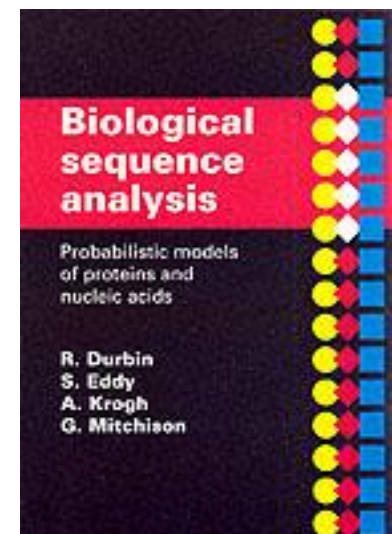


Hidden Markov Models

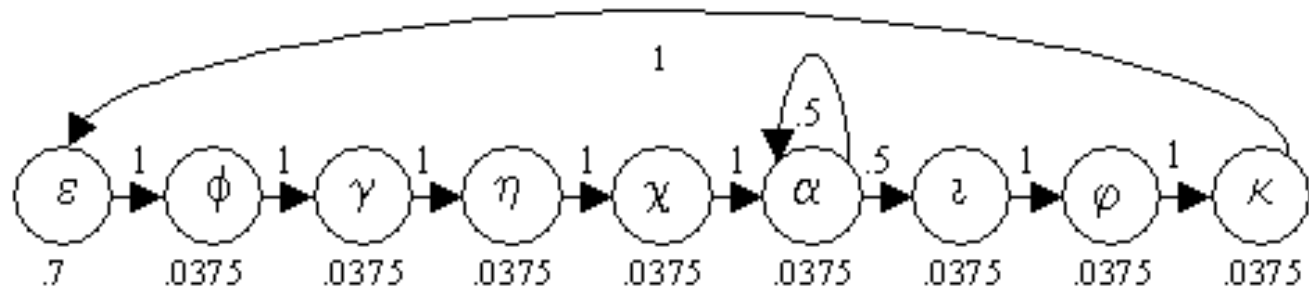
based on chapters from the book
Durbin, Eddy, Krogh and Mitchison
Biological Sequence Analysis
via Shamir's lecture notes



music recognition



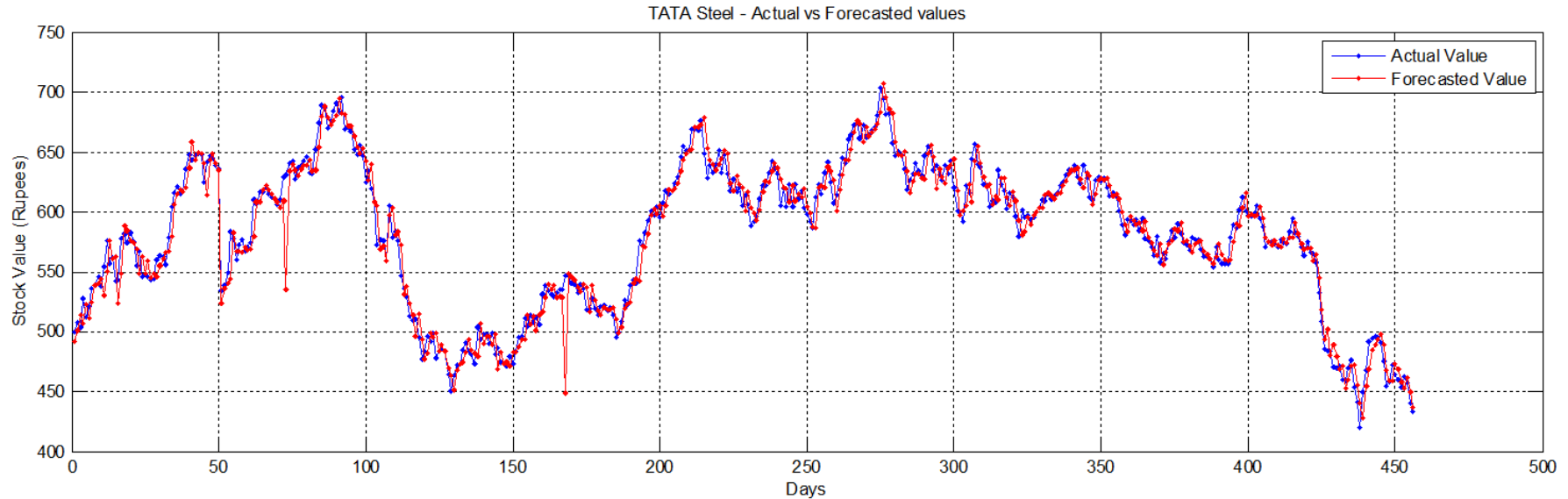
Delta Pitch:	2	2	0	-2	-2	2	2	-4	-5	5	2	2	0
IOI:	3	1	2	2	1	1	1	1	2	2	3	1	2
IOI ratio:	3	.5	1	2	1	1	1	.5	1	.66	3	.5	1
State:	ε	ϕ	γ	η	χ	α	α	ι	φ	κ	ε	ϕ	γ



deal with variations in

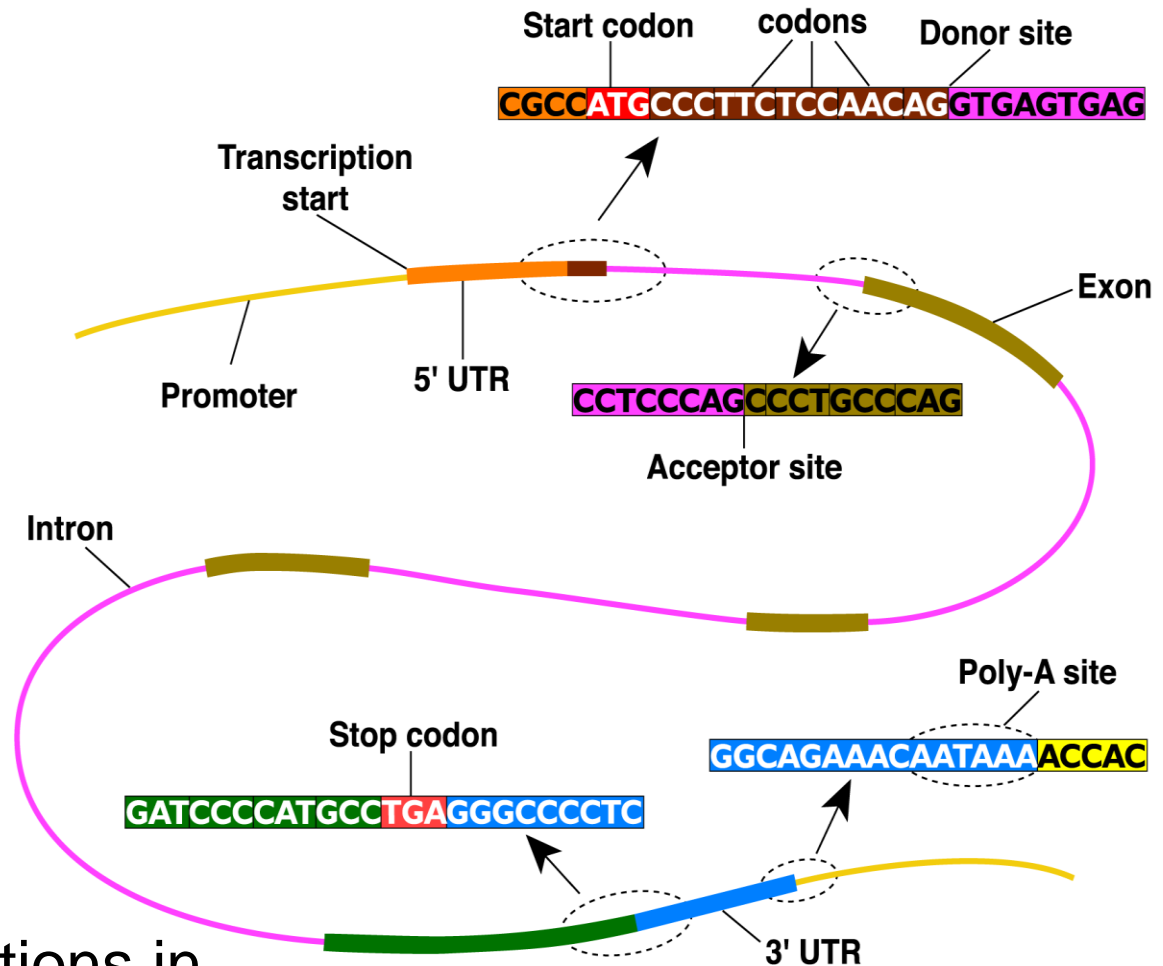
- pitch
- timing
- timbre
- ...

Stock Market Prediction



- Actual Value versus Forecasted Value for Tata Steel in Rupees over the period 5-9 2009 – 23-9 2011.
- Variations of value over time.
- From: A. Gupta, B. Dhingra, Stock Market Prediction Using Hidden Markov Models, 2011.

application: gene finding



deal with variations in

- actual sound
- timing

- actual base (match/substitutions)
- insertions/deletions

Basic Questions

Given:

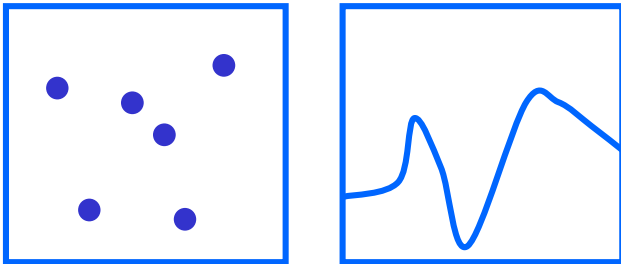
- A sequence of “observations”
- A probabilistic model of our “domain”

Questions:

- Does the given sequence belong to a certain family?
 - Markov chains
 - Hidden Markov Models (HMMs)
- Can we say something about the internal structure of the sequence? (indirect observations)
 - Hidden Markov Models (HMMs)

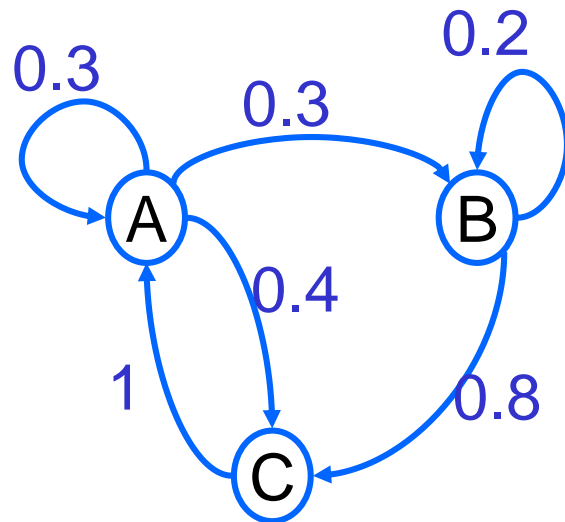
Introduction Markov Chain Model

Discrete vs Continuous



Characteristics

- Discrete time
- Discrete space
- No state History
 - Present state only
- States and transitions



Notations:

