Sequence Analysis and Genomics

4 · Hidden Markov models I: annotating a sequence

Maria Waldl Elisa Israel

Bioinformatics, Leipzig University
maria.waldl@bioinf.uni-leipzig.de

02/11/2026 · WiSe 2026/27 · module 10-202-2207

Today

After today you can ...

- draw an HMM from a verbal description and write down its transitions and emissions
- find the most probable path by hand (Viterbi) and see it as a DP like alignment
- tell "best path" (Viterbi) from "all paths" (forward), and say when each is the question
- compute posterior state probabilities and explain why they can disagree with the Viterbi path
- estimate parameters from labelled data (with pseudocounts) and explain Baum-Welch as EM
- describe gene finding as decoding, and name what a plain HMM cannot model

Last time

Your turn 3 min

- In our example, star and progressive alignment made the same error. Why could neither repair it?
- How can sequences C and D change how T-Coffee aligns A and B?

An MSA column is position-specific: some positions are conserved, others vary. Next time, each column becomes a *state* of a model (profile HMMs). Today: the machinery, on a simpler question.

A stretch of genome

```
  1 AAAGCCACTC TAGTAAGGCG ATAGTCCGGG AATCTACGTG CATTAATCCA GTGAACAGAA TACTAGGTAT TGGAACGCTT
 81 TCTCAGTATG CCTCGCGCGT CGCACAAGCG CCCCGTAACG CCTCAAGATA ATCGACCCGG CTGTTGCACC TCCCTGCTAT
161 AAGCTACCAA CGAAATGTCC GTCTCACCTG TGTCAACCTA ATTCAATTTC ACCAACACCC ATGCCACGGG CGCCCGGCCG
241 GCCCGCCCGA CCGGGTGGGA CGACACGCGG CCCATCCGCC CGACGGGCCG TATCCTAAAT CTGCATATCT GACAATTTTG
321 TATTGAATGT AACCAACCAG CCCCCCCCCC GGCGCTATAA GGGTACGGAT AGATTTGCCA TCTAACTTAC AGGGGGGTAG
```

Your turn 2 min

Where are the GC-rich regions? How sure are you about the borders?

Why care: CpG islands

- CpG: a `C` followed by a `G` on the same strand
- The C of a CpG is often methylated; methylated C tends to mutate to T
- So CpG is rare in most of the genome (about a quarter of the expected frequency)
- Except in *CpG islands*: a few hundred to a few thousand bp, often at promoters, where methylation is suppressed

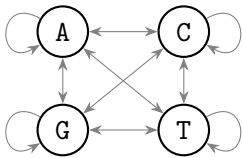
A common rule of thumb: ≥ 200 bp, GC content $> 50\,\%$, observed/expected CpG > 0.6.

Question: where are the CpG islands in a new genome?

Markov chains

Definition: Markov chain

The next letter depends only on the current one: $P(x) = P(x_1) \prod_{i=2}^{L} T(x_{i-1}, x_i)$, $T(a, b)$: probability that b follows a; each row of T sums to 1.



Simplest case (*order 0*): letters are independent, $P(x) = \prod_i P(x_i)$.

Your turn 1 min

Where have you seen "the next step depends only on the current state" in this course already?

Two models, one log-odds

What does the score mean? 4 / 5

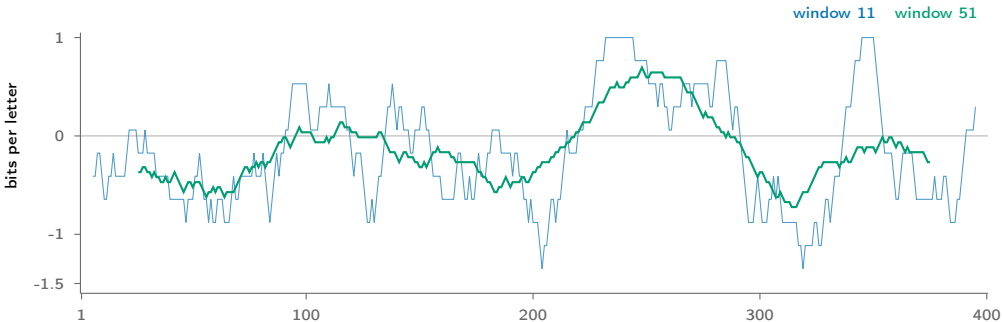
$$S(x) = \log_2 \frac{P(x \mid \text{GC-rich})}{P(x \mid \text{background})} = \sum_i \log_2 \frac{P(x_i \mid \text{GC-rich})}{P(x_i \mid \text{background})}$$: positive = evidence for GC-rich

b	A	C	G	T
$P(b \mid \text{GC-rich})$	0.1	0.4	0.4	0.1
$P(b \mid \text{background})$	0.3	0.2	0.2	0.3
$\log_2$ ratio (bits)	-1.58	$+1.00$	$+1.00$	-1.58

Your turn 2 min

Score the window `AGCCCCCC`. What is the null model here, and is it realistic?

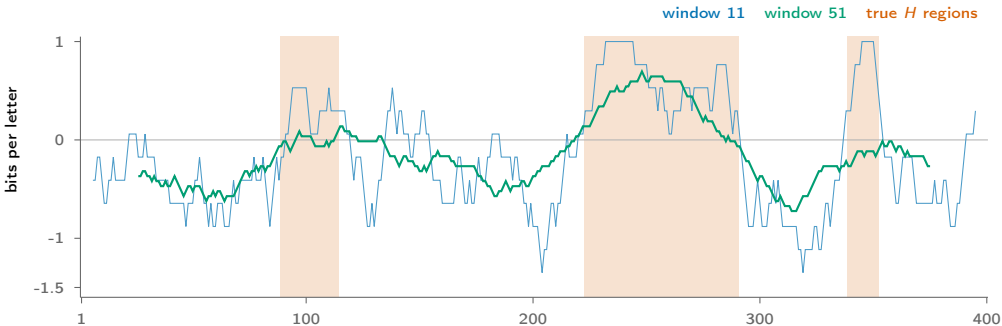
Sliding windows, and where they fail



- small window: noisy; large window: blurs borders and misses short regions
- the window size is a guess about region length

  Better: let the model decide where a region starts and ends → hidden states.

