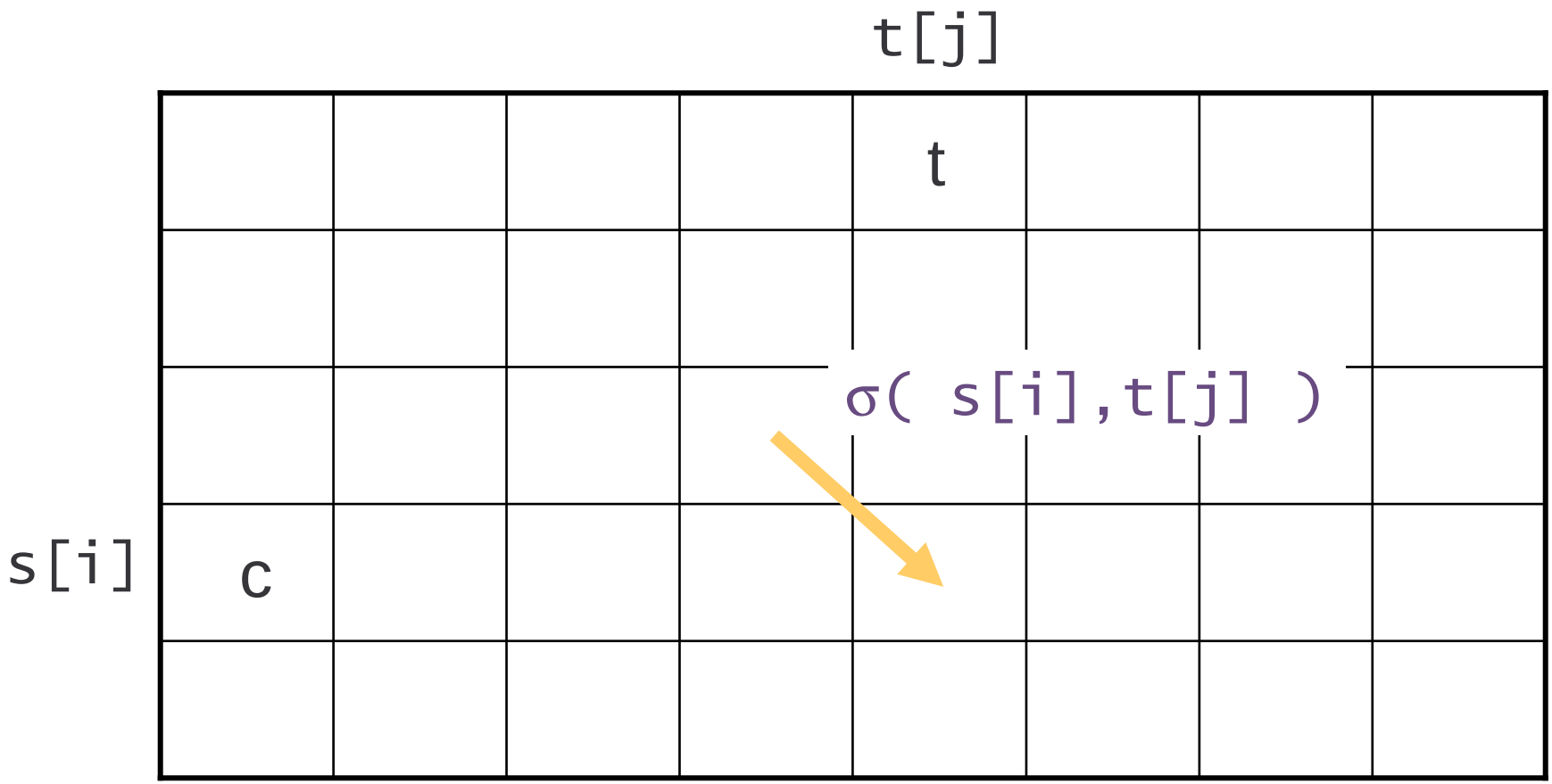
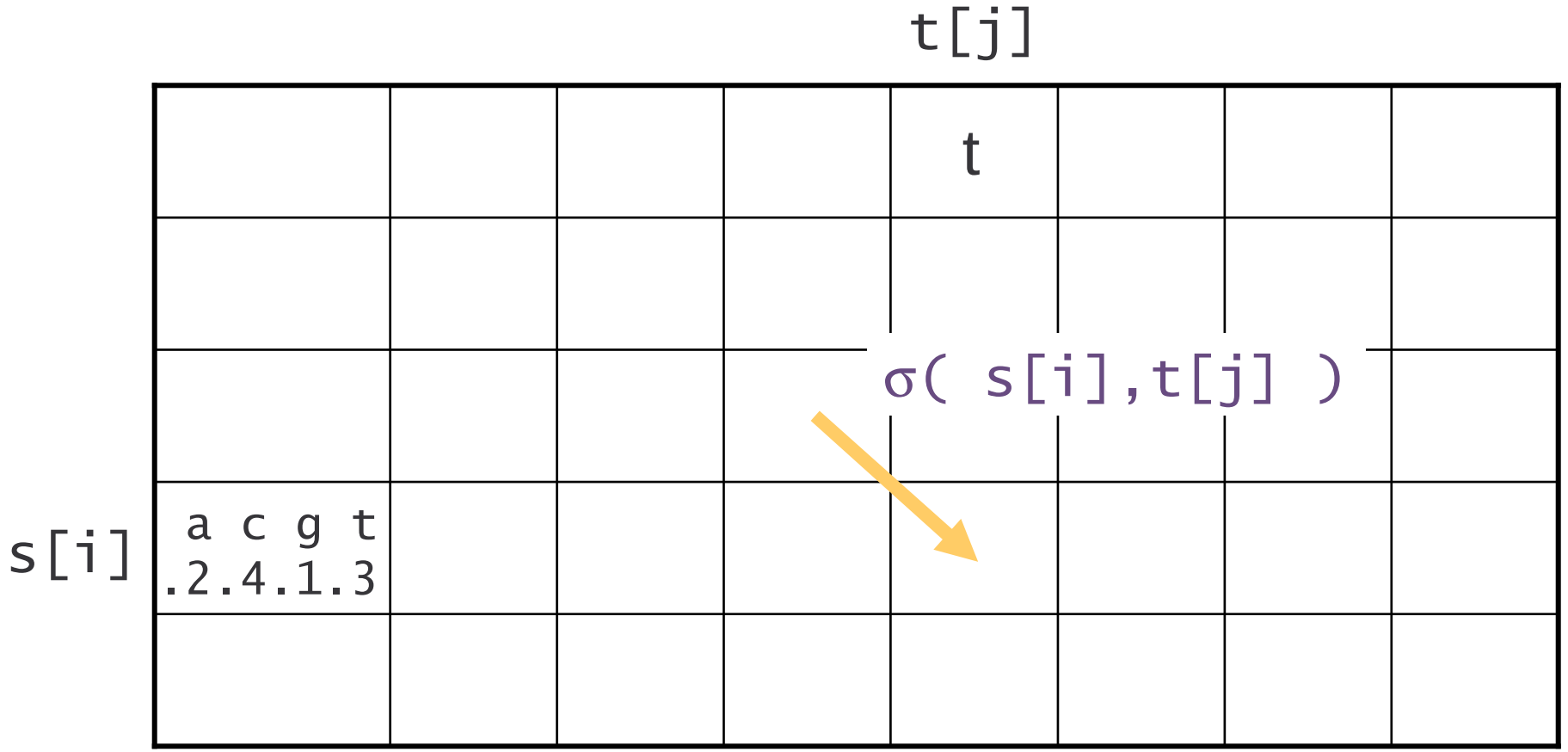
profile alignment

