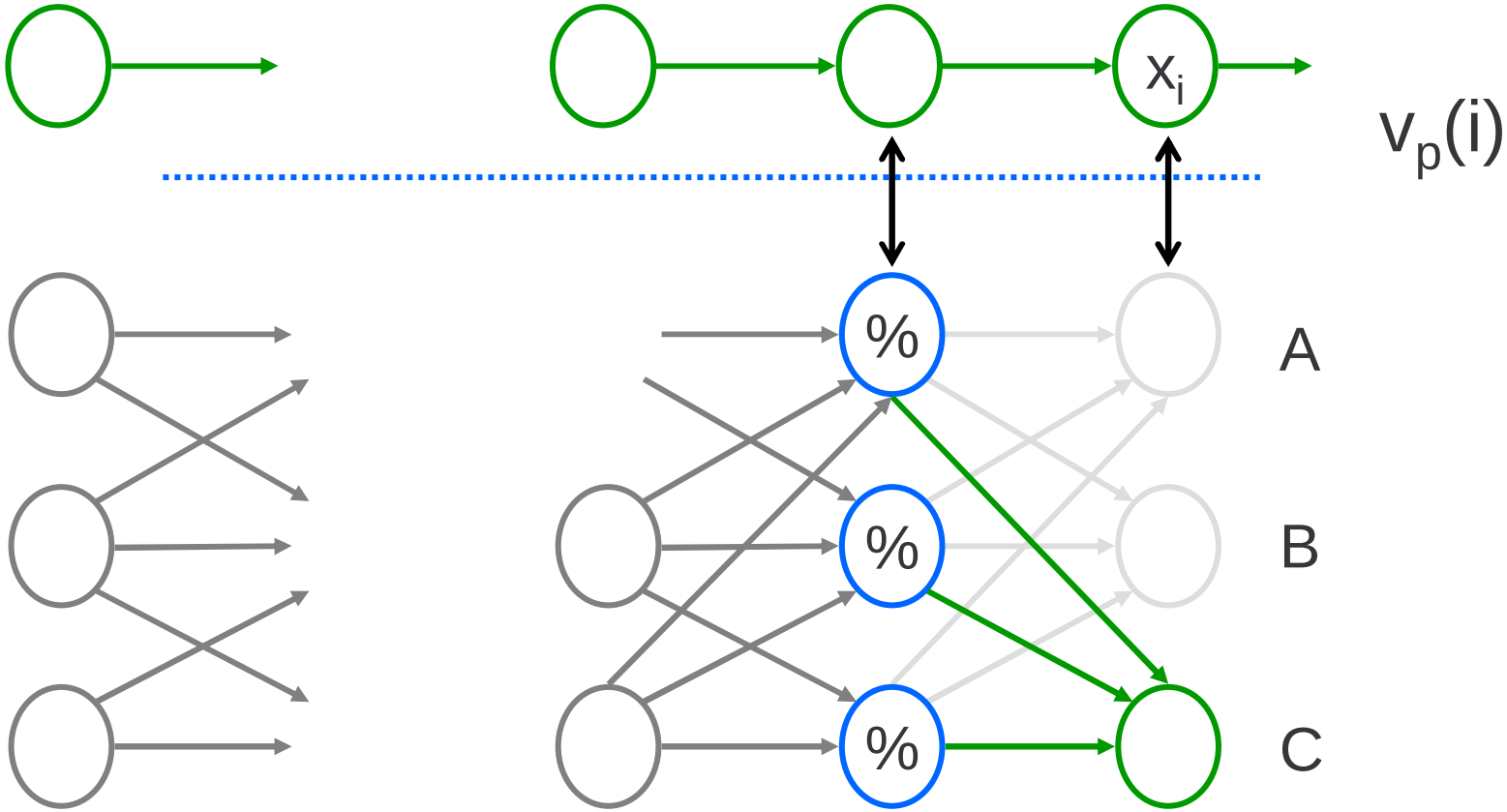
again:
cannot try all possibilities
Viterbi