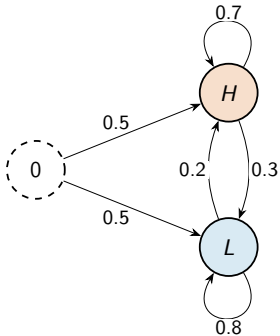
Sliding windows, and where they fail



- small window: noisy; large window: blurs borders and misses short regions
- the window size is a guess about region length

  Better: let the model decide where a region starts and ends → hidden states.

From two models to one



$E(H, b)$, GC-rich

b	A	C	G	T
	0.1	0.4	0.4	0.1

$E(L, b)$, background

b	A	C	G	T
	0.3	0.2	0.2	0.3

- emission tables: the two models of the last slides; transitions: when to switch
- self-loop $p \Rightarrow$ a state lasts $1/(1-p)$ letters on average: here 3.3 (H), 5 (L)
- genome stretch: same emissions, $T(H, H) = 0.98$, $T(L, L) = 0.99$ (regions of ≈ 50 and 100 nt)

A deliberate simplification: this model only sees GC content (see "What does this model actually detect?").

The parts of an HMM

Definition: Hidden Markov model

- states (H, L) and a silent begin state 0
- transitions $T(k, l) = P(\pi_i = l \mid \pi_{i-1} = k)$
- emissions $E(k, b) = P(x_i = b \mid \pi_i = k)$
- the path $\pi = \pi_1 \dots \pi_L$ is *hidden*, the sequence $x = x_1 \dots x_L$ is *observed*

- each row of T sums to 1; each row of E sums to 1
- silent states emit nothing: begin and end here, delete states next time
- today: no end state, the path simply stops after x_L

The model writes sequences

path π	L	L	L	L	L	L	H	L	L	L	L	L	H	H	H	H	H	H	L	L
sequence x	G	T	G	T	A	T	C	C	C	C	G	C	G	A	G	G	C	G	A	A

To generate: start in 0, pick a state with $T(0, \cdot)$, emit a letter with E, move on with T, repeat.

The model writes sequences

path π	L	L	L	L	L	L	H	L	L	L	L	L	H	H	H	H	H	H	L	L
sequence x	G	T	G	T	A	T	C	C	C	C	G	C	G	A	G	G	C	G	A	A

To generate: start in 0, pick a state with $T(0, \cdot)$, emit a letter with E, move on with T, repeat.

We see only the bottom line. The top line is what we want back.

One path, one probability

$$P(x, \pi) = T(0, \pi_1)\, E(\pi_1, x_1) \prod_{i=2}^{L} T(\pi_{i-1}, \pi_i)\, E(\pi_i, x_i)$$

Your turn 3 min

Compute $P(x, \pi)$ for $x = \texttt{TGGG}$ and $\pi = \text{HHHH}$. Then for $\pi = \text{LHHH}$. Which is more probable?

One path, one probability

$$P(x, \pi) = T(0, \pi_1)\, E(\pi_1, x_1) \prod_{i=2}^{L} T(\pi_{i-1}, \pi_i)\, E(\pi_i, x_i)$$

Your turn 3 min

Compute $P(x, \pi)$ for $x = \texttt{TGGG}$ and $\pi = \text{HHHH}$. Then for $\pi = \text{LHHH}$. Which is more probable?

$P(x, \text{HHHH}) = 0.5 \cdot 0.1 \cdot (0.7 \cdot 0.4)^3 = 1.1{\cdot}10^{-3}$
$P(x, \text{LHHH}) = 0.5 \cdot 0.3 \cdot 0.2 \cdot 0.4 \cdot (0.7 \cdot 0.4)^2 = 9.41{\cdot}10^{-4}$

Four questions

question	**algorithm**
Which path explains x best?	Viterbi (decoding)
How probable is x under the model?	forward
Which state is position i in, how sure are we?	posterior decoding
Where do the numbers come from?	counting; Baum-Welch

What does this model actually detect?

⚠ Careful: Our simplification: GC-rich is not CpG

Our emissions see single letters: the model finds GC-rich regions. `GGGCCC` and `CGCGCG` have the same probability under it.

To see CpG, the emission must depend on the previous letter: e.g. 8 states $A^+, C^+, G^+, T^+, A^-, \ldots$, one per letter and region (backup).

How many paths?

Your turn 1 min

K states, sequence of length L: how many paths? Our toy model, $L = 1000$?

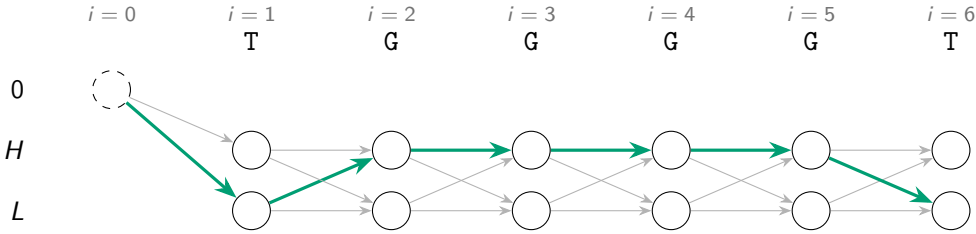
How many paths?