alignment



profile alignment



$$s[i] = (s_a[i],\ s_c[i],\ s_g[i],\ s_t[i])$$

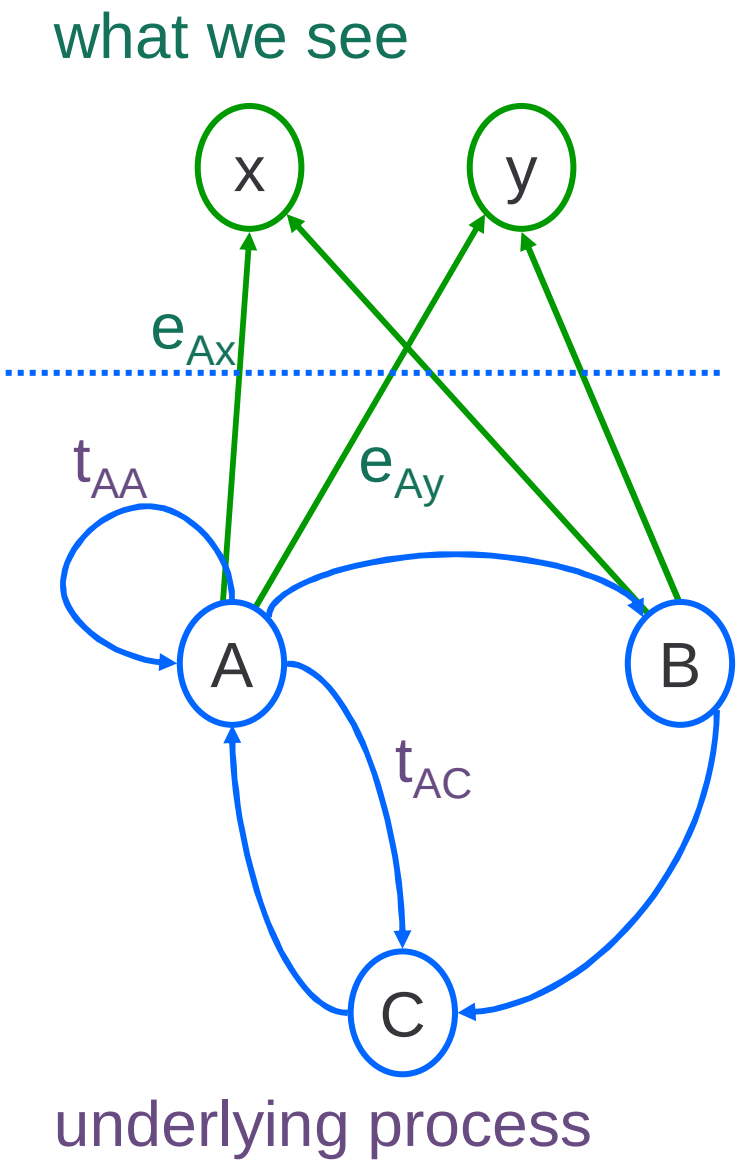
$$\sigma(s[i],t[j]) = \sum_x s_x[i] \cdot \sigma(x,t[j])$$

Markov models

review

and … your questions

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

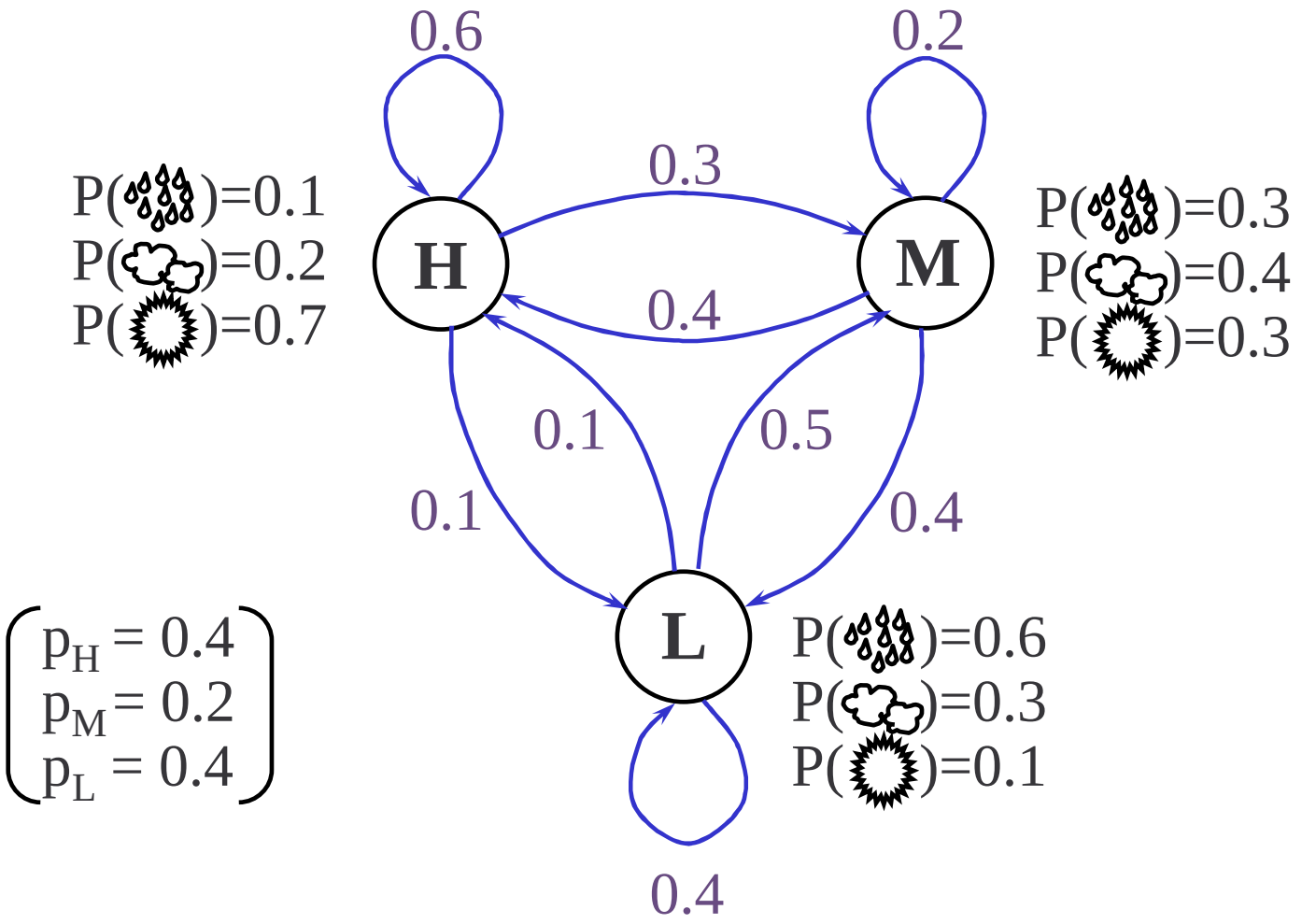
$e_{px},\ p \in Q,\ x \in \Sigma \qquad e_p(x)$

probability

observation given the model

? there may be *many* state seq's

weather



observed weather vs. pressure

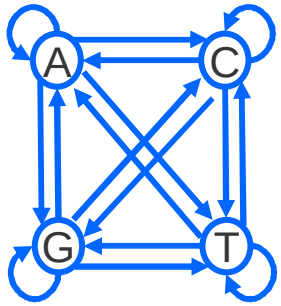
CpG islands ctd.

8 states A^+ vs A^-

unique observation each state

p

+	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



1-p

1-q

-	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

q

64 transitions!