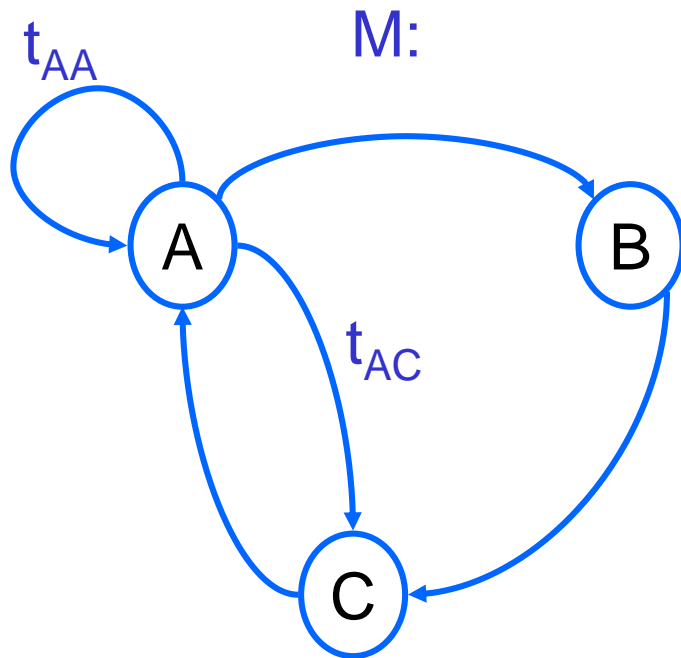
$P(X)$	probability for event X
$P(X, Y)$	event X and event Y
$P(X Y)$	event X given event Y

Definition of Markov Chain Model

- A **Markov chain**^[1] model is defined by
 - a set of states
 - some states emit symbols
 - other states (e.g., the begin state) are silent
 - a set of transitions with associated probabilities
 - the transitions emanating from a given state define a distribution over the possible next states (i.e., all positive, and sum equals 1)

[1] Марков А. А., Распространение закона больших чисел на величины, зависящие друг от друга. — Известия физико-математического общества при Казанском университете. — 2-я серия. — Том 15. (1906) — С. 135—156

Markov Model



Markov Model $M = (Q, P, T)$, with

- Q the set of states
- P the set of initial probabilities p_x for each state x in Q
- $T = (t_{xy})$ the transition probabilities matrix/graph, with t_{xy} the probability of the transition from state x to state y .

This is a **first order Markov Model**:
no history is modeled

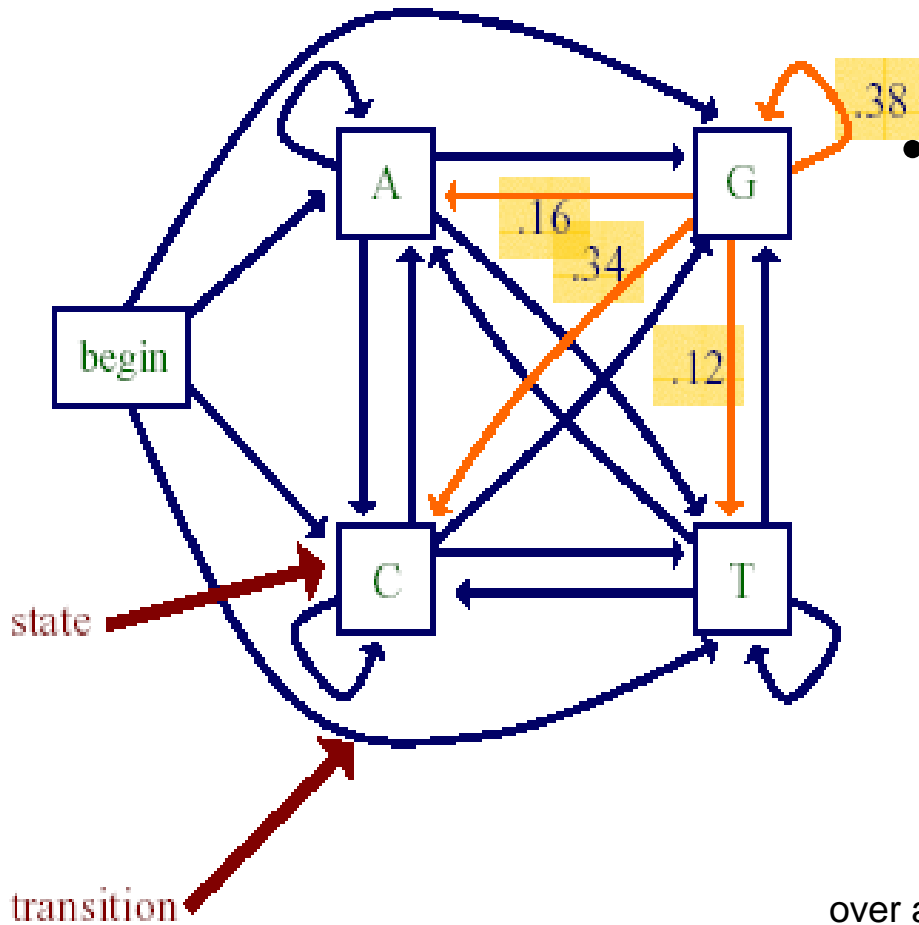
An observation X is a sequence of states:

$$X = x_1 x_2 \dots x_n$$

The probability of an observation X given the model M is equal to:

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

A Markov Chain Model Example



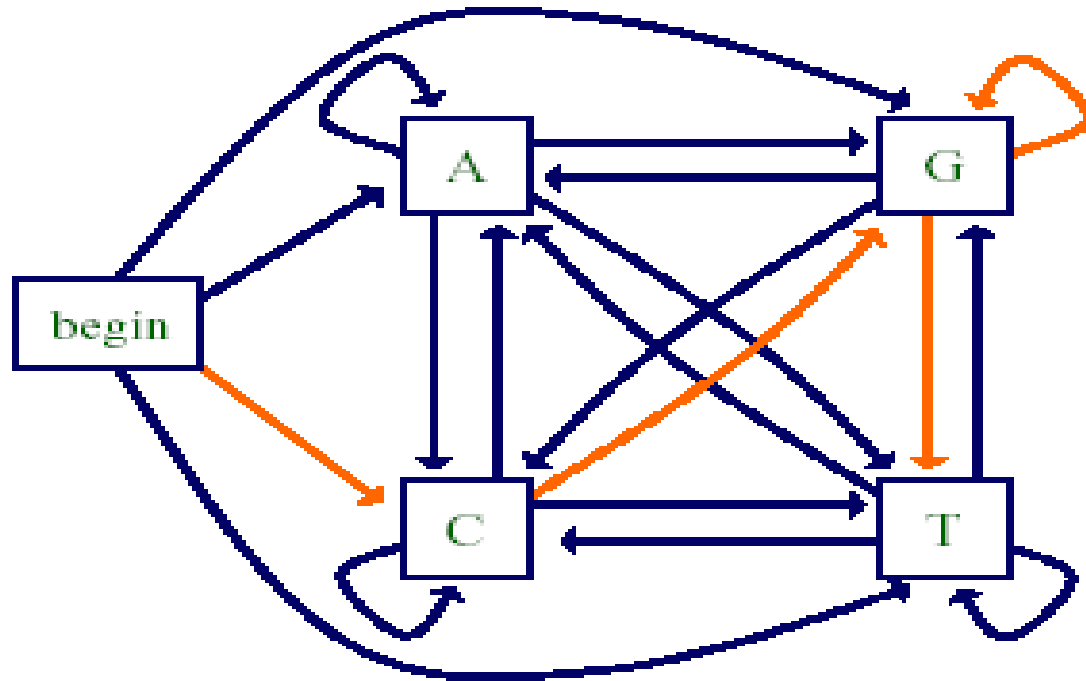
• Transition probabilities

- $\Pr(x_i=a|x_{i-1}=g)=0.16$
- $\Pr(x_i=c|x_{i-1}=g)=0.34$
- $\Pr(x_i=g|x_{i-1}=g)=0.38$
- $\Pr(x_i=t|x_{i-1}=g)=0.12$

$$\sum \Pr(x_i | x_{i-1} = g) = 1$$

over all neighbors x_i

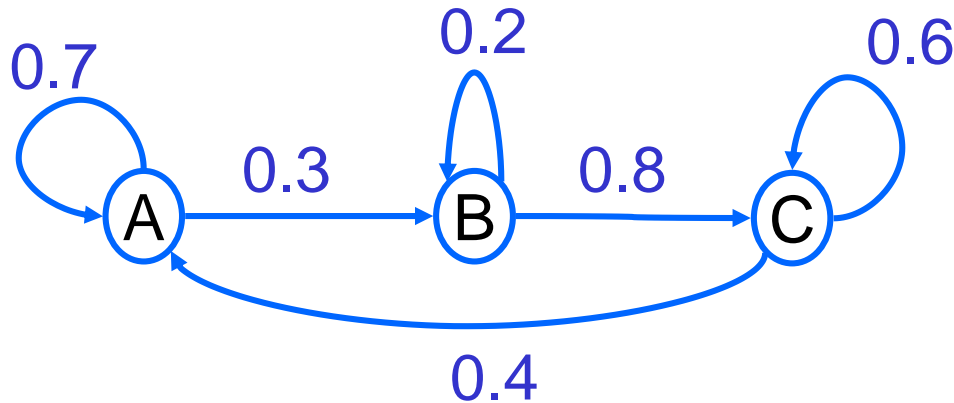
The Probability of a Sequence for a Markov Chain Model



$$\Pr(\text{CGGT}) = \Pr(\text{C})\Pr(\text{G}|\text{C})\Pr(\text{G}|\text{G})\Pr(\text{T}|\text{G})$$

Markov Chains: Another Example

M_1 :



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

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

unique starting state A

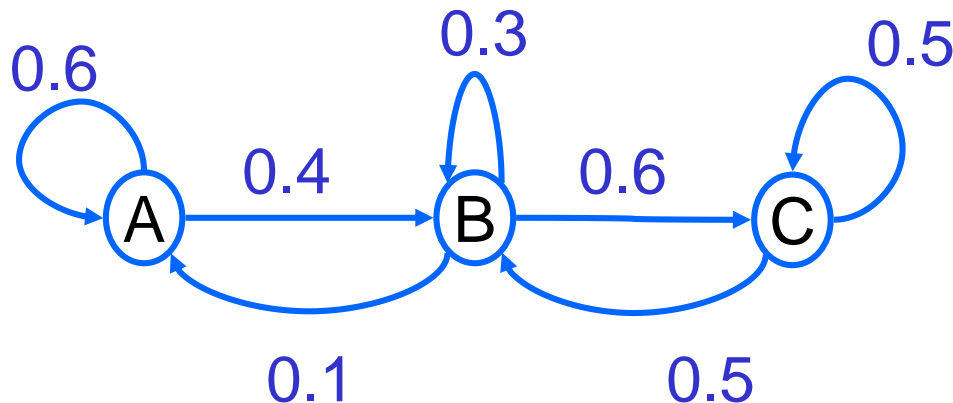
$$T = \begin{matrix} & \begin{matrix} A & B & C \end{matrix} \\ \begin{matrix} A \\ B \\ C \end{matrix} & \begin{pmatrix} .7 & .3 & 0 \\ 0 & .2 & .8 \\ .4 & 0 & .6 \end{pmatrix} \end{matrix}$$

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

$$P(AABBBCCC \mid M_1) = 1 \cdot .7 \cdot .3 \cdot .2 \cdot .8 \cdot .6 \cdot .6 \cdot 10^{-6} = 1.2 \cdot 10^{-2}$$

$$P(AABBBCCC \mid M_2) = 1 \cdot .6 \cdot .4 \cdot .3 \cdot .6 \cdot .5 \cdot .5 \cdot 10^{-6} = 1.1^{11} \cdot 10^{-2}$$

M_2 :



Markov Models: Properties

Given some sequence x of length L , we can ask:
How probable is the sequence x given our model M ?