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

Viterbi algorithm

most probable state sequence for observation

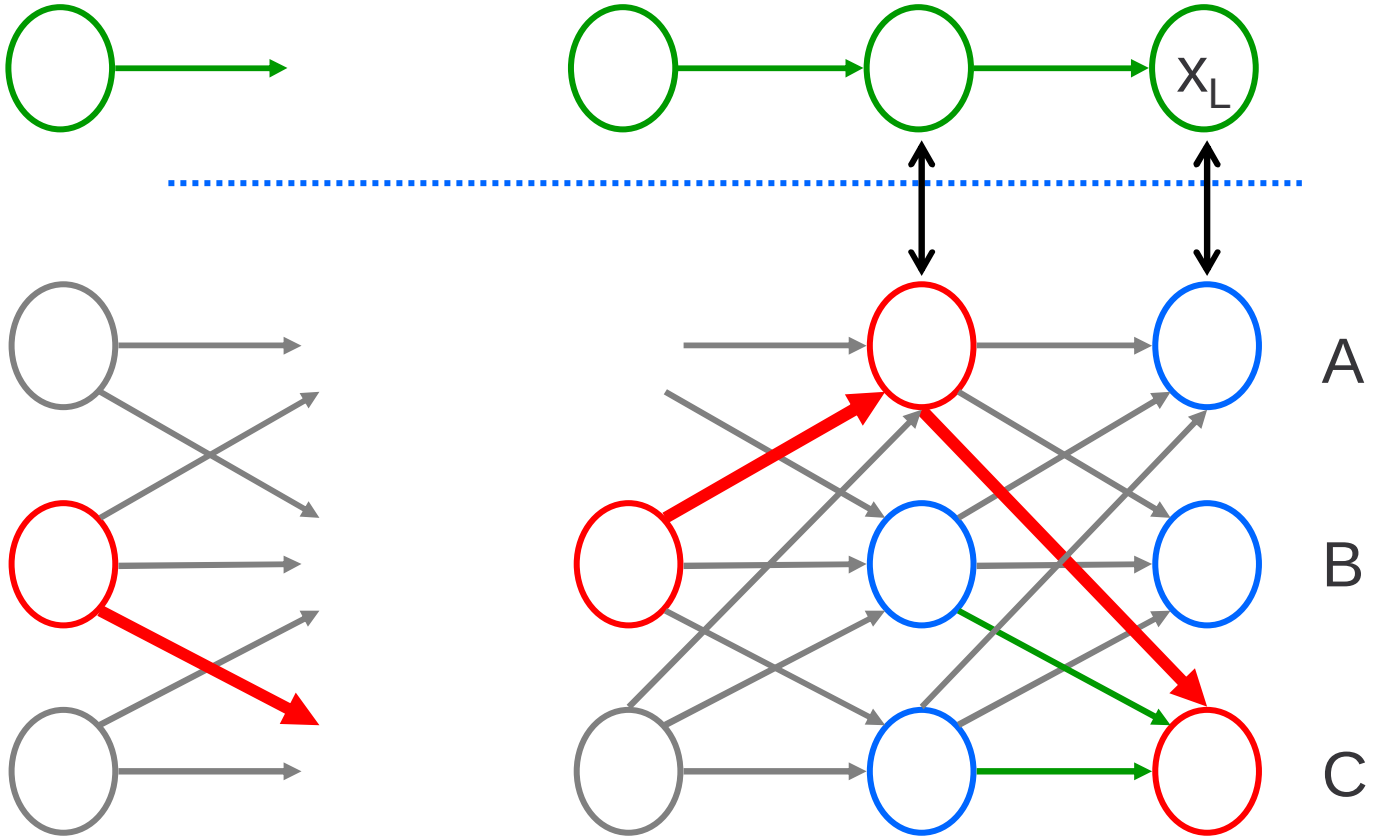
(1) *dynamic programming*: probability ending in state