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

next character C or G
positive contribution

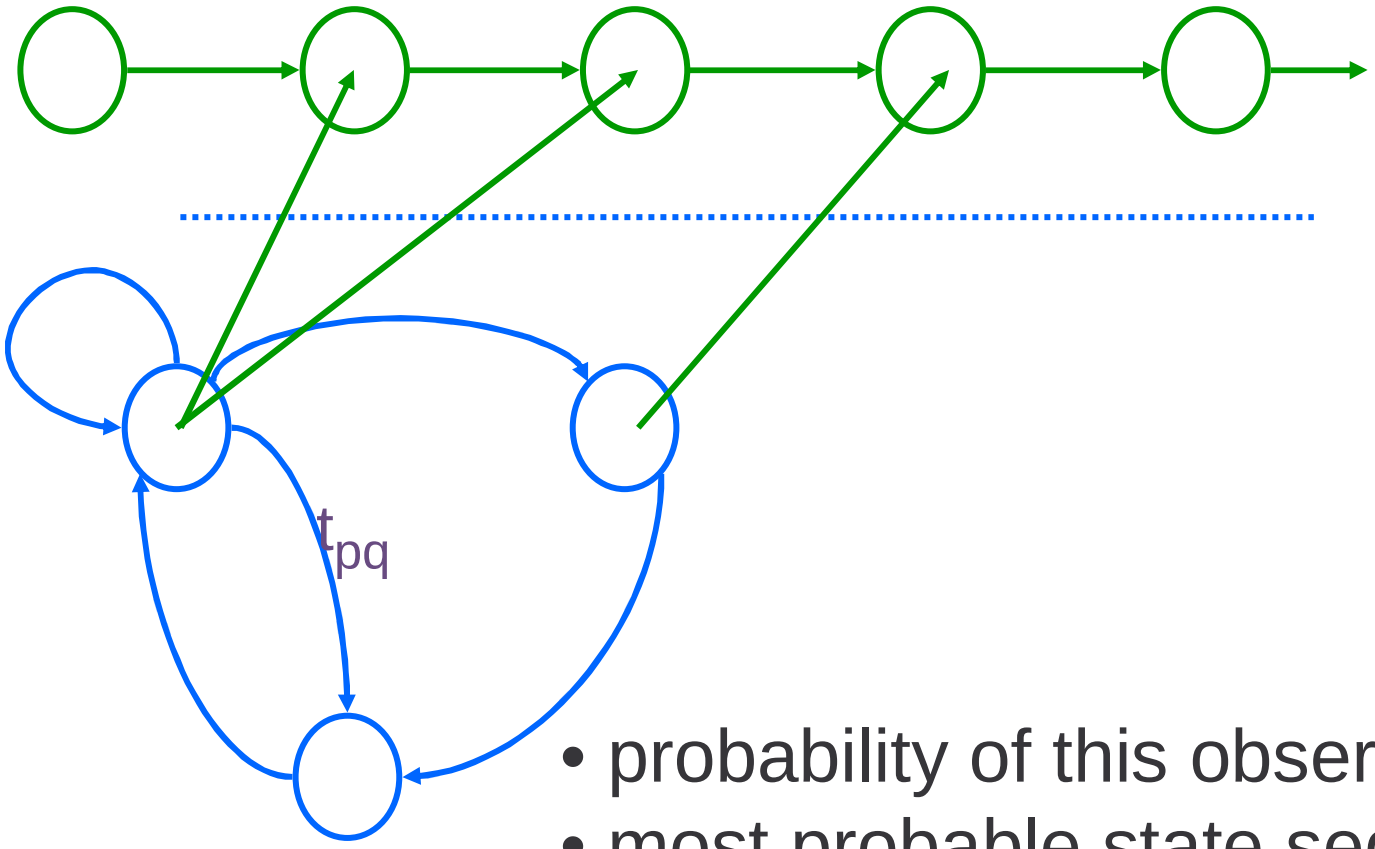
do we measure CG content?

- no answer -

a propos …

HMM main questions

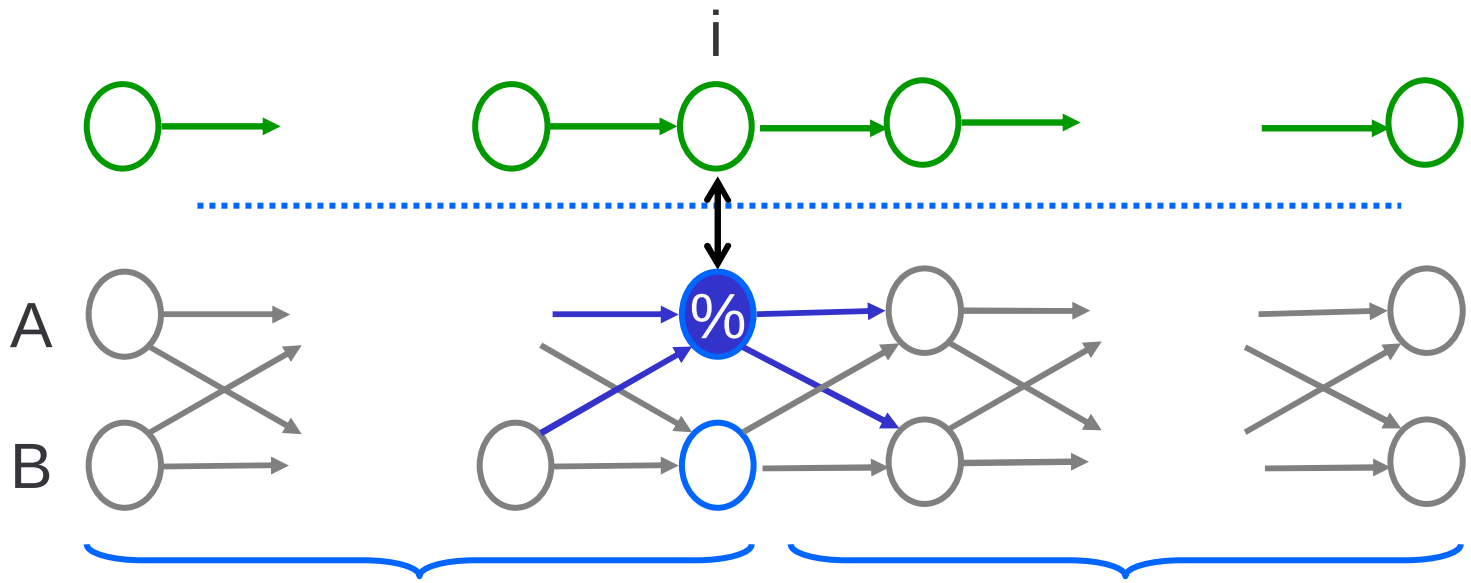
observation $X \in \Sigma^*$



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

posterior decoding

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



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

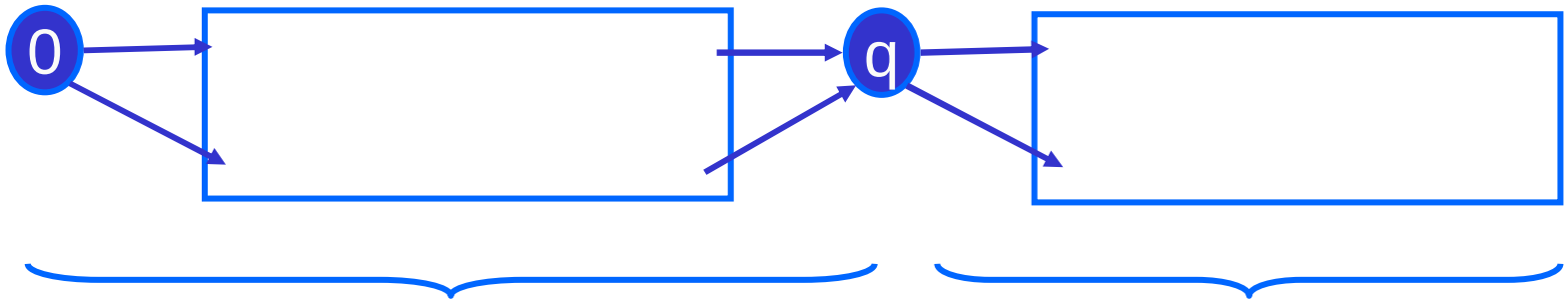
$b_q(i) = P(x_{i+1} \ldots x_n \mid \pi_i = q)$