Your turn 1 min

K states, sequence of length L: how many paths? Our toy model, $L = 1000$?

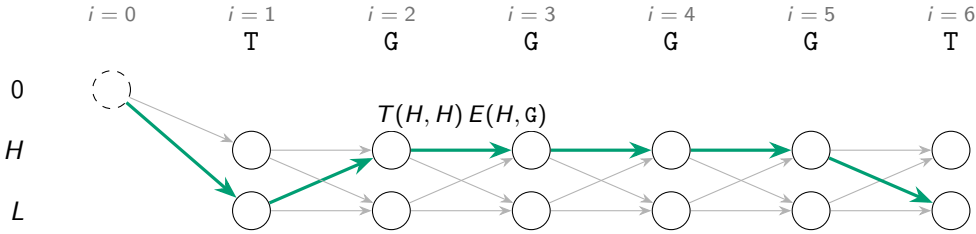
K^L paths; $2^{1000} \approx 10^{301}$. Brute force is out, as for alignments in slot 1.

Paths through a graph



- a node per (state, position); an edge from every state at $i-1$ to every state at i
- edge weight $T(k,l)\,E(l,x_i)$; a path's probability is the product of its edge weights
- decoding = best path through a layered graph, like the alignment grid

Paths through a graph



- a node per (state, position); an edge from every state at $i - 1$ to every state at i
- edge weight $T(k, l)\, E(l, x_i)$; a path's probability is the product of its edge weights
- decoding = best path through a layered graph, like the alignment grid

The recursion

Your turn 5 min

The best path that ends in state l at position i: what must its last step look like?

The recursion

Your turn 5 min

The best path that ends in state l at position i: what must its last step look like?

Definition: Viterbi recursion

$V(l, i)$: probability of the best path for $x_1 \dots x_i$ that ends in state l

$$V(l, i) = E(l, x_i) \max_k V(k, i-1)\, T(k, l)$$

Where to start, where to stop

- initialisation: $V(0,0) = 1, \quad V(k,0) = 0$ for $k \neq 0$
  so the first column is $V(I,1) = E(I,x_1)\, T(0,I)$
- recursion for $i = 1, \ldots, L$; store the winning k (pointer)
- termination: $P(x,\pi^*) = \max_k V(k,L)$ (no end state)
- traceback from the best last cell along the pointers

Filling the table

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0						
L	0						

Your turn 8 min

$x =$ TGGGGT: fill the rest of the table with your neighbour.

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Filling the table

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05					
L	0						

$V(H,1) = E(H,\mathtt{T})\, T(0,H) = 0.1 \cdot 0.5 = 0.05$

Your turn 8 min

$x =$ TGGGGT: fill the rest of the table with your neighbour.

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Filling the table

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05					
L	0	0.15					

$V(L,1) = E(L,\mathtt{T})\,T(0,L) = 0.3 \cdot 0.5 = 0.15$

Your turn 8 min

$x = \mathtt{TGGGGT}$: fill the rest of the table with your neighbour.

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Filling the table

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.014				
L	0	0.15					

$$V(H,2) = E(H,\texttt{G}) \max\{V(H,1)\,T(H,H),\ V(L,1)\,T(L,H)\} = 0.4 \cdot \max\{0.05 \cdot 0.7,\ 0.15 \cdot 0.2\} = 0.014$$

Your turn 8 min

$x = \texttt{TGGGGT}$: fill the rest of the table with your neighbour.

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Filling the table

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.014				
L	0	0.15	0.024				

$V(L,2) = E(L,\texttt{G}) \max\{V(H,1)\,T(H,L),\ V(L,1)\,T(L,L)\} = 0.2 \cdot \max\{0.05 \cdot 0.3,\ 0.15 \cdot 0.8\} = 0.024$

Your turn 8 min

$x =$ TGGGGT: fill the rest of the table with your neighbour.

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Filling the table

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05 ←	0.014	?			
L	0	0.15 ←	0.024	?			

Your turn 8 min

$x =$ TGGGGT: fill the rest of the table with your neighbour.

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Filling the table

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.014	$3.92{\cdot}10^{-3}$			
L	0	0.15	0.024	$3.84{\cdot}10^{-3}$			

Your turn 8 min

$x =$ `TGGGGT`: fill the rest of the table with your neighbour.

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Filling the table

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.014	$3.92 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$		
L	0	0.15	0.024	$3.84 \cdot 10^{-3}$	$6.14 \cdot 10^{-4}$		

Your turn 8 min

$x =$ TGGGGT: fill the rest of the table with your neighbour.

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Filling the table

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.014	$3.92 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$	$3.07 \cdot 10^{-4}$	
L	0	0.15	0.024	$3.84 \cdot 10^{-3}$	$6.14 \cdot 10^{-4}$	$9.83 \cdot 10^{-5}$	

Your turn 8 min

$x =$ `TGGGGT`: fill the rest of the table with your neighbour.

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Filling the table

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.014	$3.92 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$	$3.07 \cdot 10^{-4}$	$2.15 \cdot 10^{-5}$
L	0	0.15	0.024	$3.84 \cdot 10^{-3}$	$6.14 \cdot 10^{-4}$	$9.83 \cdot 10^{-5}$	$2.77 \cdot 10^{-5}$

Your turn 8 min

$x =$ `TGGGGT`: fill the rest of the table with your neighbour.

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Filling the table

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.014	$3.92\cdot10^{-3}$	$1.1\cdot10^{-3}$	$3.07\cdot10^{-4}$	$2.15\cdot10^{-5}$
L	0	0.15	0.024	$3.84\cdot10^{-3}$	$6.14\cdot10^{-4}$	$9.83\cdot10^{-5}$	$2.77\cdot10^{-5}$

$P(x, \pi^*) = \max_k V(k, 6) = 2.77\cdot10^{-5}$

Your turn 8 min

$x =$ `TGGGGT`: fill the rest of the table with your neighbour.