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

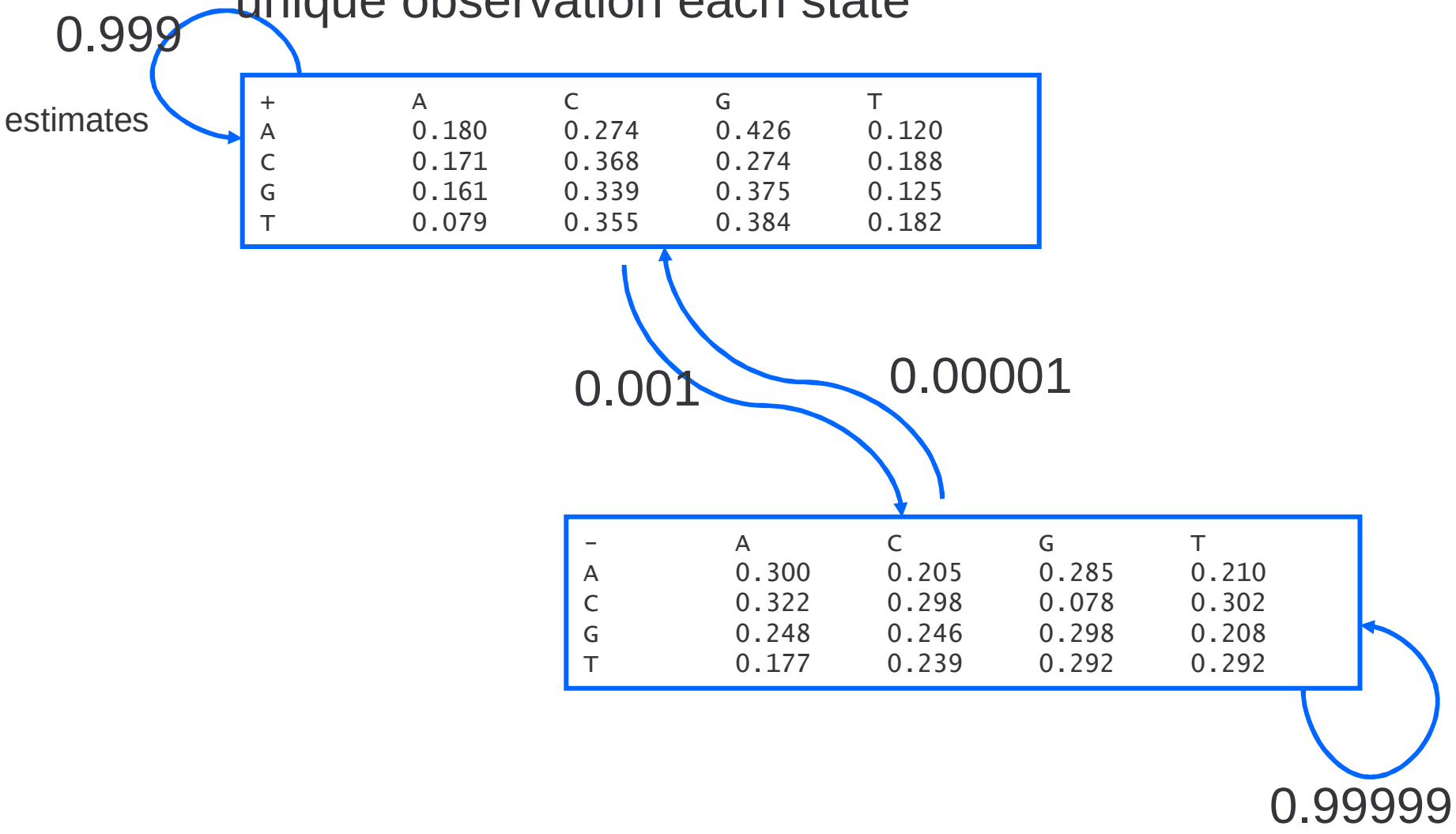
Viterbi algorithm

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

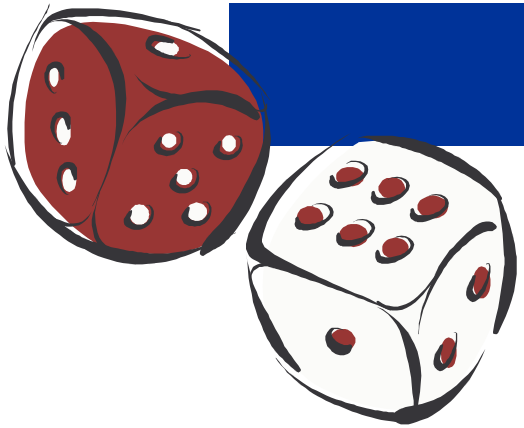


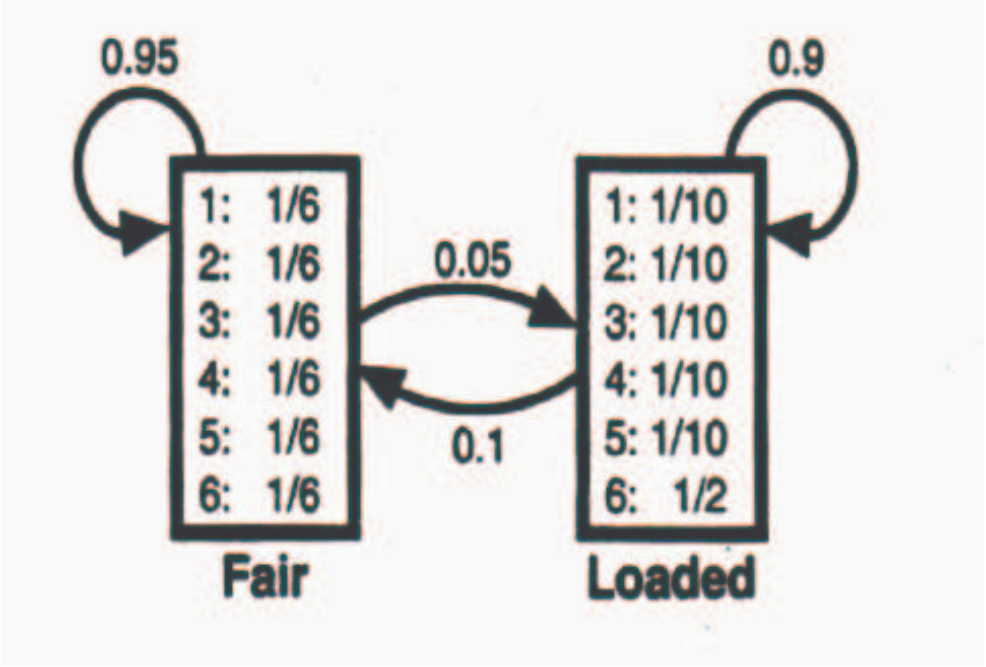
CpG islands ctd.

8 states A^+ vs A^-
unique observation each state



dishonest casino dealer





dishonest casino dealer

```
Rolls    3151162464466442453211316311641521336251445436316566266566666
Die      FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLL
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLL

Rolls    651166453132651245636664631636663162326455235266666625151631
Die      LLLLLLFFFFFFFFFFFLLLLLLLLLLLLLLLLFFFLLLLLLLLLLLLLLFFFFFFFFF
Viterbi  LLLLLLFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF

Rolls    22255544166656656356432436413151346514635341112641462625335 6
Die      FFFFFFFLLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL

Rolls    366163666466232534413661661163252562462255265252266435353336
Die      LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi  LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Rolls    233121625364414432335163243633665562466662632666612355245242
Die      FFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFF
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLFFFFFFFFFF
```

dishonest casino dealer

```
   Observation
3661636664662325344136616611632525624622552652522664353533336
   Viterbi
LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
   Forward
FFLLLLLLLLLLLLFFFFFFFFFLFLLLFLLFFFFFFFFFFFFFFFFFFFFLFFFFFFFFF
   Posterior (total)
LLLLLLLLLLLLFFFFFFFFFLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Parameter estimation

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

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 Θ
- unstable solutions !
- local maximum

tips:

- repeat for several initial Θ
- start with meaningful Θ

Viterbi training (an alternative)
determine optimal paths
recompute as if paths known

- score may decrease!