forward

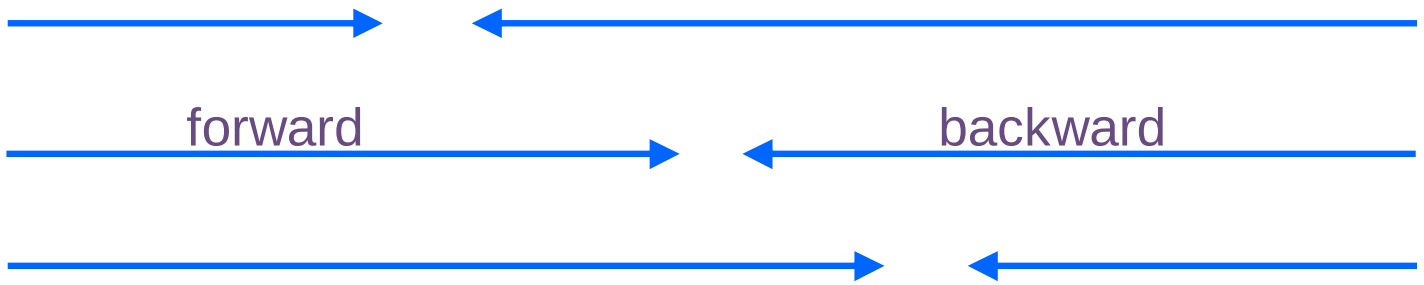
backward

posterior decoding

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

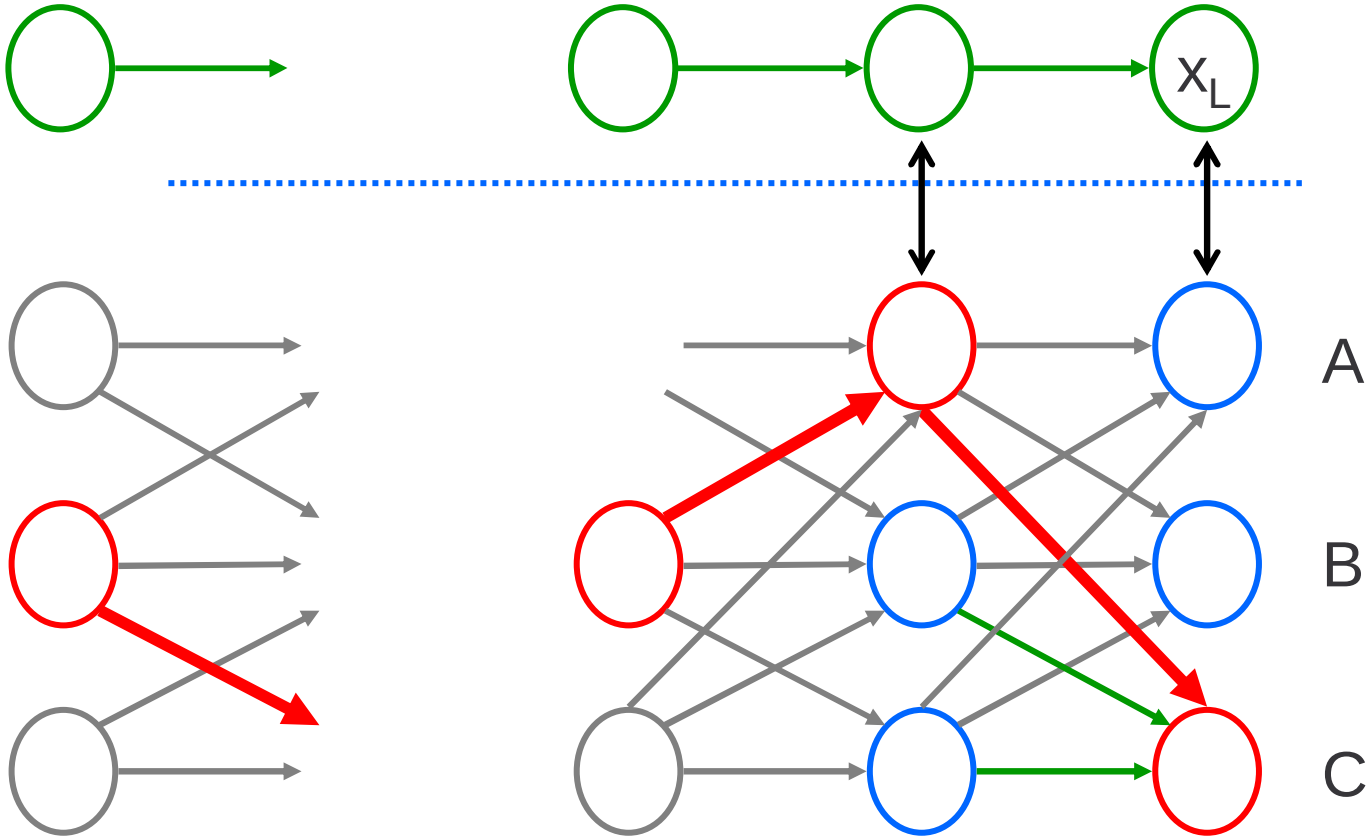


$f_q(i) = P(x_1 \dots x_i, \pi_i = q)$ $b_q(i) = P(x_{i+1} \dots x_n \mid \pi_i = q)$



Viterbi algorithm

(1) *dynamic programming*: max probability ending in state
(2) *traceback*: most probable state sequence



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

two explanations

posterior Σ

best state every position

☹ path may not be allowed by model

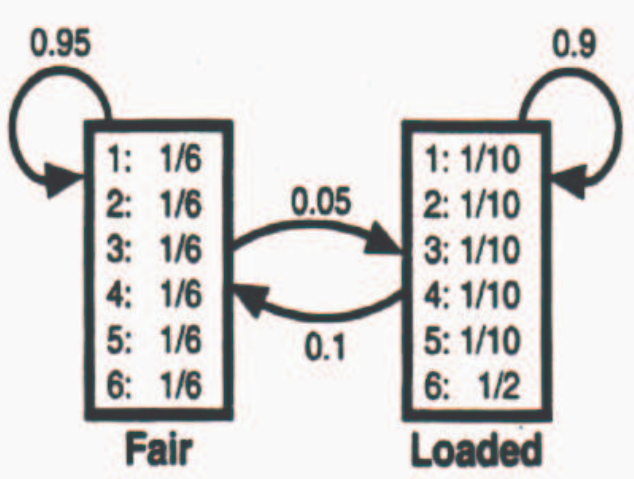
viterbi max

optimal global path

☹ many paths with similar probability

dishonest casino dealer

```
Rolls      315116246446644245321131631164152133625144543631656626566666
Die        FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLL
Viterbi    FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLL
-----------------------------------------------------------------------
Rolls      651166453132651245636664631636663162326455235266666625151631
Die        LLLLLLFFFFFFFFFFFLLLLLLLLLLLLLLLLFFFLLLLLLLLLLLLLLFFFFFFFFF
Viterbi    LLLLLLFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF
-----------------------------------------------------------------------
Rolls      222554416665665635643243641315134651463534111264146262533 56
Die        FFFFFFFLLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL
Viterbi    FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL
-----------------------------------------------------------------------
Rolls      366163666466232534413661661163252562462255265252266435353336
Die        LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi    LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFF
-----------------------------------------------------------------------
Rolls      233121625364414432335163243633 66
Die        FFFFFFFFFFFFFFFFFFFFFFFFFLLLLI
Viterbi    FFFFFFFFFFFFFFFFFFFFFFFFFFFLLI
```



experiment: rolls, die

reconstruction: rolls → die (viterbi)