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Traceback

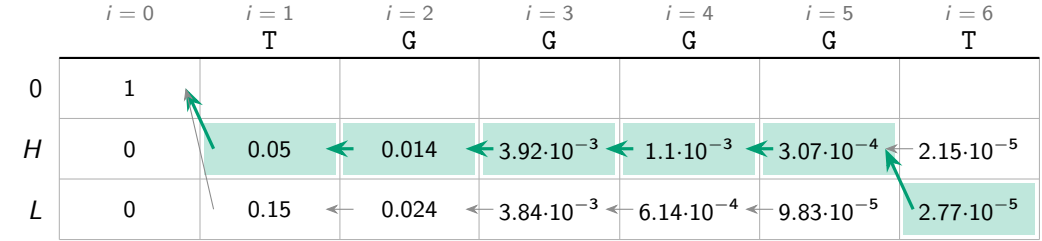
	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.014	$3.92 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$	$3.07 \cdot 10^{-4}$	$2.15 \cdot 10^{-5}$
L	0	0.15	0.024	$3.84 \cdot 10^{-3}$	$6.14 \cdot 10^{-4}$	$9.83 \cdot 10^{-5}$	$2.77 \cdot 10^{-5}$

Pointers store the winning predecessor. The best path, HHHHHL, starts like HHHH from "One path, one probability".

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Traceback



$P(x, \pi^*) = \max_k V(k, 6) = 2.77 \cdot 10^{-5}, \quad \pi^* = \mathrm{HHHHHL}$

Pointers store the winning predecessor. The best path, HHHHHL, starts like HHHH from "One path, one probability".

$T(k, l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k, b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Products become very small

- $0.25^{1000} \approx 10^{-602}$; the smallest positive double is $\approx 10^{-308}$ (with subnormals 10^{-324}): underflow to 0
- log space: products $\rightarrow$ sums; max is unchanged by log (monotone)

$$\log V(l, i) = \log E(l, x_i) + \max_k \left[\log V(k, i-1) + \log T(k, l) \right]$$

Careful

These log scores are log probabilities: the result is still $\log P(x, \pi^*)$. Other additive scores also run through Viterbi, but then the result is a best score, not a probability (section 10).

Viterbi and Needleman-Wunsch

	Needleman-Wunsch	**Viterbi**
cell	$F(i, j)$: best alignment of two prefixes	$V(k, i)$: best path for $x_1..x_i$ ending in k
predecessors	3 neighbouring cells	all K states of column $i-1$
combine	max of sums	max of products (sums in log space)
result	$F(n, m)$	$\max_k V(k, L)$
traceback	pointers	pointers
cost	$O(nm)$	$O(LK^2)$

A different question

How probable is x under the model, summed over all paths?

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

What does the score mean? 4 / 5

$\log \frac{P(x \mid \text{HMM})}{P(x \mid \text{background})}$: does x look like the model at all? (HMMER, next time)

Max becomes sum

Definition: Forward recursion

$\alpha(l, i)$: probability of $x_1 \ldots x_i$, summed over all paths that end in state l

$$\alpha(l, i) = E(l, x_i) \sum_k \alpha(k, i-1)\, T(k, l), \qquad P(x) = \sum_k \alpha(k, L)$$

Same graph, same table, same cost $O(LK^2)$; no traceback.

Your turn 1 min

PAM: $(P^2)_{AA} = \sum_c P_{Ac} P_{cA}$. Where is the forward algorithm in this?

Forward on our example

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0						
L	0						

Your turn 5 min

Fill the forward table. Compare $P(x)$ with $P(x, \pi^*) = 2.77{\cdot}10^{-5}$ from Viterbi. What fraction of the probability sits on the best path?

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Forward on our example

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05					
L	0						

$\alpha(H,1) = E(H,\mathtt{T})\, T(0,H) = 0.1 \cdot 0.5 = 0.05$

Your turn 5 min

Fill the forward table. Compare $P(x)$ with $P(x,\pi^*) = 2.77{\cdot}10^{-5}$ from Viterbi. What fraction of the probability sits on the best path?

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Forward on our example

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05					
L	0	0.15					

$\alpha(L,1) = E(L,\mathtt{T})\, T(0,L) = 0.3 \cdot 0.5 = 0.15$

Your turn 5 min

Fill the forward table. Compare $P(x)$ with $P(x,\pi^*) = 2.77{\cdot}10^{-5}$ from Viterbi. What fraction of the probability sits on the best path?

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Forward on our example

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.026				
L	0	0.15					

$$\alpha(H,2) = E(H,\texttt{G})\,(\alpha(H,1)\,T(H,H) + \alpha(L,1)\,T(L,H)) = 0.4 \cdot (0.05 \cdot 0.7 + 0.15 \cdot 0.2) = 0.026$$

Your turn 5 min

Fill the forward table. Compare $P(x)$ with $P(x,\pi^*) = 2.77{\cdot}10^{-5}$ from Viterbi. What fraction of the probability sits on the best path?

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Forward on our example

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.026				
L	0	0.15	0.027				

$$\alpha(L,2) = E(L,\texttt{G})\,(\alpha(H,1)\,T(H,L) + \alpha(L,1)\,T(L,L)) = 0.2 \cdot (0.05 \cdot 0.3 + 0.15 \cdot 0.8) = 0.027$$

Your turn 5 min

Fill the forward table. Compare $P(x)$ with $P(x,\pi^*) = 2.77{\cdot}10^{-5}$ from Viterbi. What fraction of the probability sits on the best path?

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Forward on our example

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.026	?			
L	0	0.15	0.027	?			