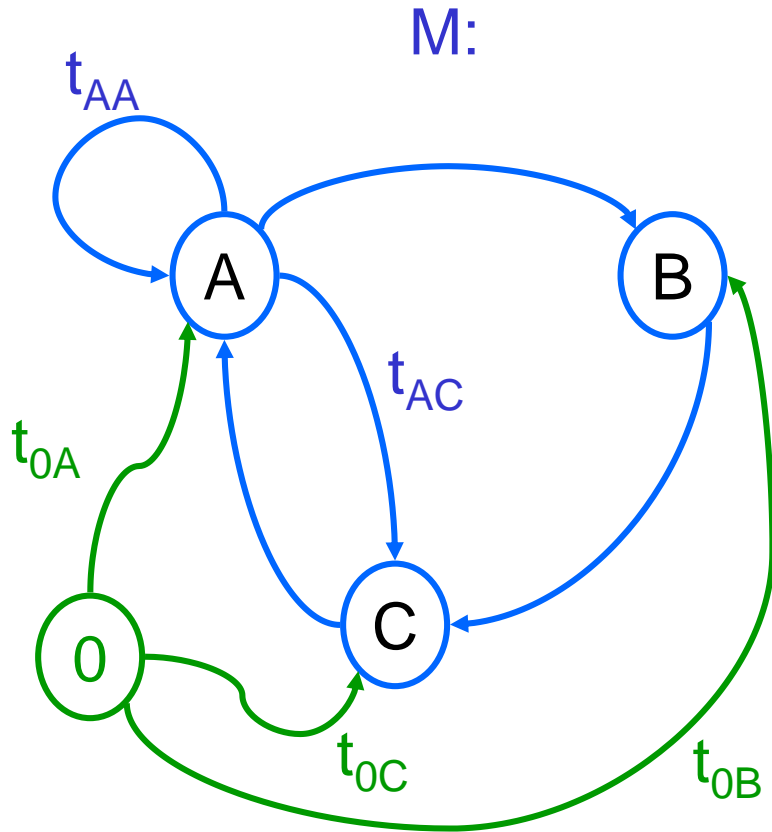
- For **any probabilistic model of sequences**, we can write this probability as

$$\begin{aligned}\Pr(x) &= \Pr(x_L x_{L-1} \dots x_1) \\ &= \Pr(x_L | x_{L-1} \dots x_1) \Pr(x_{L-1} | x_{L-2} \dots x_1) \dots \Pr(x_1)\end{aligned}$$

- key property of a **(1st order)** Markov chain: the probability of each x_i depends only on the value of

$$\begin{aligned}x_{i-1} \quad \Pr(x) &= \Pr(x_L | x_{L-1}) \Pr(x_{L-1} | x_{L-2}) \dots \Pr(x_2 | x_1) \Pr(x_1) \\ &= \Pr(x_1) \prod_{i=2}^L \Pr(x_i | x_{i-1})\end{aligned}$$

Markov Model: Underflow Problem



$$t_{0x} = p_x$$

- initial state x_0 fixed
~ initial probabilities
- final state [not depicted]

$$X = x_1 x_2 \dots x_n$$

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

small values: **underflow**

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

Markov Model: Comparing Models

Given:

$$X = x_1 x_2 \dots x_n$$

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

Question: X best explained by which model?

We can calculate: $P(X | M_1)$ vs. $P(X | M_2)$

We want to know: $P(M_1 | X)$ vs. $P(M_2 | X)$!!

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

$$\frac{P(M_1|X)}{P(M_2|X)} = \frac{P(X|M_1) \cdot P(M_1)}{P(X|M_2) \cdot P(M_2)}$$

bases are not random

Motivation for Markov Models in Computational Biology

- There are many cases in which we would like to represent the **statistical regularities** of some *class of sequences*
 - genes
 - various **regulatory sites** in DNA (e.g., where RNA polymerase and transcription factors bind)
 - **proteins** in a given family
- **Markov models** are well suited to this type of task

Markov Chain: An Example Application

- CpG islands
 - CG di-nucleotides are *rarer* in eukaryotic genomes than expected given the marginal probabilities of C and G
 - but the regions upstream of genes (*reading is from 5' to 3'*) are *richer* in CG di-nucleotides than elsewhere – so called CpG islands
 - useful evidence for finding genes
- Application: Predict CpG islands with Markov chains
 - a Markov chain to represent CpG islands
 - a Markov chain to represent the rest of the genome

Markov Chains for Discrimination

- Suppose we want to distinguish CpG islands from other sequence regions
- Given sequences from CpG islands, and sequences from other regions, we can construct
 - a model to represent CpG islands
 - a null model to represent the other regions
- We can then score a test sequence X by:

$$score(X) = \log \frac{\Pr(X | CpGModel)}{\Pr(X | nullModel)}$$

Markov Chains for Discrimination

Why can we use the scoring function:

$$\text{score}(X) = \log \frac{\Pr(X | CpGModel)}{\Pr(X | nullModel)}$$

- According to Bayes' rule we have:

$$\Pr(CpG | X) = \frac{\Pr(X | CpG) \Pr(CpG)}{\Pr(X)}$$

$$\Pr(null | X) = \frac{\Pr(X | null) \Pr(null)}{\Pr(X)}$$

- If we are not taking into account prior probabilities ($\Pr(CpG)$ and $\Pr(null)$) of the two classes, then from Bayes' rule it is clear that we just need to compare $\Pr(X|CpG)$ and $\Pr(X|null)$ as is done in our scoring function `score()`.

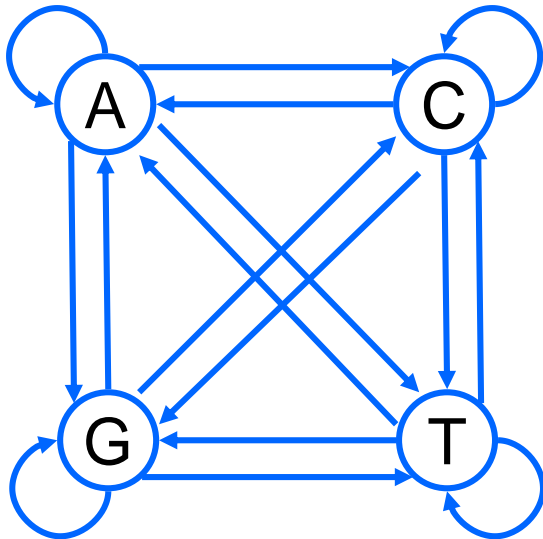
Markov Chain Application: CpG islands

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

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



In general consecutive CG pairs
 CG → CG are rare, although 'islands'
 Occur in signal (e.g.) promotor regions.

- observation: DNA sequence
 - model 1: CpG islands
 - model 2: 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?
Markov Chains: windowing
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

A->C

C->G

G->T

Note: A score > 1 is an
Indication of a CpG island.

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

application: CpG islands

LLR = Log-Likelihood Ratio

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

'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

$$\log_2(0.274/0.078) = 1.81$$

log-score (\log_2)

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

X = ACGT

$$\log_2 \frac{0.274 \cdot 0.274 \cdot 0.125}{0.205 \cdot 0.078 \cdot 0.208} = 0.42 + 1.81 - 0.73 = 1.50$$

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

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

$$1.5/4 = 0,375$$

- 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*
- ! What would be the length of window?

empirical data

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

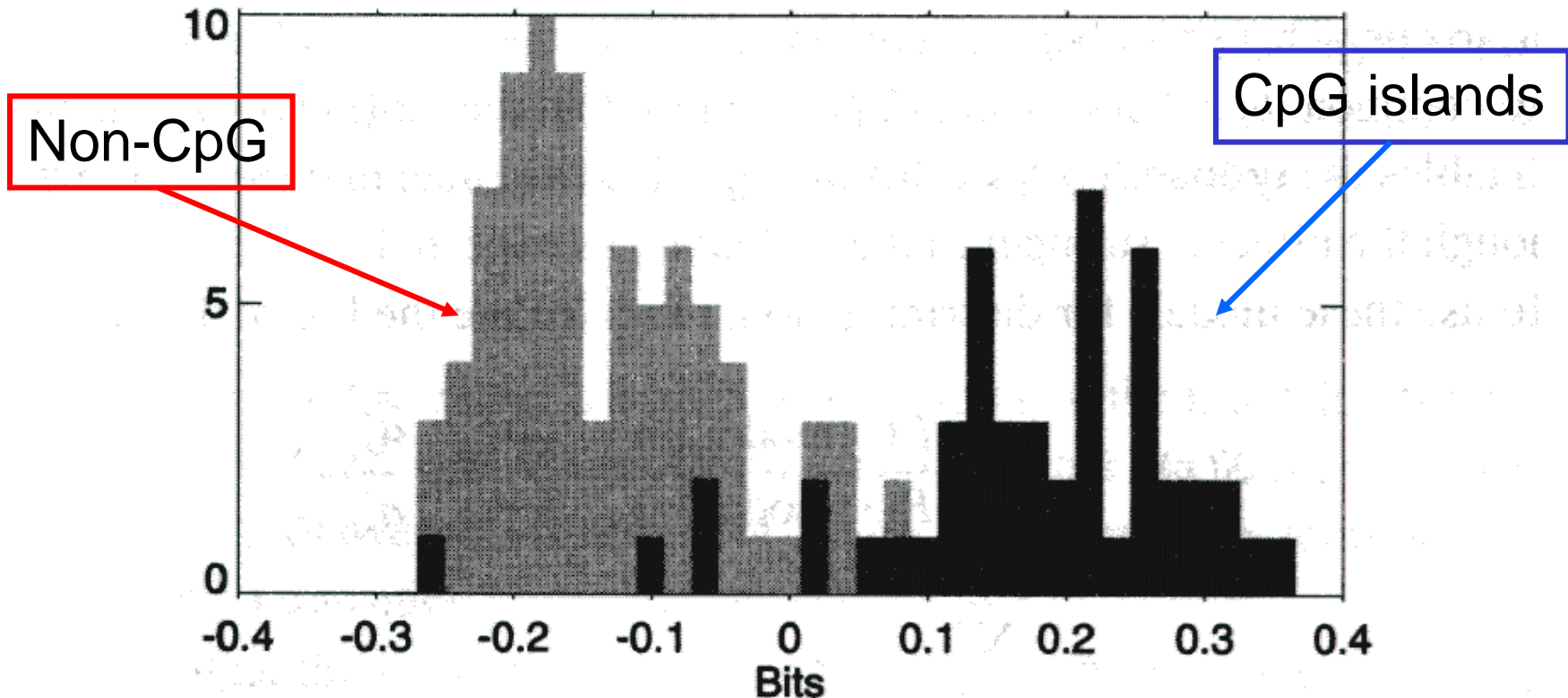


Figure 3.2 *The histogram of the length-normalised scores for all the sequences. CpG islands are shown with dark grey and non-CpG with light grey.*

- 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 [isochore](#) plots GC content over a sequence.