Baum-Welch

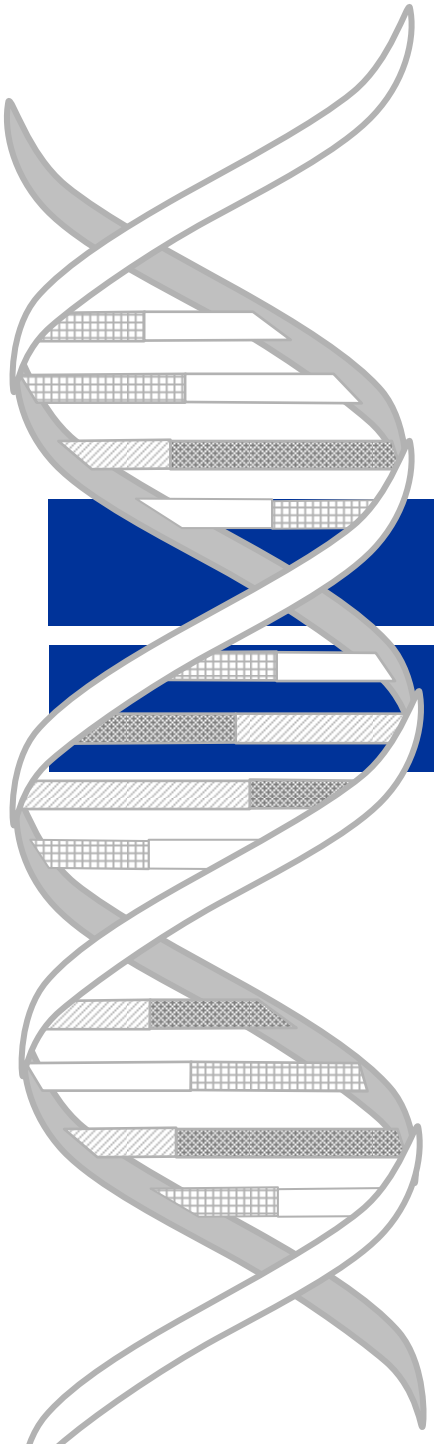
state sequences unknown

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

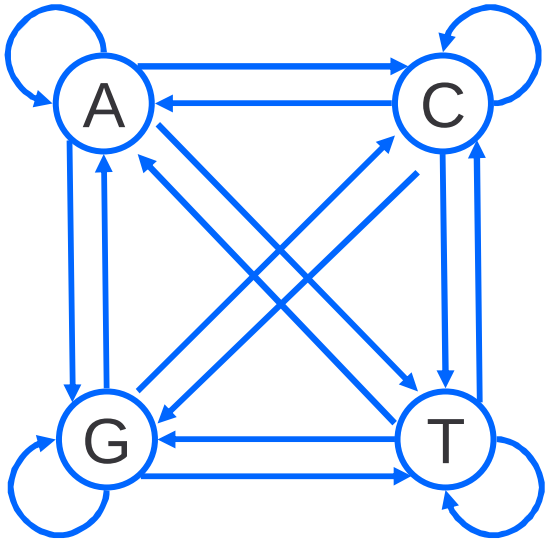
$E_p(b)$ sum over all training sequences X
sum over all positions i with x_i=b



Hidden Markov Models II

applications

model structure



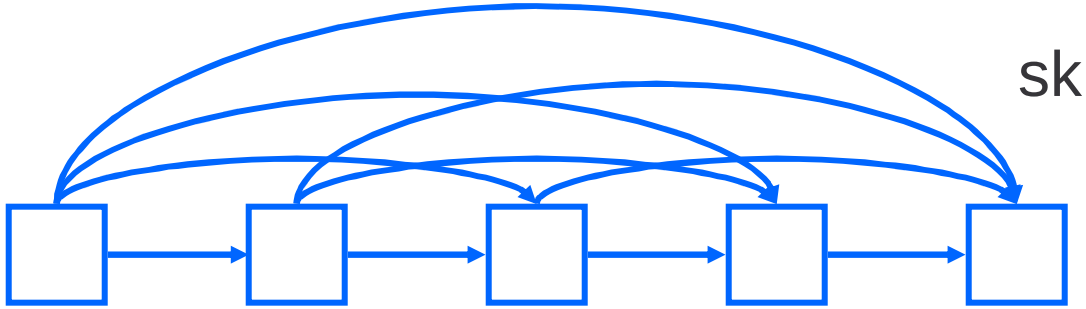
many states & fully connected
training seldom works
local maxima

use knowledge about the problem

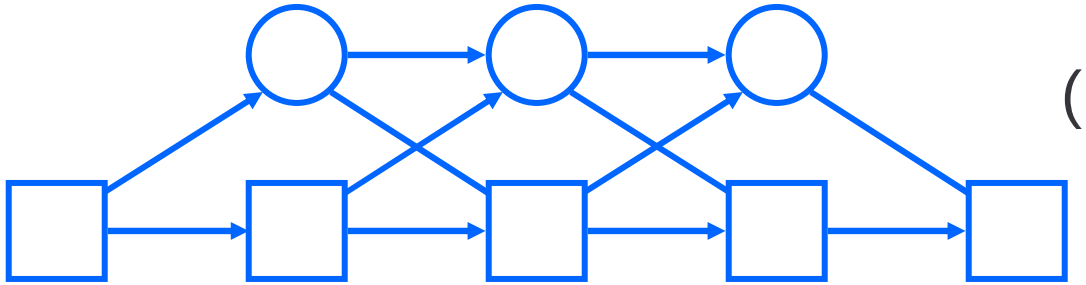
e.g. linear model for *profile alignment*
(using HMM, later)

silent states

skipping nodes

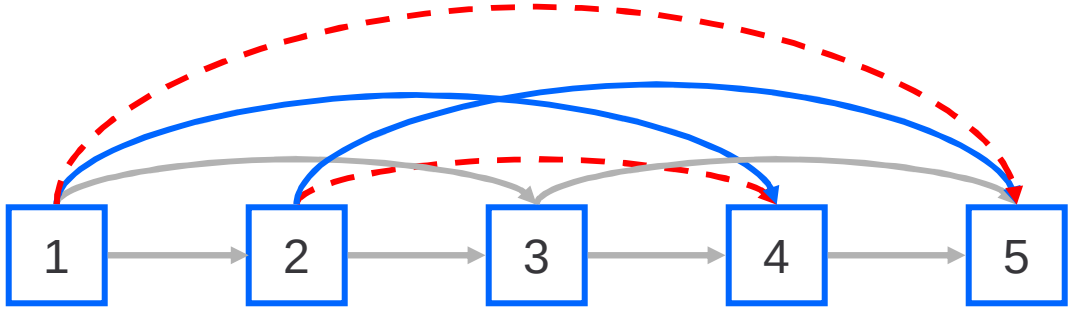


[square states]



silent states
(no emission)

quadratic vs. linear size
but less modelling possibilities

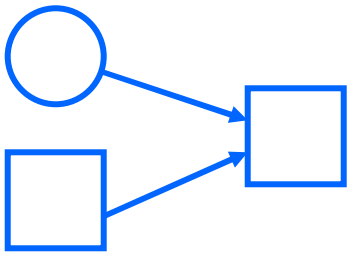


high/low

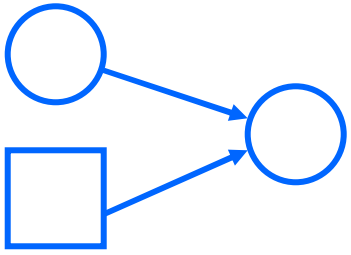
silent states: algorithm

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

transition / emission



as before



for silent states

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

no silent loops (!):
update in ‘topological order’

profile alignment

ungapped
transition prob's 1
trivial alignment HMM to sequence

VGAHAGEY
VTGNVDEV
VEADVAGH
VKSNDVAD
VYSTVETS
FNANIPKH
IAGNGAGV

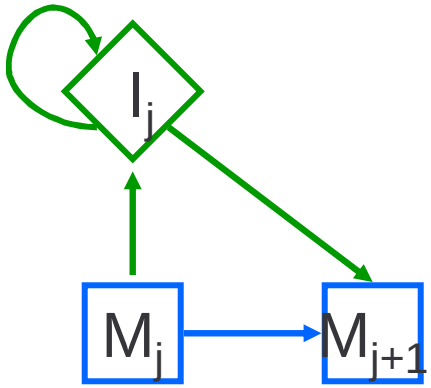
profile HMM $\mathcal{P}$ 'dedicated topology'

$$X = x_1 x_2 \ldots x_L$$

$e_i(b)$ observing symbol b at pos i

$$P(X|\mathcal{P}) = \prod_{i=1}^{L} e_i(x_i)$$

profile alignment (2)



insert state

match states

```
VGA--HAGEY
VNA--NVDEV
VEA--DVAGH
VKG--NYDED
VYS--TYETS
FNA--NIPKH
IAGADNGAGV
```

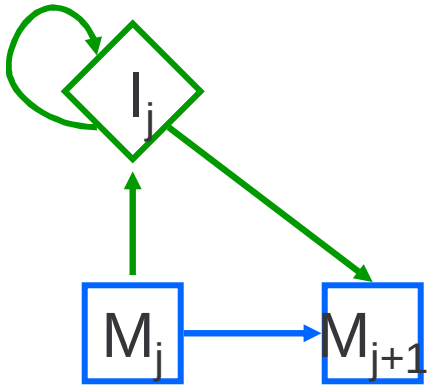
emission: background probabilities
or based on alignment

affine model

$$t_{M_j I_j} \cdot t_{I_j M_{j+1}} \cdot t_{I_j I_j}^{h-1}$$

open gap extension

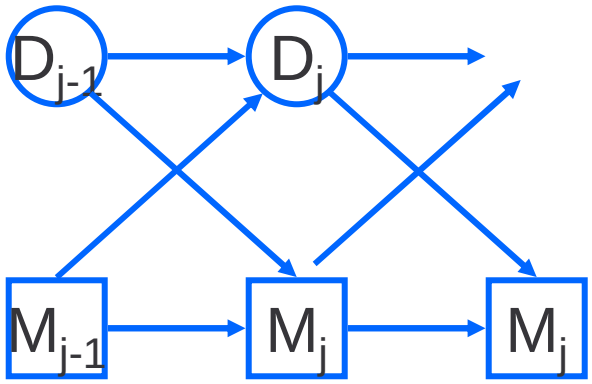
profile alignment (3)