Your turn 5 min

Fill the forward table. Compare $P(x)$ with $P(x, \pi^*) = 2.77{\cdot}10^{-5}$ from Viterbi. What fraction of the probability sits on the best path?

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Forward on our example

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.026	$9.44{\cdot}10^{-3}$			
L	0	0.15	0.027	$5.88{\cdot}10^{-3}$			

Your turn 5 min

Fill the forward table. Compare $P(x)$ with $P(x, \pi^*) = 2.77{\cdot}10^{-5}$ from Viterbi. What fraction of the probability sits on the best path?

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Forward on our example

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.026	$9.44{\cdot}10^{-3}$	$3.11{\cdot}10^{-3}$		
L	0	0.15	0.027	$5.88{\cdot}10^{-3}$	$1.51{\cdot}10^{-3}$		

Your turn 5 min

Fill the forward table. Compare $P(x)$ with $P(x, \pi^*) = 2.77{\cdot}10^{-5}$ from Viterbi. What fraction of the probability sits on the best path?

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Forward on our example

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.026	$9.44{\cdot}10^{-3}$	$3.11{\cdot}10^{-3}$	$9.92{\cdot}10^{-4}$	
L	0	0.15	0.027	$5.88{\cdot}10^{-3}$	$1.51{\cdot}10^{-3}$	$4.28{\cdot}10^{-4}$	

Your turn 5 min

Fill the forward table. Compare $P(x)$ with $P(x, \pi^*) = 2.77{\cdot}10^{-5}$ from Viterbi. What fraction of the probability sits on the best path?

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Forward on our example

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.026	$9.44{\cdot}10^{-3}$	$3.11{\cdot}10^{-3}$	$9.92{\cdot}10^{-4}$	$7.8{\cdot}10^{-5}$
L	0	0.15	0.027	$5.88{\cdot}10^{-3}$	$1.51{\cdot}10^{-3}$	$4.28{\cdot}10^{-4}$	$1.92{\cdot}10^{-4}$

Your turn 5 min

Fill the forward table. Compare $P(x)$ with $P(x, \pi^*) = 2.77{\cdot}10^{-5}$ from Viterbi. What fraction of the probability sits on the best path?

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Forward on our example

	$i=0$	$i=1$ T	$i=2$ G	$i=3$ G	$i=4$ G	$i=5$ G	$i=6$ T
0	1						
H	0	0.05	0.026	$9.44{\cdot}10^{-3}$	$3.11{\cdot}10^{-3}$	$9.92{\cdot}10^{-4}$	$7.8{\cdot}10^{-5}$
L	0	0.15	0.027	$5.88{\cdot}10^{-3}$	$1.51{\cdot}10^{-3}$	$4.28{\cdot}10^{-4}$	$1.92{\cdot}10^{-4}$

$$P(x) = \alpha(H,6) + \alpha(L,6) = 7.8{\cdot}10^{-5} + 1.92{\cdot}10^{-4} = 2.7{\cdot}10^{-4}$$

Your turn 5 min

Fill the forward table. Compare $P(x)$ with $P(x,\pi^*) = 2.77{\cdot}10^{-5}$ from Viterbi. What fraction of the probability sits on the best path?

$T(k,l)$	H	L
0	0.5	0.5
H	0.7	0.3
L	0.2	0.8

$E(k,b)$	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

Log space for sums

⚠ Careful: $\log(a+b) \neq \log a + \log b$

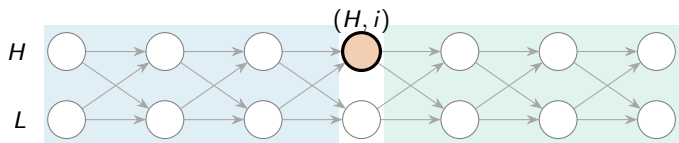
Forward needs sums of probabilities. In log space:

$$\log(a+b) = \log a + \log\left(1 + e^{\log b - \log a}\right) \quad (a \geq b)$$

("log-sum-exp"), or rescale each column to sum 1 and keep the factors (backup).

How sure are we about position i?

The question: $P(\pi_i = H \mid x)$, not "is i in H on the best path?"



$P(\pi_i = H, x) = \alpha(H, i)\,\beta(H, i)$: every path through (H, i) is a way in followed by a way out.

Backward

Definition: Backward recursion

$\beta(k, i)$: probability of $x_{i+1} \dots x_L$, given state k at position i

$$\beta(k, i) = \sum_l T(k, l)\, E(l, x_{i+1})\, \beta(l, i+1), \qquad \beta(k, L) = 1$$

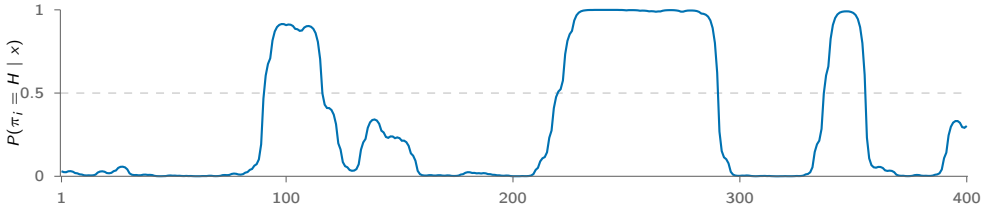
Definition: Posterior probability

$$P(\pi_i = k \mid x) = \frac{\alpha(k, i)\, \beta(k, i)}{P(x)}$$

Your turn 2 min

Why divide by $P(x)$? What do the posteriors in one column add up to?