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

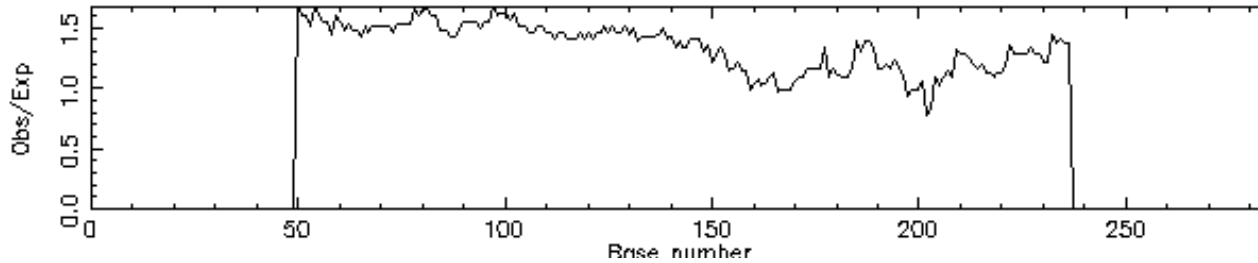
Enter or Paste a nucleic acid Sequence (at least 100bp) in any format: [Help](#)

```

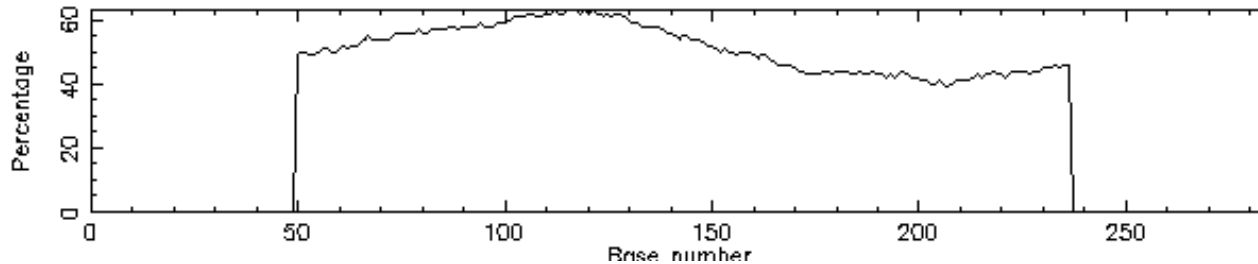
ACCGATACGATGAGAATGAGCAATGTAGTGAATCGTTTTAGCTACT
CTCTATCGTAGCATTACTATGCAGTCAGTGATGCGCGCTAGCCGCG
TAGCTCGCGGTCGCATCGCTGGCCGTAGCTGCGTACGATCTGCTGT
ACGCTGATCGGAGCGCTGCATCTCAACTGACTCATACTCATATGTC
TACATCATCATTCATGTGAGTCTAGCATACTATTATCGACGAC
TGATCGATCTGACTGCTAGTAGACGTACCGAGCCAGGCATACGACA
TCAGTCGACT
  
```

Upload a file:

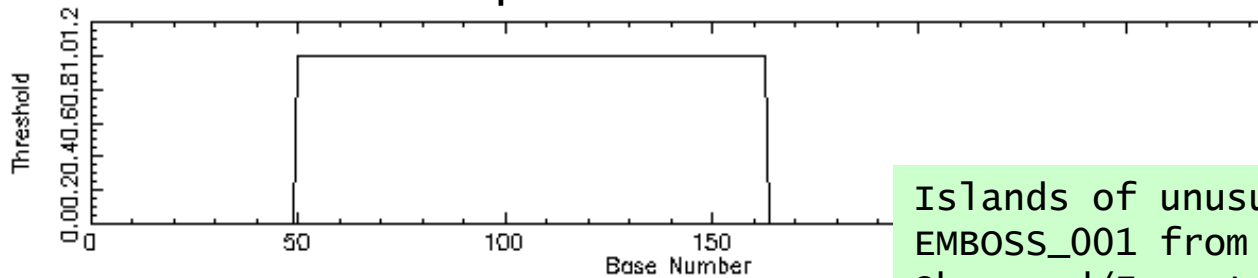
observed vs. expected



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)

Some Notes on: Higher Order Markov Chains

- The Markov property specifies that the probability of a state depends **only** on the probability of the previous state
- But we can build more “memory” into our states by using a **higher order** Markov model
- In an **n-th** order Markov model

$$\Pr(x_i \mid x_{i-1}, x_{i-2}, \dots, x_1) = \Pr(x_i \mid x_{i-1}, \dots, x_{i-n})$$

The probability of the current state depends on the previous **n** states.

Selecting the Order of a Markov Chain Model

- But the number of parameters we need to estimate for an **n-th** order Markov model grows exponentially with the order
 - for modeling DNA we need $O(4^{n+1})$ parameters (# of state transitions) for an **n-th** order model
- The higher the order, the less reliable we can expect our parameter estimates to be
 - estimating the parameters of a **2nd** order Markov chain from the complete genome of E. Coli (5.44×10^6 bases) , we would see each (length 3) word ~ 85.000 times on average (divide by 4^3)
 - estimating the parameters of a **9th** order chain, we would see each (length 10) word ~ 5 times on average (divide by $4^{10} \sim 10^6$)

Higher Order Markov Chains

- An n -th order Markov chain over some alphabet A is equivalent to a first order Markov chain over the alphabet of n -tuples: A^n
- Example: a 2nd order Markov model for DNA can be treated as a 1st order Markov model over alphabet

AA, AC, AG, AT

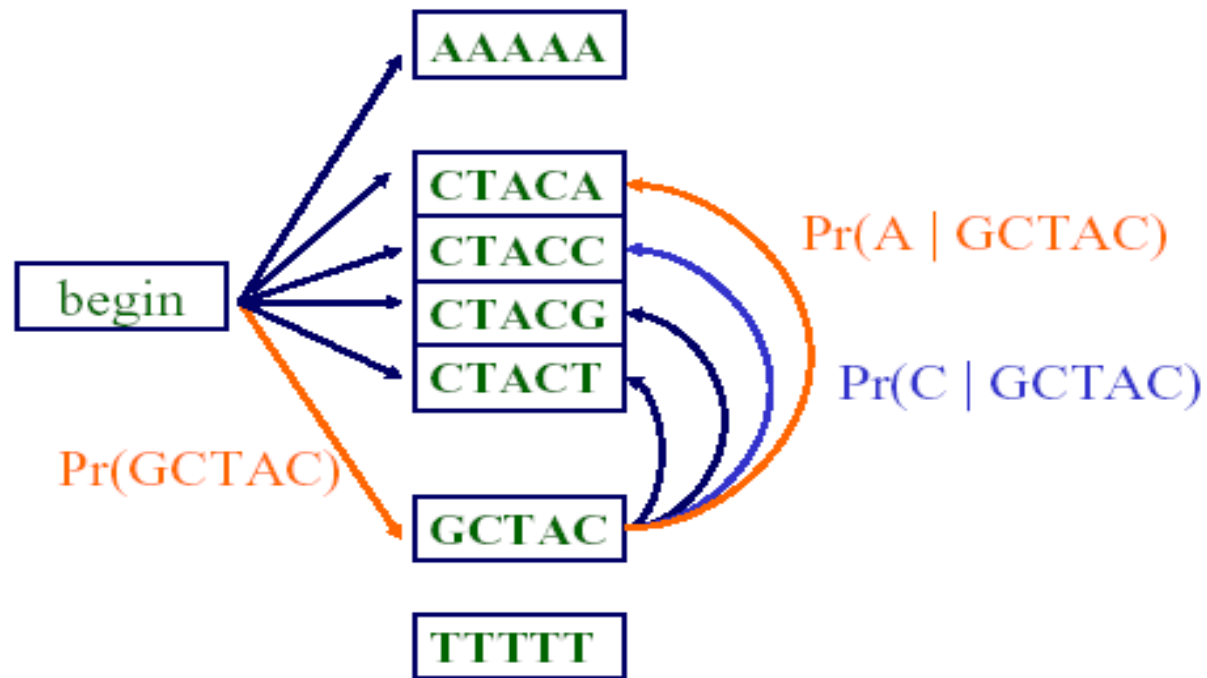
CA, CC, CG, CT

GA, GC, GG, GT

TA, TC, TG, TT

Transition probabilities:
 $P(A|AA)$, $P(A|AC)$, etc.

A Fifth Order Markov Chain Equivalent



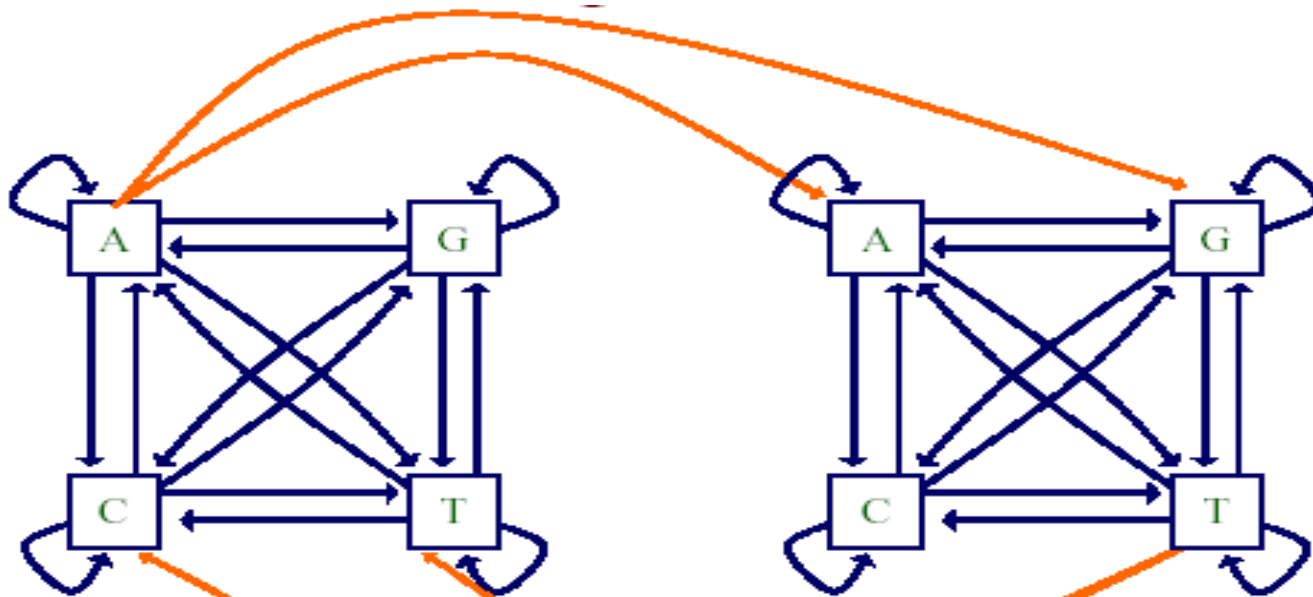
$$\Pr(\text{GCTACA}) = \Pr(\text{GCTAC})\Pr(\text{A} \mid \text{GCTAC})$$

hidden Markov model

Where (in long sequence) are CpG islands?

- first approach: Markov Chains + windowing
- second approach: *hidden Markov model*

Hidden Markov Model: A Simple HMM

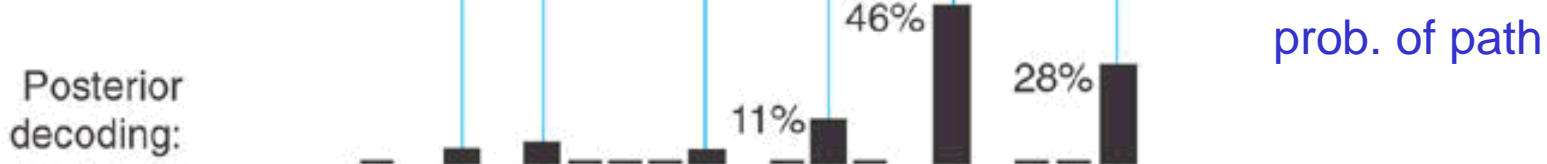
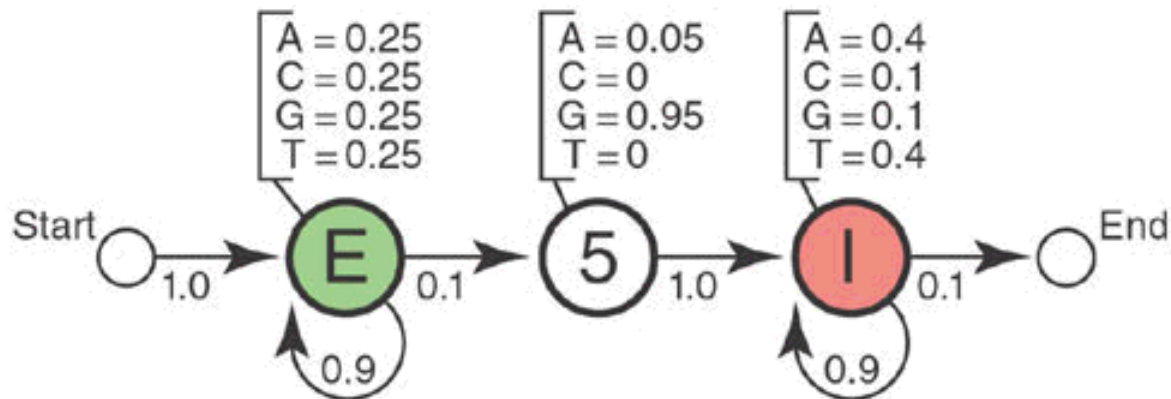


Model 1

Model 2

Given observed sequence AGGCT, which state emits which item?