insert state

match states

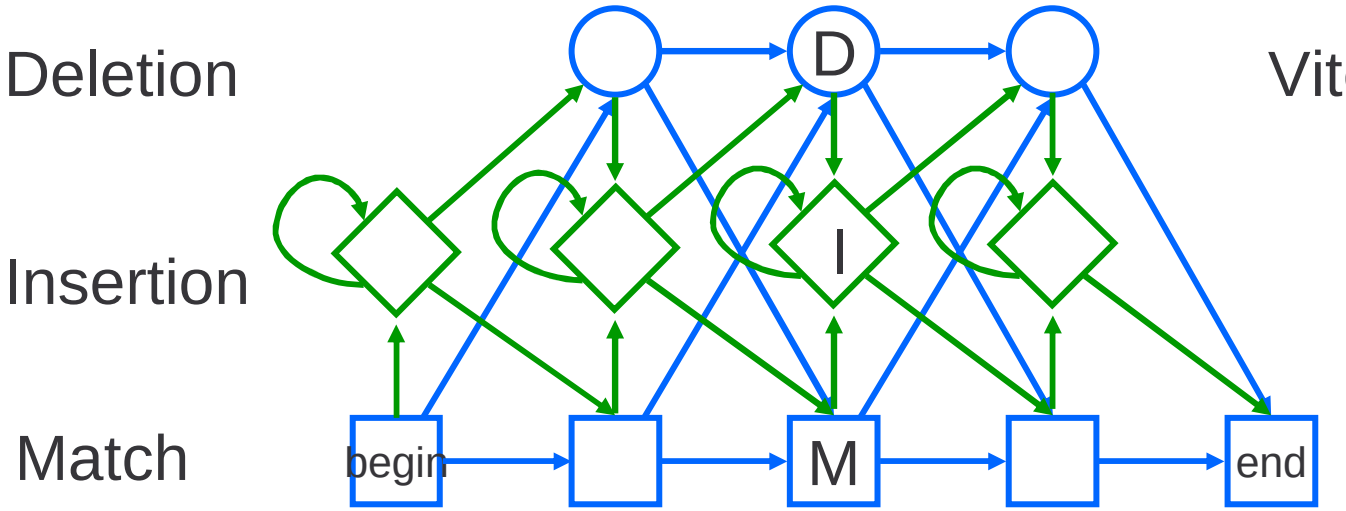
```
VGA--HAGEY
V----NVDEV
VEA--DVAGH
VKG------D
VYS--TYETS
FNA--NIPKH
IAGADNGAGV
```



delete state
(*silent*)

adapt Viterbi

HMM for profiles / multiple alignment



$$v_j^M(i) = e_{M_j}(x_i) \cdot \max_{Y=M,I,D} v_{j-1}^Y(i-1)\, t_{Y_{j-1}M_j}$$

$$v_j^I(i) = p(x_i) \cdot \max_{Y=M,I,D} v_j^Y(i-1)\, t_{Y_jM_j}$$

same level

$$v_j^D(i) = \max_{Y=M,I,D} v_{j-1}^Y(i)\, t_{Y_{j-1}M_j}$$

same position

profile alignment

given multiple alignment
Insertion / Deletion states

```
VGA--HAGEY
V----NVDEV
VEA--DVAGH
VKG------D
VYS--TYETS
FNA--NIPKH
IAGADNGAGV
123  45678
```

counting

transitions

$M_1 \rightarrow M_2$	6+1	$7/_{10}$
$M_1 \rightarrow I_1$	0+1	$1/_{10}$
$M_1 \rightarrow D_1$	1+1	$2/_{10}$

emissions

F	1+1	$2/_{27}$
I	1+1	$2/_{27}$
V	5+1	$6/_{27}$
other 17x	0+1	$1/_{27}$

Laplace correction

multiple alignment with profile

```
IAGADNGAGV
123  45678
```

```
VGAHAGEY
12345678
```

```
FNAPNI-KH
123 45678
```

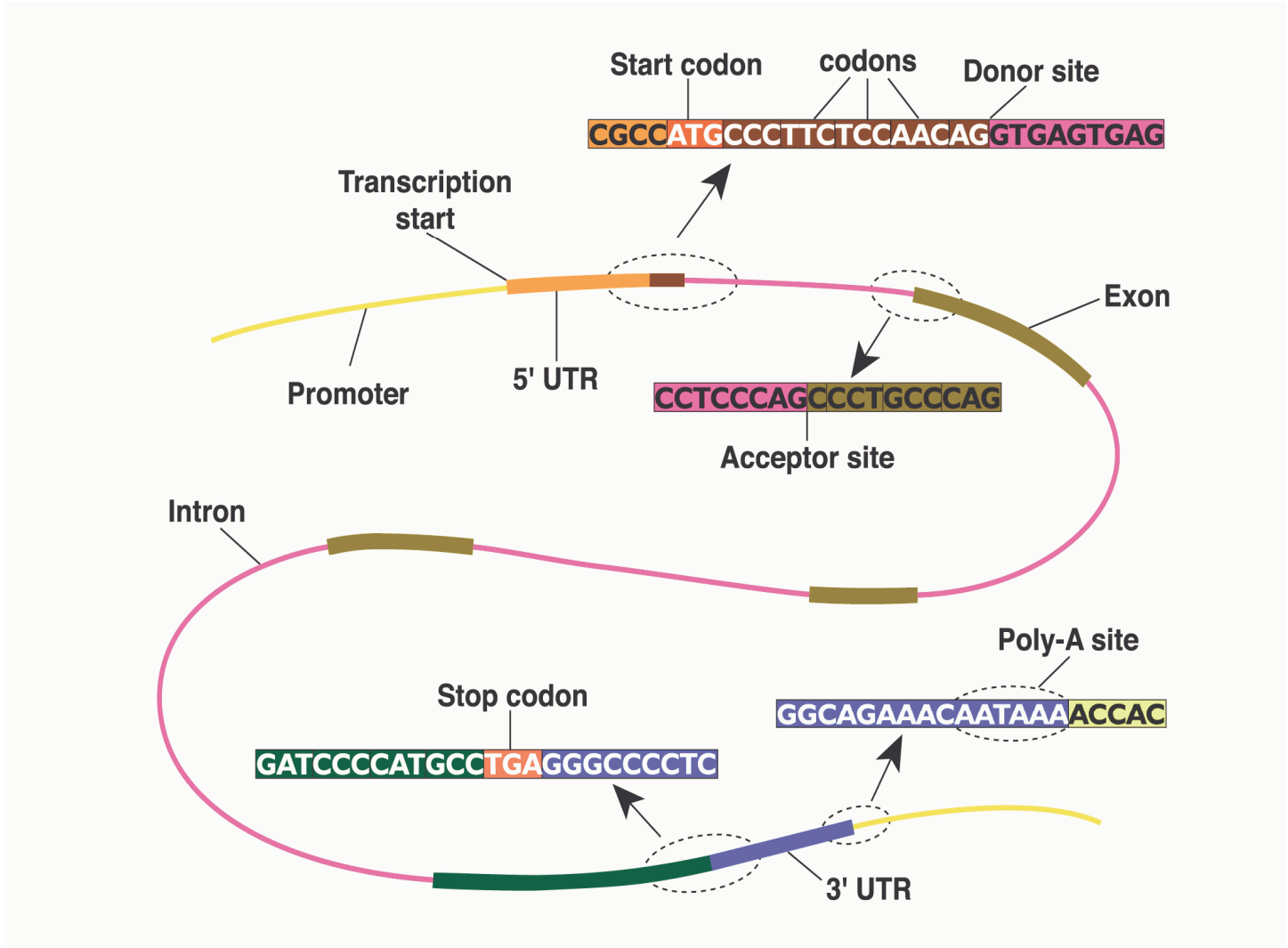
```
VGA--HAGEY
FNAP-NI-KH
IAGADNGAGV
123  45678
```

align each sequence separately

accumulate alignments
M and D positions

align inserts leftmost
I positions

application: gene finding



gene finding

central dogma:

DNA transcription RNA translation protein

only 2%-3% coding … find these regions

Prokaryotes vs. Eukariotes
- no nucleus
- most of genome is coding
- continuous genes vs. introns & exons

‘signals’

ORFs

open reading frames

3 (or 6)

start AUG
stop UAA, UAG, UGA

3/64 stops (random)
average protein 1000bp [much longer]

search for long ORFs
- miss short genes
- too many found

genes are not random

motto

codon frequencies

Leu	Leucine	6 codons	6.9
Ala	Alanine	4	6.5
Trp	Tryptophan	1	1
		‘random’	coding

A or T in 2nd position sometimes 90%

Markov models

- codon triplets as states [64 states]
  ~ 3rd order (but no overlap)
- triplet frequencies only
  keep 3 frames in sliding window cf. CpG

promotor regions

‘consensus’ sequence i.e. not exact

… n TTGAC n^{18} TATAAT n^{6} N n …

TATA box

start coding

position weight matrix

pos	1	2	3	4	5	6
A	2	95	26	59	51	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96

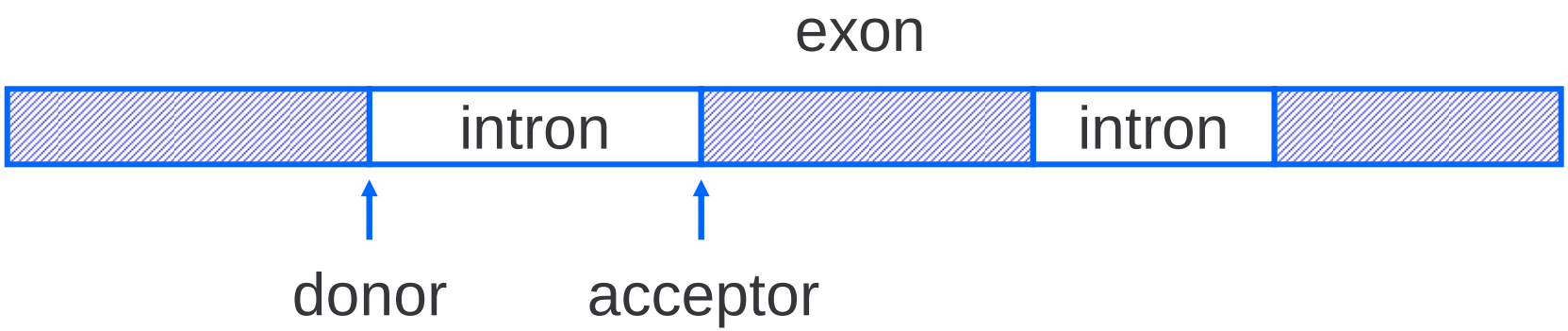
cf. ‘profile’

wmm weight matrix model

wam + dependencies between adjacent positions

Eukaryotes

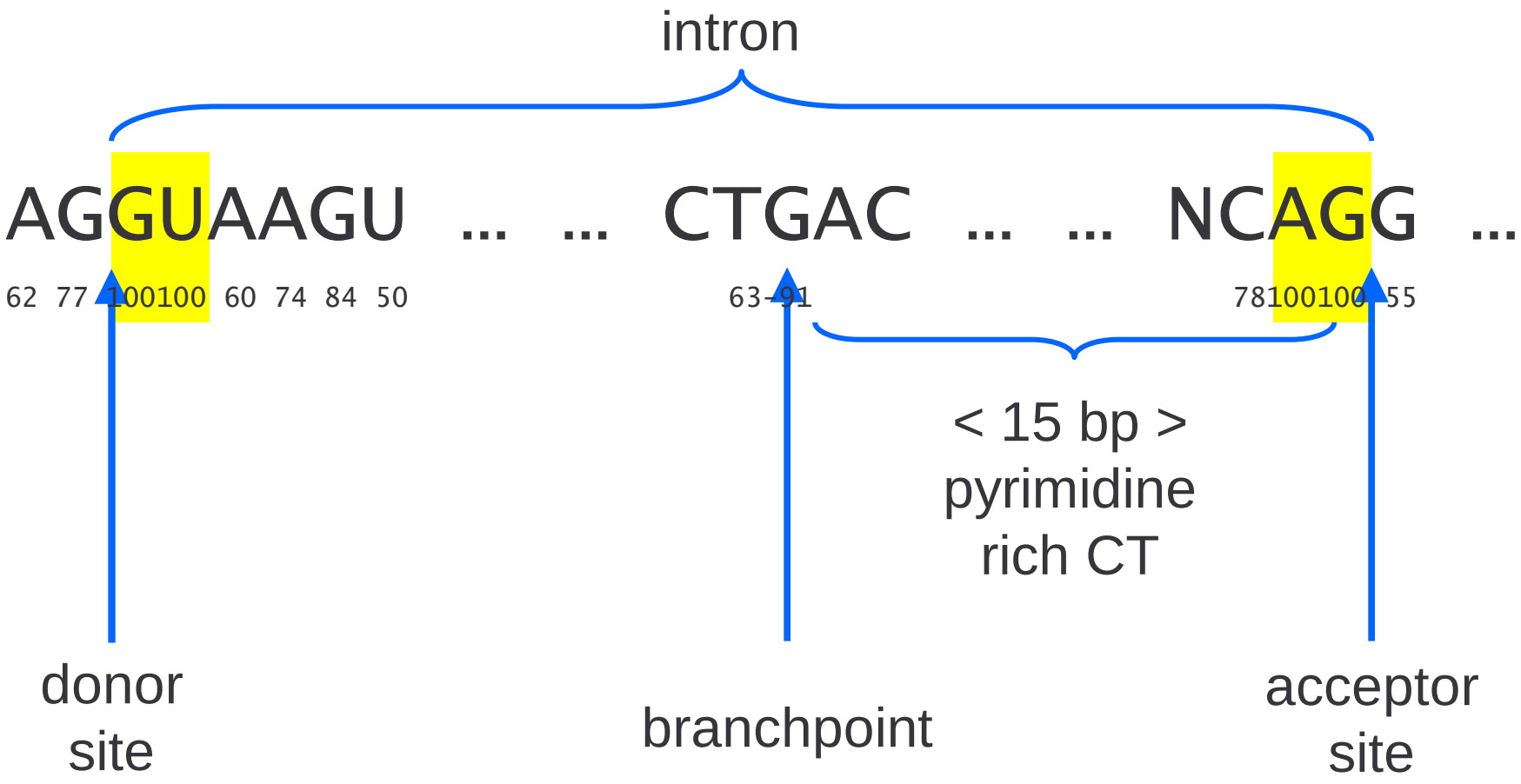
exons expressed
introns noncoding
(alternative) splicing



tss transcription start site
polyA polyadenylation
utr untranslated region
5’ tss and start codon
3’ stop codon and polyA