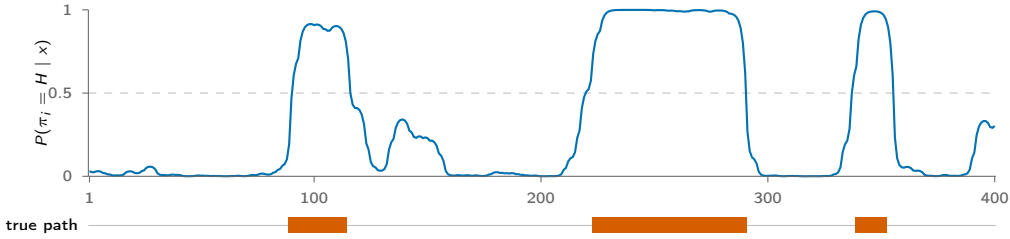
Posteriors along the sequence



Your turn 2 min

Compare with your guesses from the first slide. Where is the model unsure?

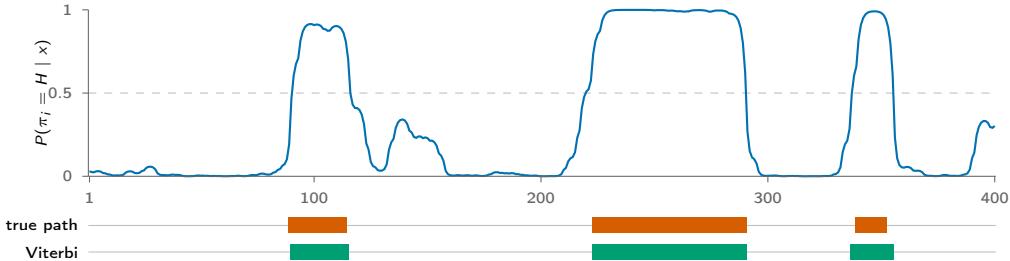
Posteriors along the sequence



Your turn 2 min

Compare with your guesses from the first slide. Where is the model unsure?

Posteriors along the sequence



Your turn 2 min

Compare with your guesses from the first slide. Where is the model unsure?

When Viterbi and posterior disagree

i	1	2	3	4	5	6
x_i	T	G	G	G	G	T
Viterbi path π_i^*	H	H	H	H	H	L
$P(\pi_i = H \mid x)$	0.38	0.62	0.71	0.70	0.59	0.29
most probable state	L	H	H	H	H	L

- Viterbi: the best *single path*, HHHHHL
- posterior: the best guess *per position*, LHHHHL
- position 1: most of the probability puts the T in L $(1 - 0.38)$, but the single best path does not pay for a switch there

The posterior-best states are not always a path

⚠ Careful

Picking the most probable state per position can use a transition with $T(k, l) = 0$: a "path" the model forbids.

Example: a gene model with codon positions $1 \to 2 \to 3 \to 1$. Two reading frames are almost equally likely, so position i is "most likely codon position 1" (frame a) and position $i + 1$ is also "most likely codon position 1" (frame b): the transition $1 \to 1$ does not exist.

Fix: if you need one consistent annotation, use Viterbi (or a path decoder that respects the zeros).

Which decoding when?

you need ...	**use**
one consistent annotation (gene structure)	Viterbi
per-position confidence, uncertain borders	posterior
is x from this model at all?	forward, log-odds against a null model

⚠ Careful: A confident answer on any input

Viterbi returns a path for a random sequence too. The posteriors and the log-odds against the null model show whether to believe it.

Where do the numbers come from?

path π	L	L	L	L	L	L	H	L	L	L	L	L	H	H	H	H	H	H	L	L
sequence x	G	T	G	T	A	T	C	C	C	C	G	C	G	A	G	G	C	G	A	A

With known states: count transitions A_{kl} and emissions $B_k(b)$, then normalise (maximum likelihood):

$$T(k,l) = \frac{A_{kl}}{\sum_{l'} A_{kl'}}, \qquad E(k,b) = \frac{B_k(b)}{\sum_{b'} B_k(b')}$$

Your turn 3 min

Estimate T and E from the labelled sequence above.

Where do the numbers come from?

path π	L	L	L	L	L	L	H	L	L	L	L	L	H	H	H	H	H	H	L	L
sequence x	G	T	G	T	A	T	C	C	C	C	G	C	G	A	G	G	C	G	A	A

With known states: count transitions A_{kl} and emissions $B_k(b)$, then normalise (maximum likelihood):

$$T(k,l) = \frac{A_{kl}}{\sum_{l'} A_{kl'}}, \qquad E(k,b) = \frac{B_k(b)}{\sum_{b'} B_k(b')}$$

A_{kl}	H	L	$B_k(b)$	A	C	G	T
H	5	2	H	1	2	4	0
L	2	10	L	3	4	3	3

$T(k,l)$	H	L	$E(k,b)$	A	C	G	T
H	5/7	2/7	H	1/7	2/7	4/7	0/7
L	2/12	10/12	L	3/13	4/13	3/13	3/13

Pseudocounts

- $E(H, \mathtt{T}) = 0/7$: under this model, no GC-rich region can ever contain a T
- a probability of 0 forbids forever: one T sends any region to L
- add a pseudocount r to every count (Laplace: $r = 1$):

$$E(H, \mathtt{T}) = \frac{0+1}{7+4} = \frac{1}{11}$$

- prior knowledge can set r per entry; zeros you *want* (forbidden transitions) stay 0

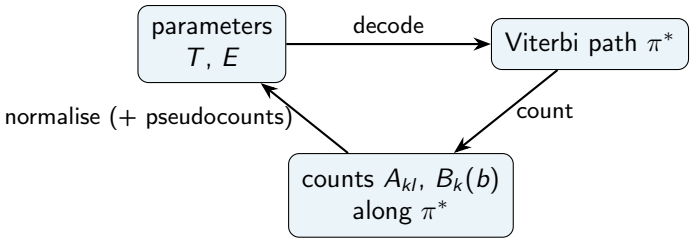
Chicken and egg

- to count, we need the paths
- to decode the paths, we need the parameters

Your turn 2 min

You have an unlabelled genome and a rough guess of the parameters. What would you do?