Another example: Eddy (2004)



$$P(s_i=E | X)$$

An (toy) HMM for 5' splice site recognition.

Figure from: What is a hidden Markov model?

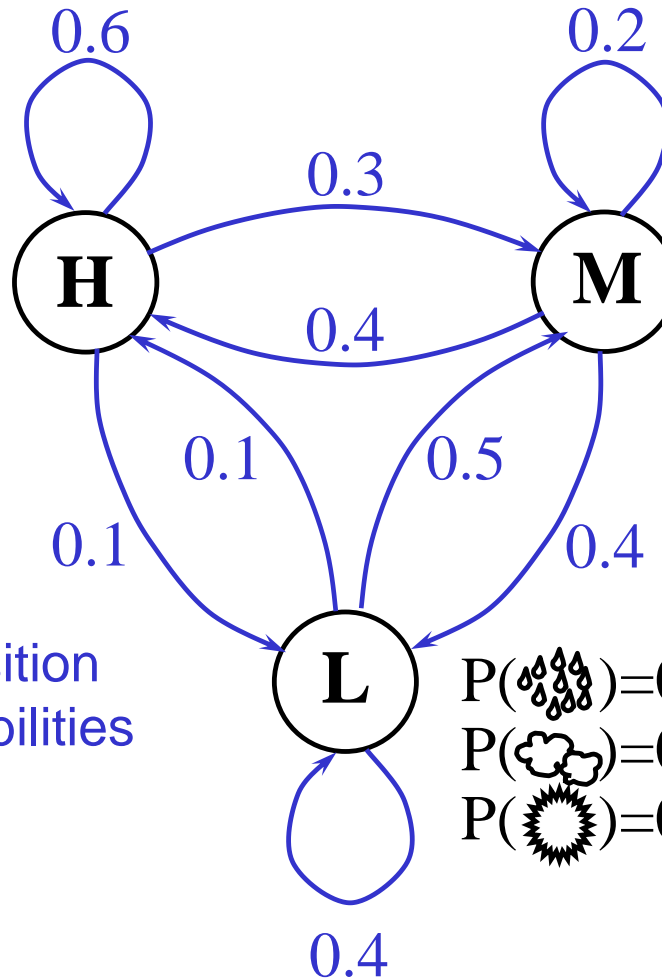
Sean R Eddy. *Nature Biotechnology* **22**, 1315 - 1316 (2004)



Example: weather

emission probabilities

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



$$\begin{aligned}
 P(\text{rain}) &= 0.3 \\
 P(\text{cloud}) &= 0.4 \\
 P(\text{sun}) &= 0.3
 \end{aligned}$$

initial probabilities

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

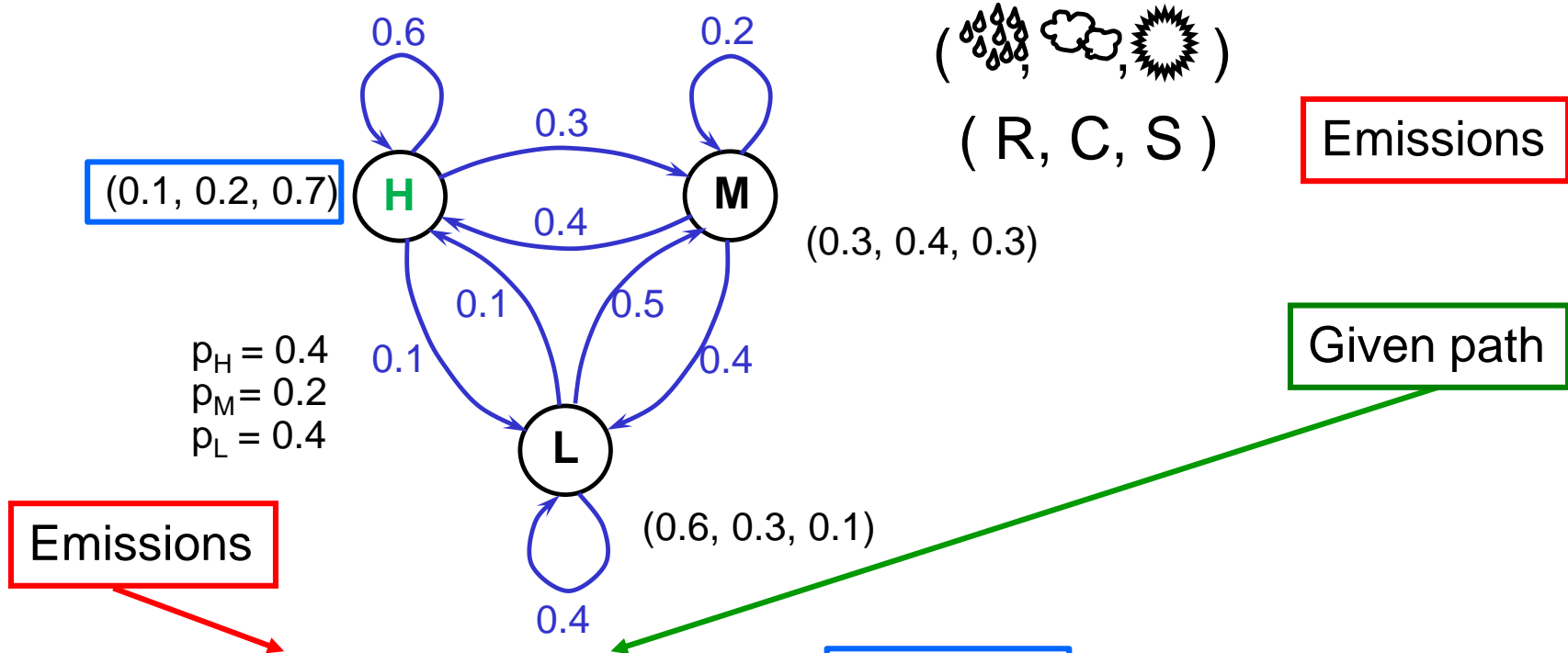
transition probabilities

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

observed weather vs. pressure



Example: weather



Emissions

Given path

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

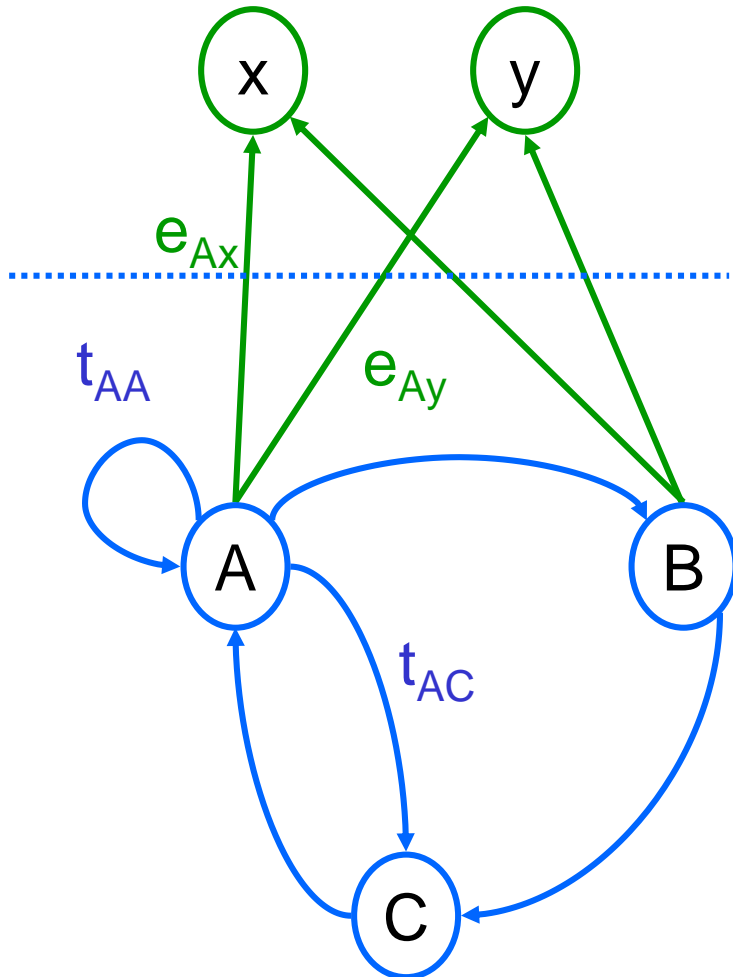
$$P(\text{RCCSS} \mid \text{MMMMM}) = 3 \cdot 4 \cdot 4 \cdot 3 \cdot 3 = 432 \quad (\times 10^{-5})$$

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

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

hidden Markov model

what we see



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

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

observation $X = x_1x_2 \dots x_n \in \Sigma^*$

observe states *indirectly* 'hidden'