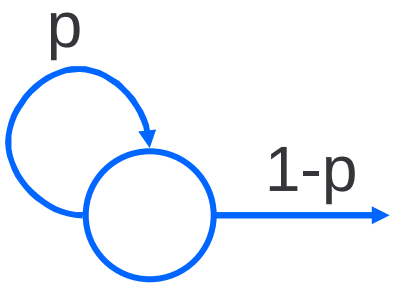
introns: splicing

consensus sequences / weight matrices

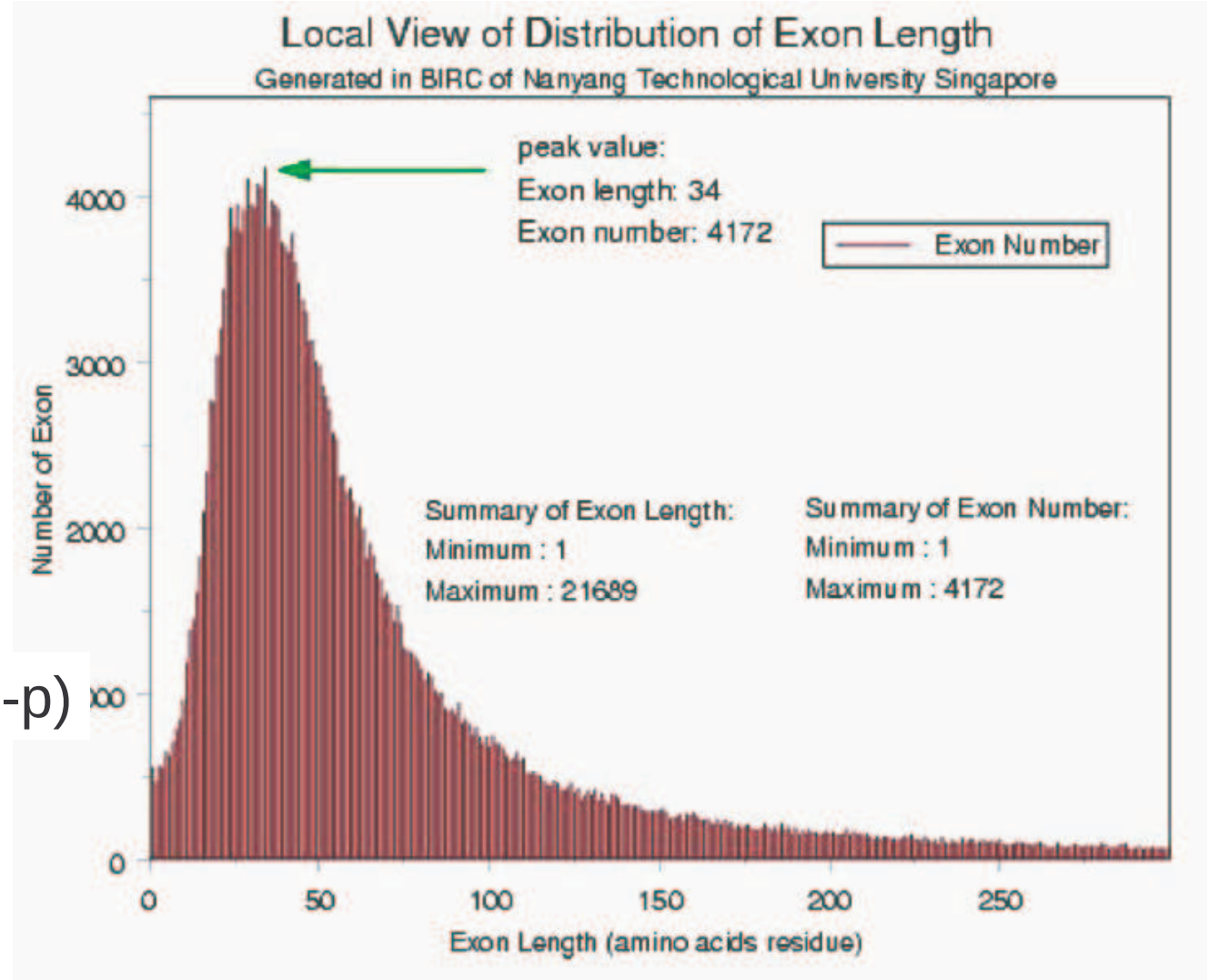


exon lengths

HMM cannot model arbitrary length distributions



$P(len=k) = p^k(1-p)$



generalized HMM

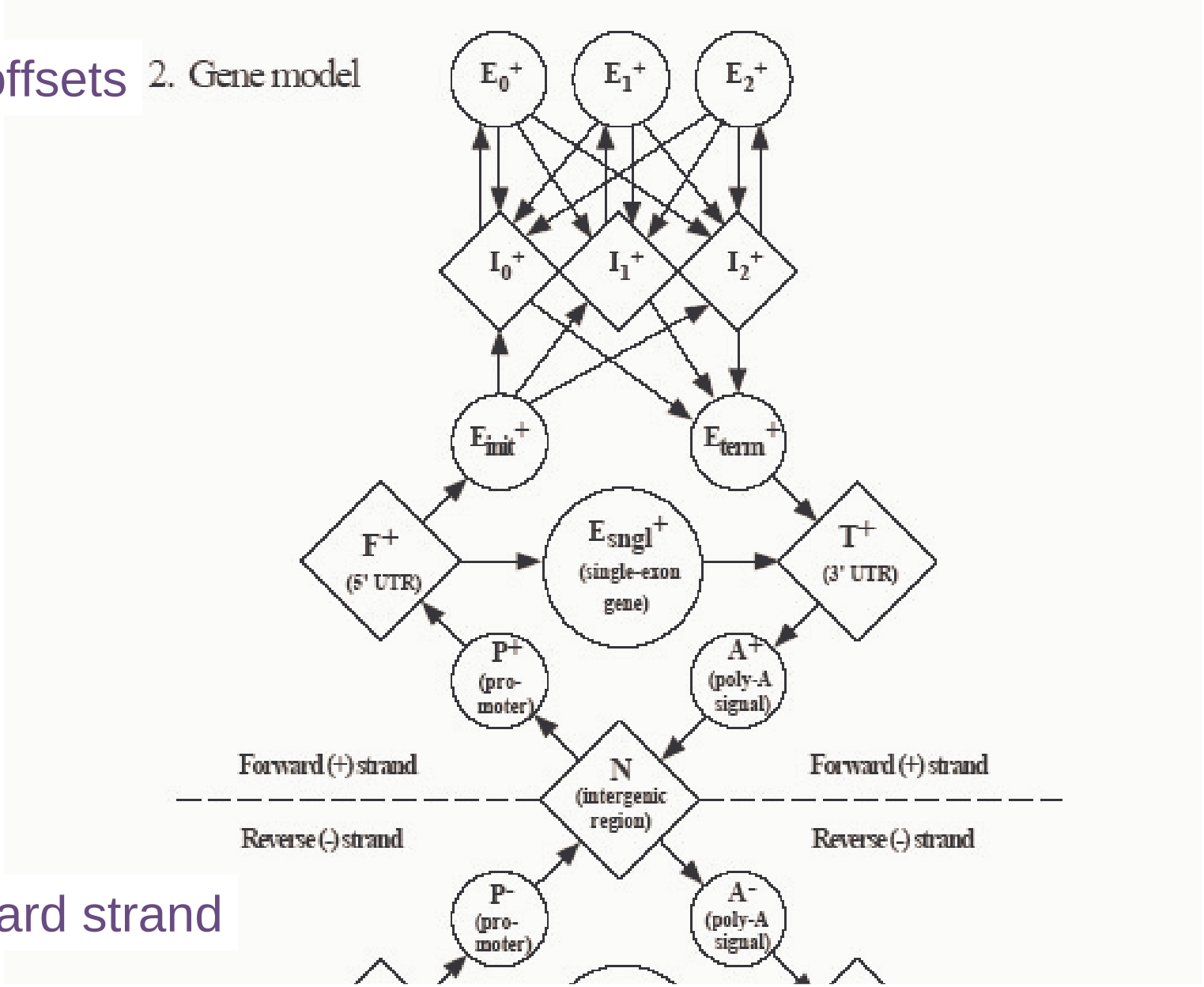
states emit *strings* of symbols
+ length distribution

parse of observation
assigns subsequences to states
Viterby like
time consuming, hard to train

GenScan:
models for subsequences in genome
transitions biologically consistent
statistics depending on C+G content

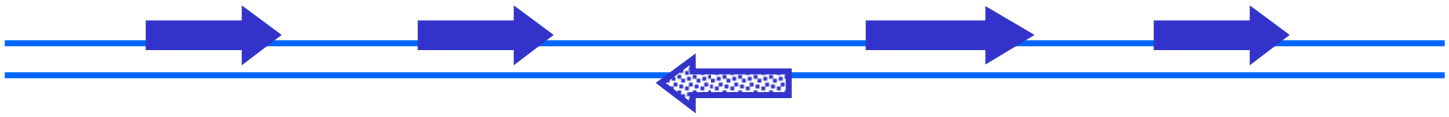
GenScan

phase offsets



backward strand

however …



Intron 22 of the FVIII gene is unusual; it is very large (32kb) and contains a CpG island. This CpG island acts as a bi-directional promoter for two genes within the FVIII gene, F8A and F8B. F8A is transcribed in the opposite direction to factor VIII, is intronless and completely nested within intron 22. The functions of F8A and F8B are unknown although the genes are expressed ubiquitously.

see http://www.ich.ucl.ac.uk/cmgs/fviii99.htm
(page moved!)