Viterbi training



- stop when the path no longer changes
- *hard* assignment: each position counts fully for one state

Baum-Welch: count with posteriors

Soft assignment: position i counts $P(\pi_i = k \mid x)$ for state k.

- expected emissions: $B_k(b) = \sum_{i:\, x_i = b} P(\pi_i = k \mid x)$
- expected transitions: $A_{kl} = \frac{1}{P(x)} \sum_i \alpha(k, i)\, T(k, l)\, E(l, x_{i+1})\, \beta(l, i+1)$

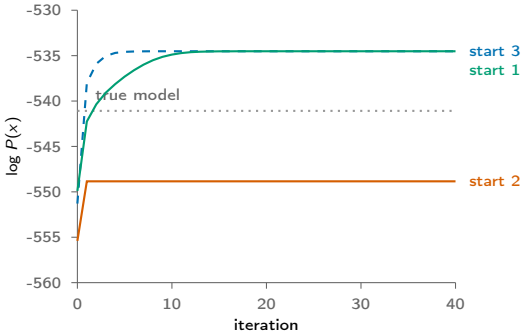
Then normalise as with known paths.

Expectation-maximisation

- **E-step:** expected counts under the current model (forward-backward)
- **M-step:** re-estimate as with known paths
- repeat; the likelihood $P(x)$ never decreases
- stops at a local optimum (or a saddle), not necessarily the global one

EM again later: reads that map to several transcripts in RNA-seq quantification (slot 6).

EM on the example



	$T(H,H)$	$T(L,L)$	$E(H,b)$ A	C	G	T
true	0.980	0.990	0.10	0.40	0.40	0.10
start 1	0.961	0.985	0.12	0.51	0.33	0.04
start 2	0.504	0.499	0.24	0.33	0.23	0.20
start 3	0.985	0.961	0.30	0.25	0.19	0.27

Baum-Welch on the 400-nt stretch, three starting guesses, begin probabilities fixed.

Your turn 2 min

Starts 1 and 3 reach the same log $P(x)$. Are they the same model? Why is start 2 stuck?

What EM does not tell you

⚠ Careful: A local optimum

Different starts $\rightarrow$ different models. Run several and compare $\log P(x)$.

⚠ Careful: States have no names

After training, "state 1" may be the GC-rich one or not (label switching). The topology and the number of states are your choice, not EM's.

A gene as a path

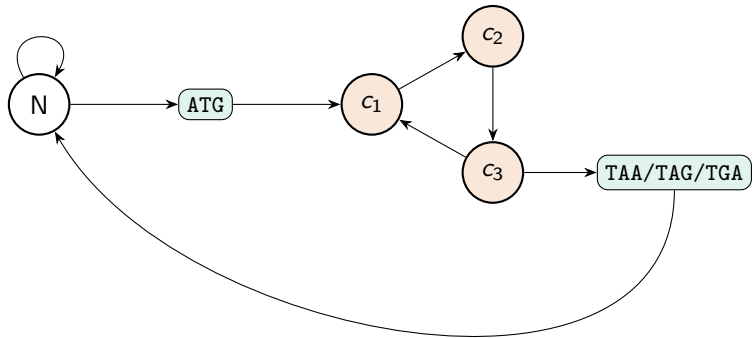
Your turn 3 min

Sketch an HMM that can only produce whole codons between a start and a stop.

A gene as a path

Your turn 3 min

Sketch an HMM that can only produce whole codons between a start and a stop.



Eukaryotic genes

- exons and introns; an intron can interrupt a codon: three intron *phases*
- signals at the borders: splice donor (`GT`), acceptor (`AG`), start, stop, promoter, poly-A
- genes on both strands: the model is mirrored

N	intergenic region
P, A	promoter, poly-A signal
F, T	5′ UTR, 3′ UTR
E_{init}, E_k, E_{term}	initial, internal (phase k), terminal exon
E_{sngl}	single-exon gene
I_k	intron of phase k
$+$, $-$	forward, reverse strand

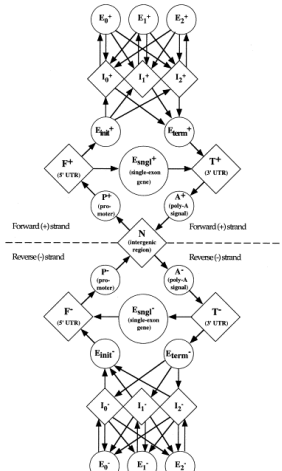


Figure: Burge & Karlin (1997), J. Mol. Biol. 268:78–94, doi:10.1006/jmbi.1997.0951; licence CC BY-NC-ND 4.0; unchanged.

How long does a state last?

- self-loop p: duration $P(d) = p^{d-1}(1-p)$, *geometric*; most likely length 1
- introns (a): roughly geometric, a self-loop is enough
- exons (b–d): clear peak, internal exons around 100–150 bp, hardly any very short ones
- fix: explicit length distributions per state (generalised / semi-Markov HMMs: GENSCAN, AUGUSTUS)

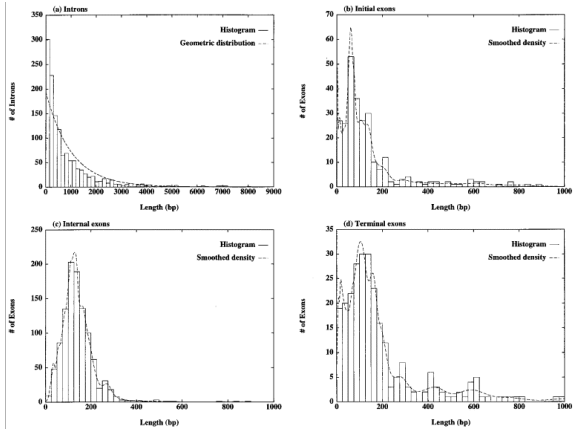


Figure: Burge & Karlin (1997), J. Mol. Biol. 268:78–94, doi:10.1006/jmbi.1997.0951; licence CC BY-NC-ND 4.0; unchanged.