- emission probabilities

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

probability

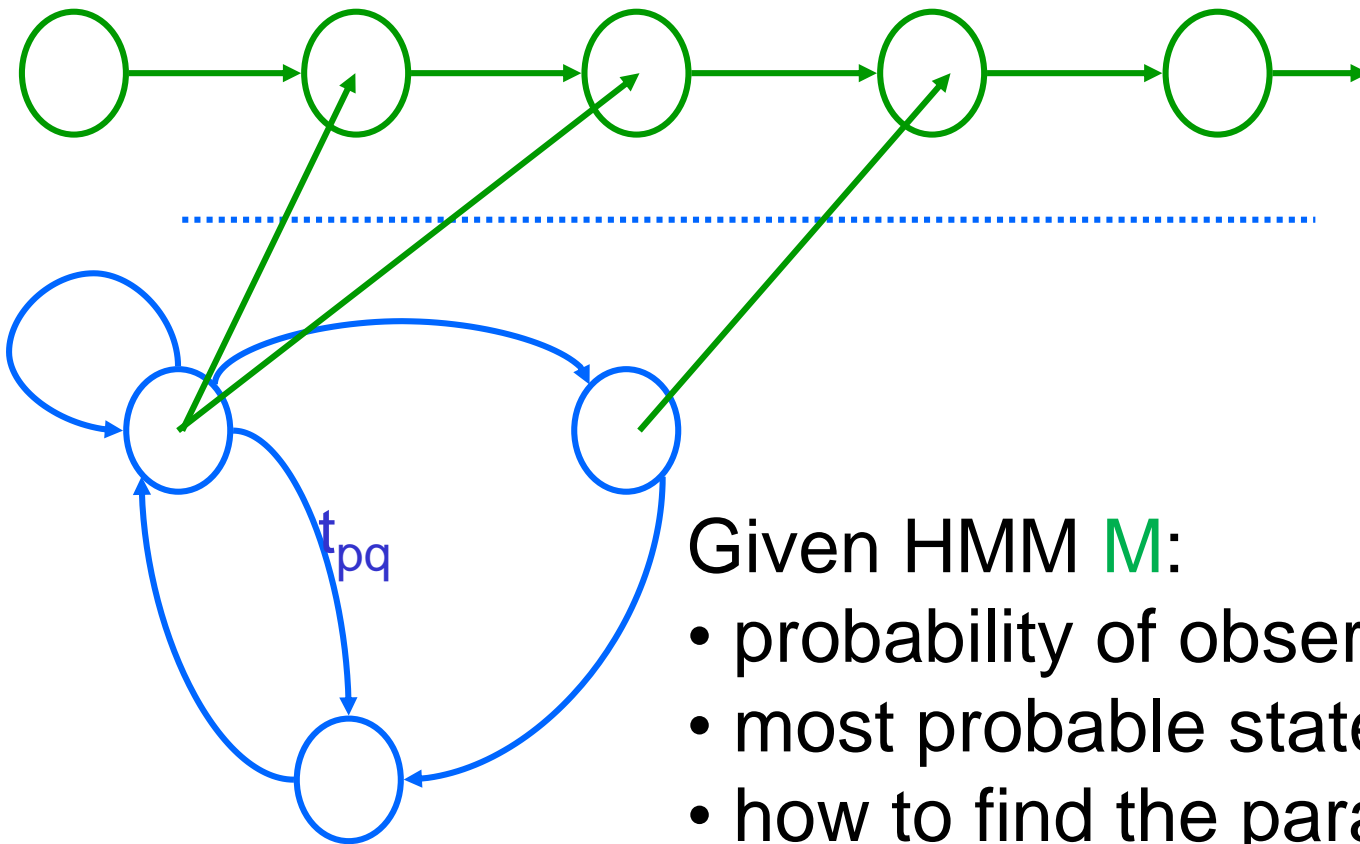
observation given the model

? there may be *many* state seq's

underlying process

HMM main questions

observation $X \in \Sigma^*$

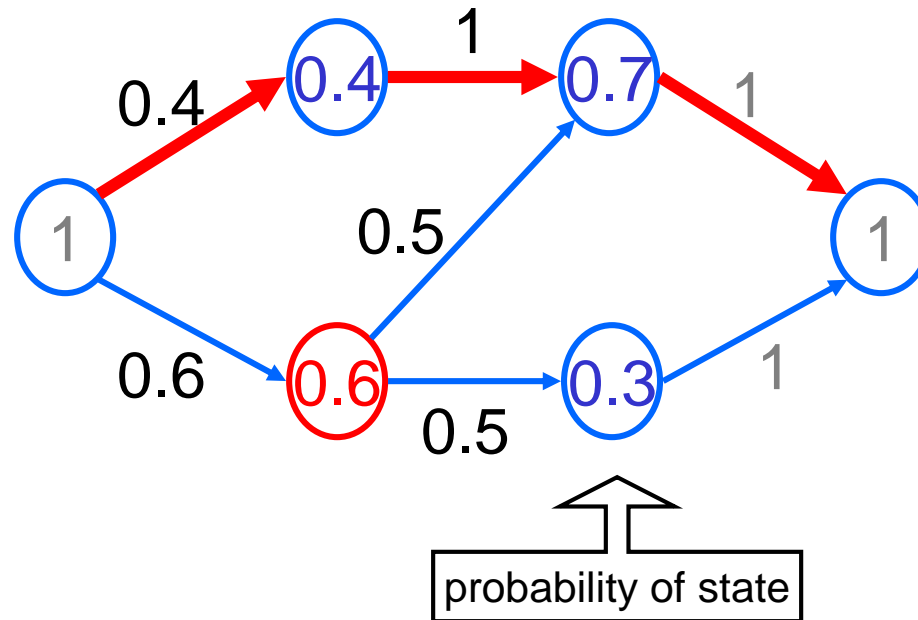


Given HMM M :

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

probability ... !

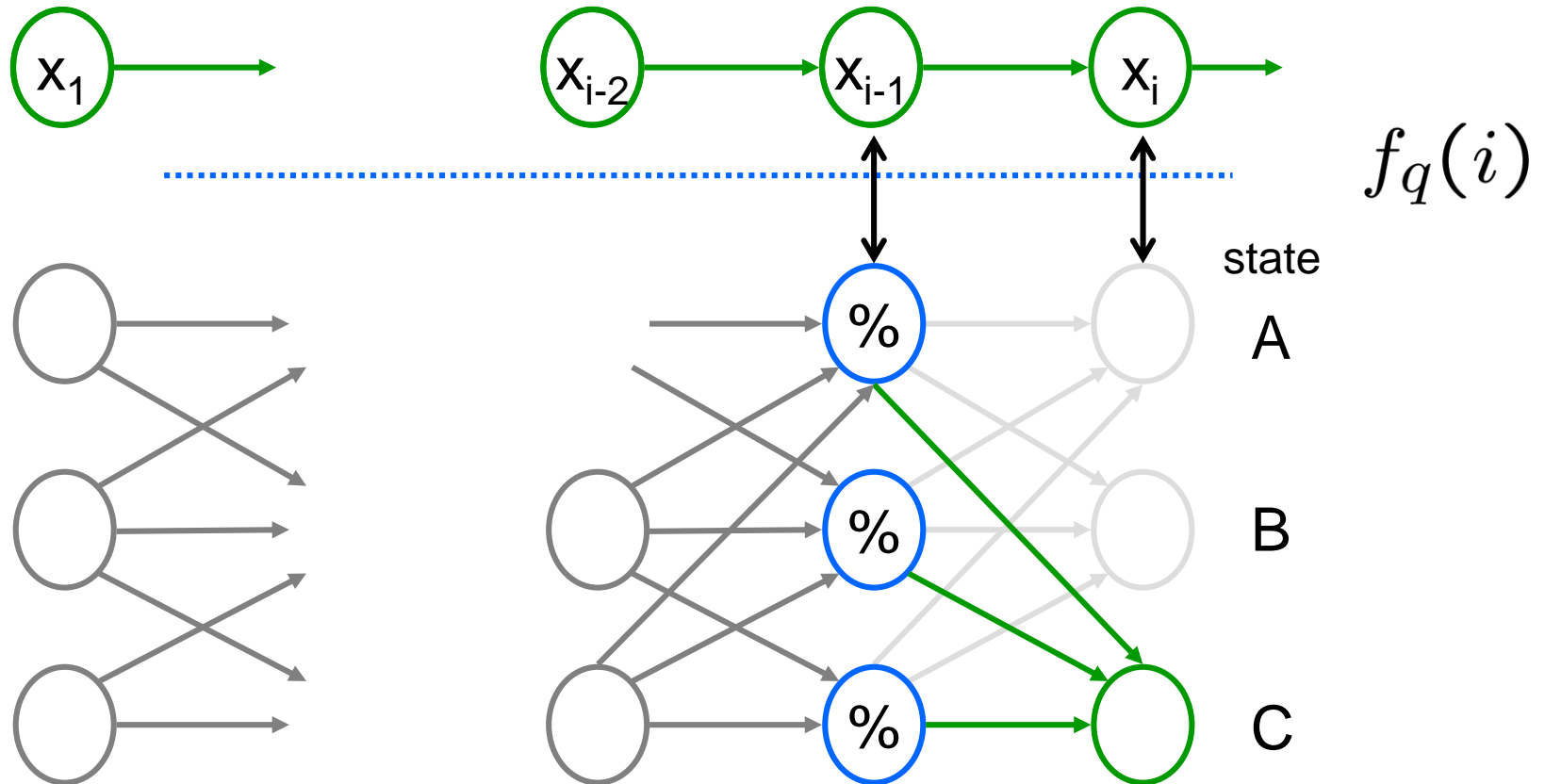
Given sequence X : most probable state vs. optimal path



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

probability of observation X

dynamic programming: probability ending in state q emitting symbol x_i



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

probability of observation X

probability observing x_1, \dots, x_i and ending in state q :

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

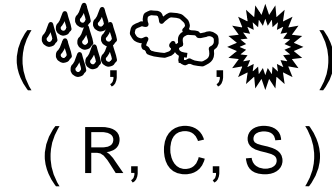
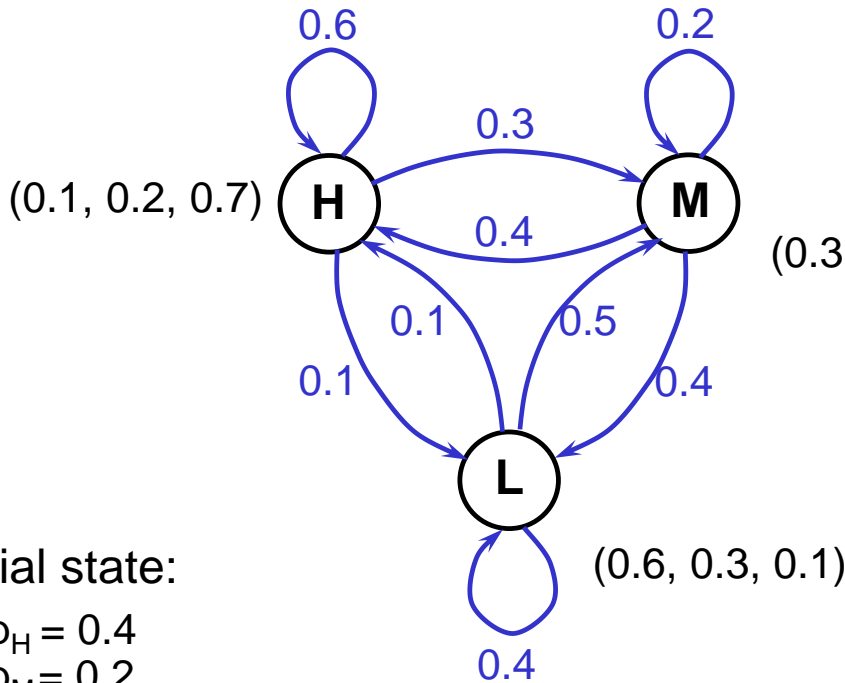
$$f_q(i) = \sum_{p \in Q} f_p(i-1) t_{pq} e_q(x_i)$$

‘forward’ probability

$$P(X) = \sum_{p \in Q} f_p(n) t_{p*} \quad * = \text{end-state}$$



Probability of observation: weather



(0.3, 0.4, 0.3)

Transitions:

- Remain in H
- Coming from M
- Coming from L

Initial state:

$p_H = 0.4$
 $p_M = 0.2$
 $p_L = 0.4$

$$P(RCCSS) = P(RC\dots)$$

1:R

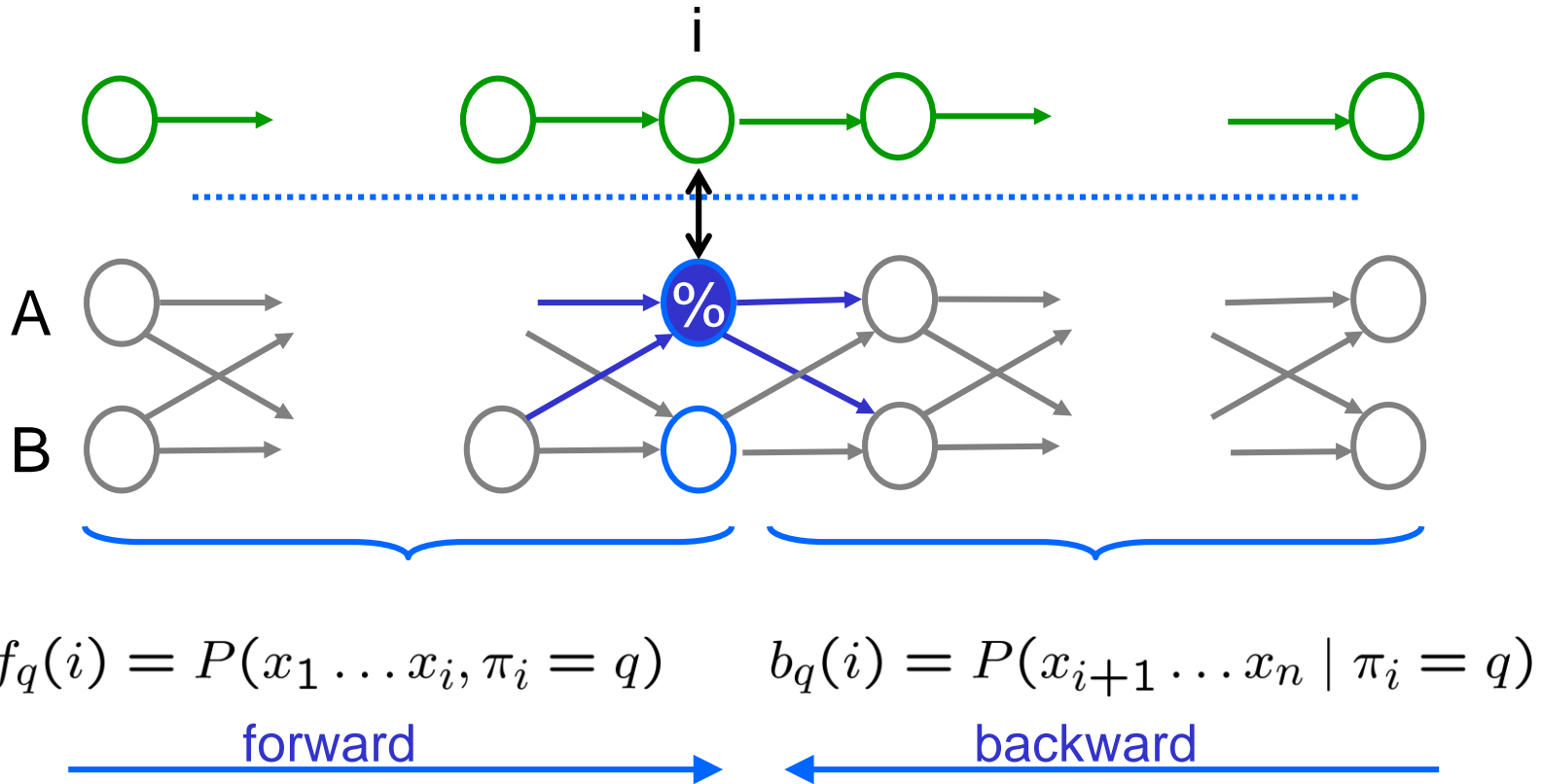
2:C

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	$4 \cdot 6 = 24$	$(4 \cdot 1 + 6 \cdot 4 + 24 \cdot 4) \cdot 3 = 372$	$(\times 10^{-4})$

Start: 0 1

posterior decoding

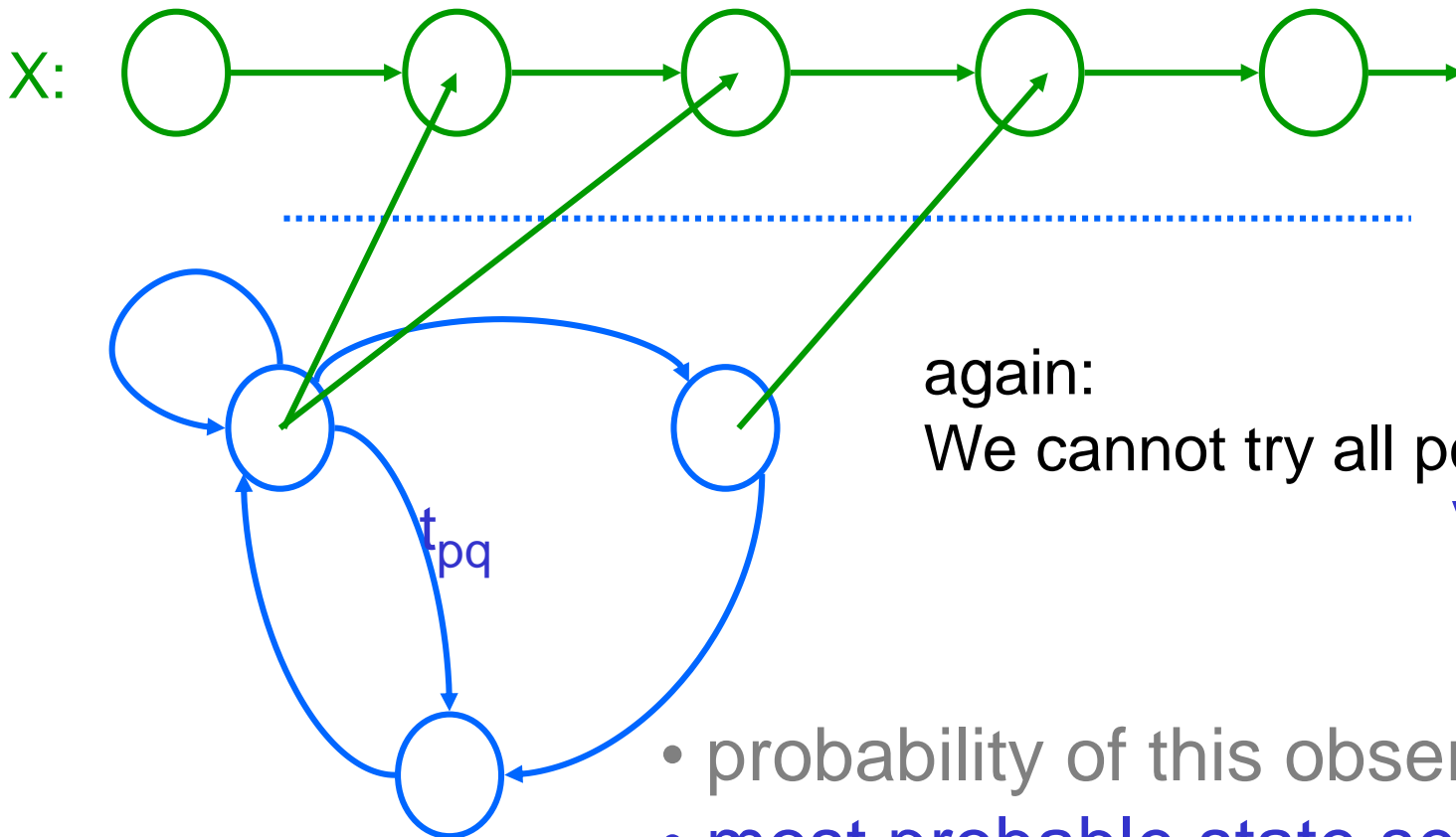
$P(\pi_i = q \mid X)$ i-th state equals q



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

HMM main questions

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



again:

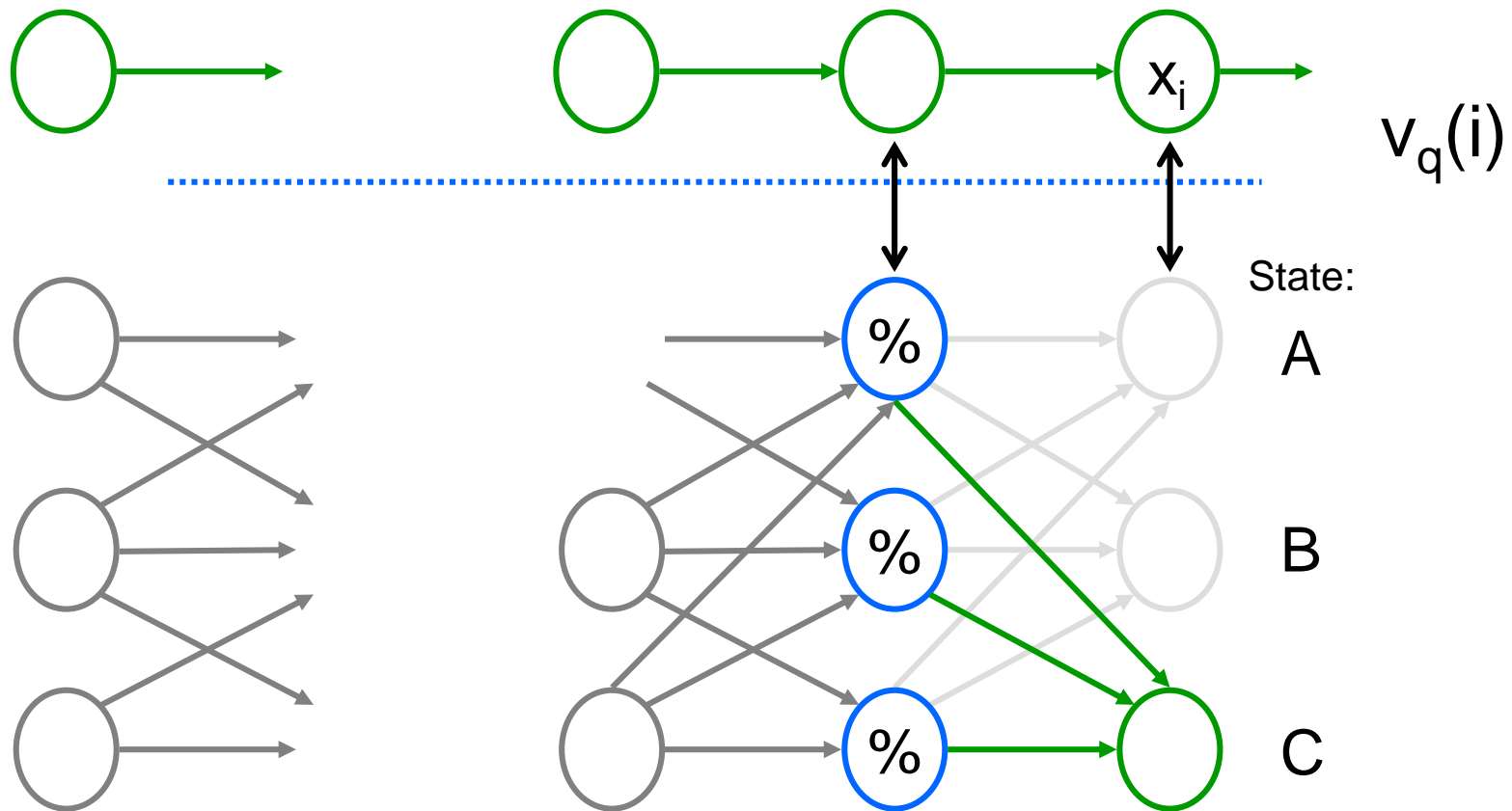
We cannot try all possibilities
Viterbi

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

Viterbi algorithm

most probable state sequence for observation X

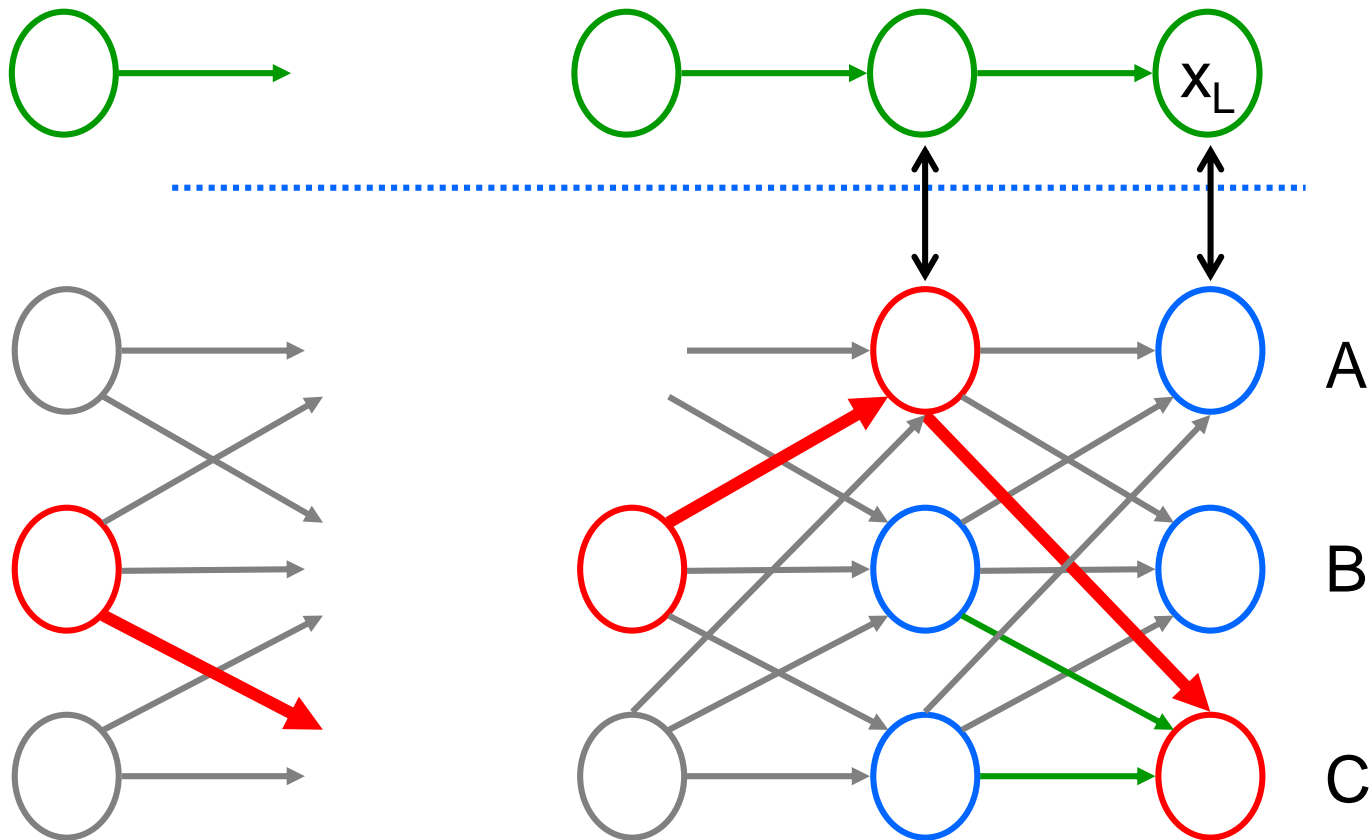
(1) *dynamic programming*: $v_q(i)$ probability ending in state q and emitting x_i



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

Viterbi algorithm

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



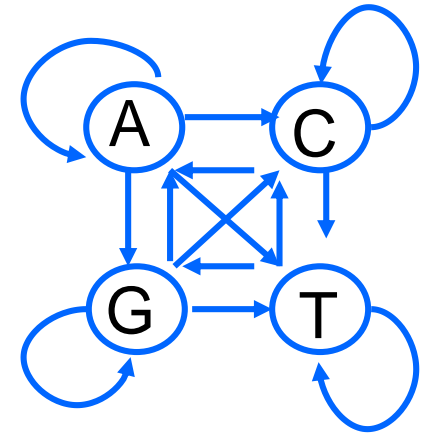
HMM Example: CpG islands

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

0.999

estimates

+	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



0.001

0.00001

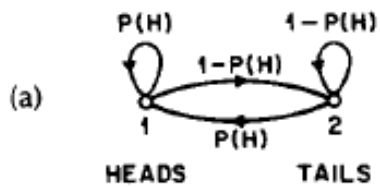
Transition Matrix

-	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

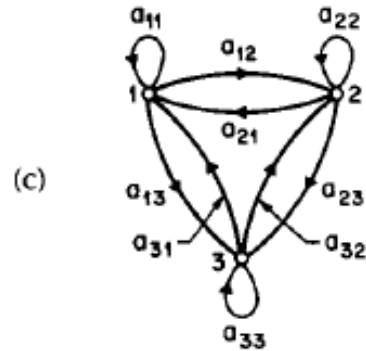
0.99999

45

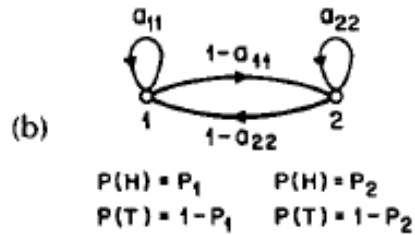
HMM for Hidden Coin Tossing



O = HHTTHTHHTTH...
 S = 11221211221...



O = HHTTHTHHTTH...
 S = 31233112313...

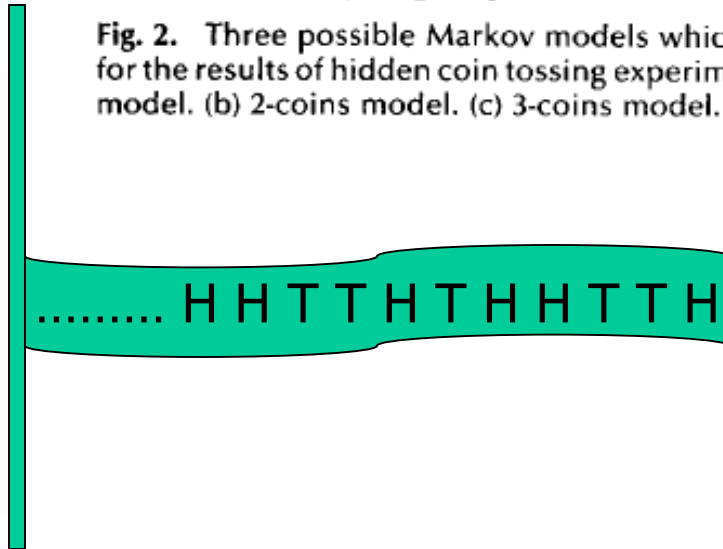
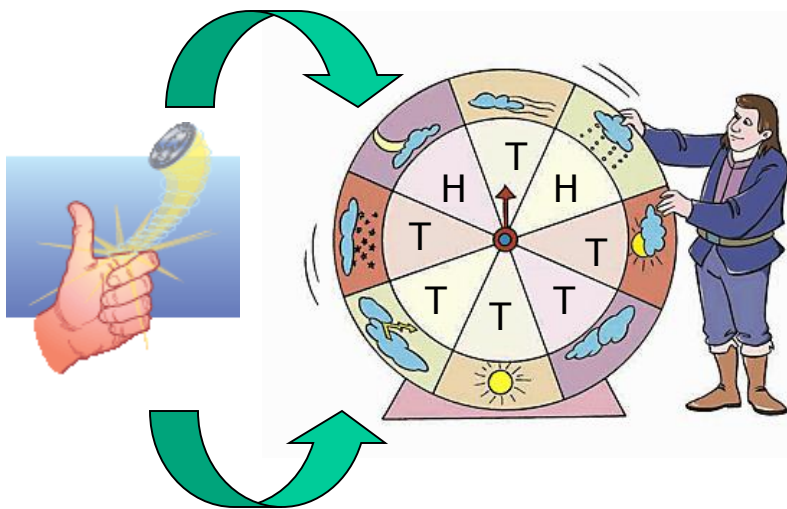


O = HHTTHTHHTTH...
 S = 21122212212...

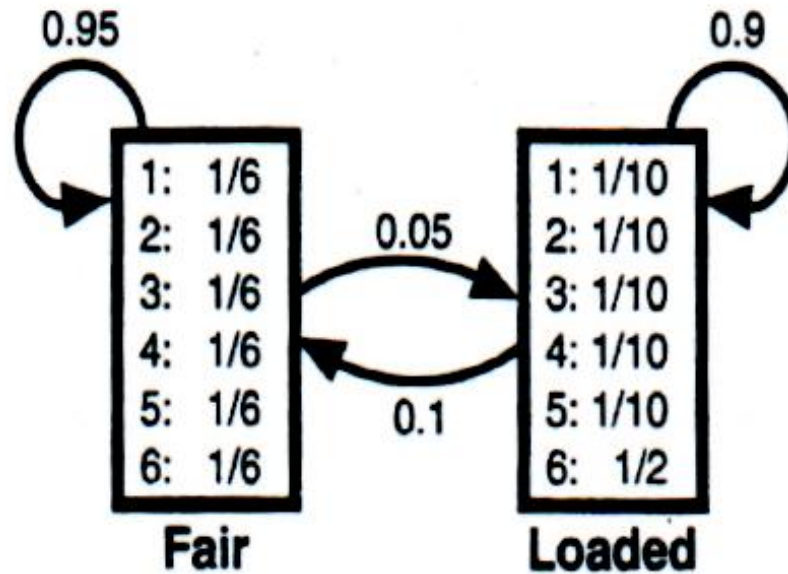
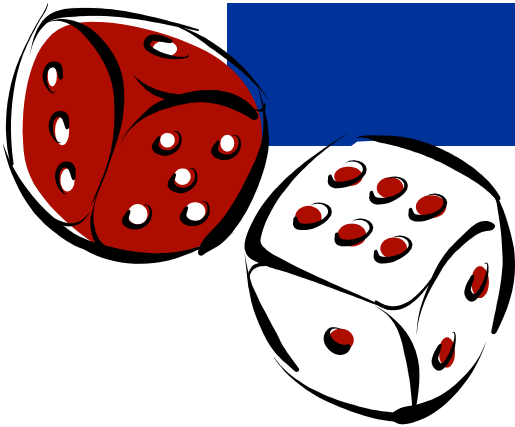
STATE

	1	2	3
P(H)	P_1	P_2	P_3
P(T)	$1-P_1$	$1-P_2$	$1-P_3$

Fig. 2. Three possible Markov models which can account for the results of hidden coin tossing experiments. (a) 1-coin model. (b) 2-coins model. (c) 3-coins model.



dishonest casino dealer



dishonest casino dealer

Rolls	315116246446644245321131631164152133625144543631656626566666
Die	FFL
Viterbi	FFL
Rolls	651166453132651245636664631636663162326455235266666625151631
Die	LLLLLLFFFFFFFFFFFFFFFFL
Viterbi	LLLLLLFFFFFFFFFFFFFFFFL
Rolls	222555441666566563564324364131513465146353411126414626253356
Die	FFFFFFFFLLL
Viterbi	FFL
Rolls	366163666466232534413661661163252562462255265252266435353336
Die	LLLLLLLLLFFF
Viterbi	LLLLLLLLLFFF
Rolls	233121625364414432335163243633665562466662632666612355245242
Die	FFL
Viterbi	FFL

dishonest casino dealer

Observation

366163666466232534413661661163252562462255265252266435353336

Viterbi

LLLLLLLLLLLLLFF

Compare to:

Forward

FFLLLLLLLLLLLLLFFFFFFFFLFLLLFLLFFFFFFFFFFFFFFFFFFFFFFFFLFFFFFFFF

Posterior (total)

LLLLLLLLLLLLLFFFFFFFFLFLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Sketch: Parameter estimation

training sequences $X^{(i)}$

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

Markov Chain: 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

HMM: 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 (sketch):

- determine optimal paths
- re-compute as if paths are known
- score may decrease!

Important Papers on HMM

L.R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceeding of the IEEE, Vol. 77, No. 22, February 1989.

Krogh, I. Saira Mian, D. Haussler, A Hidden Markov Model that finds genes in E. coli DNA, Nucleid Acids Research, Vol. 22 (1994), pp 4768-4778

Furthermore:

R. Hassan, A combination of hidden Markov model and fuzzy model for stock market forecasting, Neurocomputing archive, Vol. 72 , Issue 16-18, pp 3439-3446, October 2009.