Gene finders today

- prokaryotes: GeneMark (HMM), Prodigal (dynamic programming, no HMM)
- eukaryotes: AUGUSTUS (generalised HMM), BRAKER (AUGUSTUS + GeneMark trained with RNA-seq and protein evidence)
- deep learning: Helixer; Tiberius (neural network with an HMM layer)

The model is the biology

The algorithms are fixed. States, topology and zeros are the claim about biology.

- GC-rich regions, CpG islands (today)
- genes: exons, introns, signals (GENSCAN, AUGUSTUS)
- transmembrane helices in proteins (TMHMM)
- protein and domain families: profile HMMs, Pfam (next time)
- repeat families (Dfam)
- chromatin states from histone marks (ChromHMM)
- copy-number changes from read depth; local ancestry along a chromosome
- conserved elements in a genome alignment (phylo-HMMs, PhastCons)

Viterbi beyond biology

The pattern: a sequence of hidden decisions, local scores, an observed signal → best path through a layered graph.

- decoding noisy radio and phone signals (Viterbi 1967, convolutional codes)
- speech recognition: hidden words and sounds, observed audio
- part-of-speech tagging: hidden word classes, observed words
- GPS map matching: which road is the car on?
- nanopore basecalling: hidden bases, observed current

With probabilities: Viterbi, forward, posteriors. With other additive scores (costs, penalties): the same DP, a best score without a probability.

Your turn 2 min

Find a problem from your own field: what are the hidden states, what is observed, what would T and E be?

Assumptions are limits

- first order: the next state depends only on the current one
- emissions are independent given the state
- durations are geometric (previous section)
- no long-range dependencies: RNA base pairs couple distant positions

Beyond HMMs: stochastic context-free grammars, covariance models (Infernal, Rfam).

Your turn 2 min

Why can no HMM recognise "a sequence followed by its reverse complement" of any length?

HMMs and deep learning

- why HMMs are still used: interpretable states, probabilities and log-odds against a null model, little training data, exact algorithms
- hybrids: a neural network for the emissions, an HMM for the structure (Tiberius)

Take-home

1. An HMM: a hidden path, an observed sequence, transitions and emissions; it can generate sequences
2. Viterbi: the best path, a DP over a layered graph like alignment; work in log space
3. Forward: all paths, max $\rightarrow$ sum; log-odds against a null model
4. Posterior: per-position uncertainty; it can disagree with Viterbi
5. Training: counting with pseudocounts; Baum-Welch = EM with expected counts, local optima
6. The model's topology is the biology, and its assumptions are its limits

Could you answer these?

- Why is a 2-state single-letter model a GC-content detector and not a CpG detector?
- What changes between the Viterbi and the forward recursion, and what stays?
- Why can the most probable state at each position fail to form a valid path?
- Why do you need a pseudocount, and what happens without one?
- Two Baum-Welch runs give different models. Which do you trust?
- Why does a self-loop produce geometric lengths, and why does that matter for exons?

Next time

- exercise today: draw the HMM, Viterbi and forward by hand; optional: posteriors, coding in log space
- slot 5: HMMs for alignment: pair-HMMs (Gotoh again), profile HMMs, HMMER and Pfam
- optional primer: Eddy (2004), *What is a hidden Markov model?*; Rosalind BA10

Backup: a real CpG model

- 8 states A^+, C^+, G^+, T^+ (island) and A^-, C^-, G^-, T^- (outside); each state always emits its letter
- transitions within the $+$ block: the order-1 chain of islands; within the $-$ block: the chain of the rest; small probabilities between the blocks
- now $T(C^+, G^+)$ and $T(C^-, G^-)$ can differ: the model sees CpG

$+$	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	.274	.188
G	.161	.339	.375	.125
T	.079	.355	.384	.182

$-$	A	C	G	T
A	.300	.205	.285	.210
C	.322	.298	.078	.302
G	.248	.246	.298	.208
T	.177	.239	.292	.292

Row: current letter, column: next letter. Estimated from 48 human CpG islands; Durbin et al. (1998), *Biological Sequence Analysis*, Sect. 3.1.

Backup: notation in the books

	this course	**Durbin et al.**	**Compeau & Pevzner / BA10**
transition	$T(k, l)$	a_{kl}	$\text{transition}_{k,l}$
emission	$E(k, b)$	$e_k(b)$	$\text{emission}_k(b)$
Viterbi	$V(k, i)$	$v_k(i)$	$s_{k,i}$
forward, backward	$\alpha(k, i)$, $\beta(k, i)$	$f_k(i)$, $b_k(i)$	$\text{forward}_{k,i}$, $\text{backward}_{k,i}$
counts	A_{kl}, $B_k(b)$	A_{kl}, $E_k(b)$	$T_{l,k}$, $E_k(b)$

Backup: scaling instead of logs

- normalise each forward column: $\hat{\alpha}(\cdot, i) = \alpha'(\cdot, i)/c_i$ with $c_i = \sum_k \alpha'(k, i)$
- α': one recursion step from the normalised previous column
- $P(x) = \prod_i c_i$, so $\log P(x) = \sum_i \log c_i$
- backward uses the same c_i; the posteriors come out directly: $P(\pi_i = k \mid x) = \hat{\alpha}(k, i)\,\hat{\beta}(k, i)$