Hidden Markov models

Today:

- Using HMMs to model Variable length patterns and solving boundary detection problems
- Model design
- Parameter estimation from labeled data

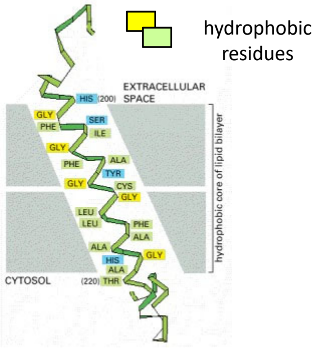
Problems with PSSMs

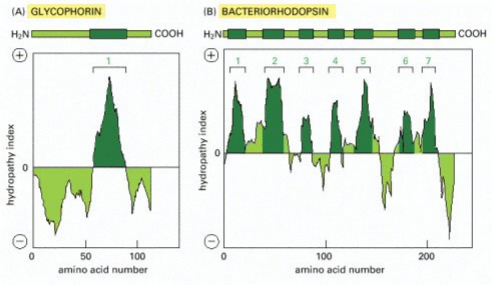
- Do not capture positional dependencies
- Hard to recognize pattern instances that contain indels
- Variable length motifs
- Do not handle boundary detection problems well

HMMs can model variable length patterns that are not position specific

Patterns characterized by changes in sequence composition,

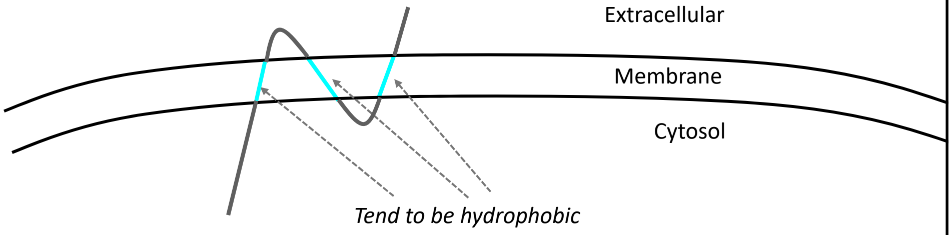
e.g. CpG islands, transmembrane domains





Molecular Biology of the Cell. 4th edition. Alberts B, Johnson A, Lewis J, *et al*.
New York: Garland Science; 2002. https://www.ncbi.nlm.nih.gov/books/NBK26878/

An example: transmembrane regions



Does a given sequence encode a transmembrane protein?

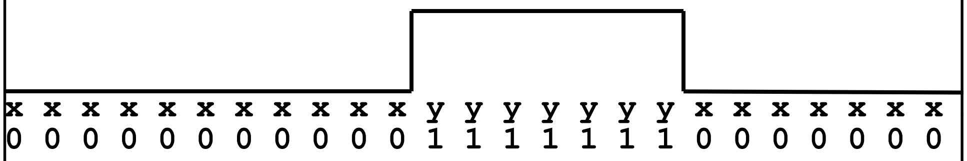
Boundary detection problem:

Find all transmembrane regions in a given sequence

Requires labeling each residue with its location in the cell

Boundary Detection

Goal: label every element in the sequence with a zero (not in membrane) or a one (in membrane)



```
x x x x x x x x x x x y y y y y y y x x x x x x x
0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0
```

HMMs

States: $E_1, E_2, \ldots E_N$

Initial state probabilities: $\pi(i)$

Transition probabilities: a_{ij}

Alphabet, Σ

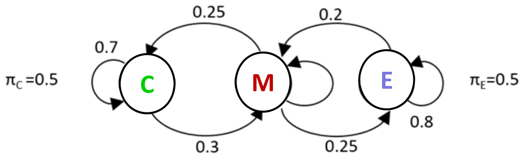
Emission probabilities: e_i

The parameters of the HMM $\lambda=(a_{ij}, e_i(\sigma), \pi)$ are "learned" from known examples ("labeled data").

An HMM is a *generative* model: we say

"the model emitted sequence $O = O_1 O_2 O_3 \ldots O_T$ via state path $Q = q_1 q_2 q_3 \ldots. q_T$ "

A three state transmembrane HMM:



$e_C(H)$	0.3
$e_C(L)$	0.7

$e_M(H)$	0.9
$e_M(L)$	0.1

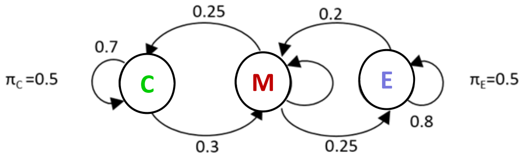
$e_E(H)$	0.2
$e_E(L)$	0.8

Emits amino acid sequences recoded in a two letter alphabet, Σ={H,L}

- H: hydrophobic residues
- L: hydrophilic residues

...**HHHLHLHLHLLLHHLHLHHHHHHHLHHHHHHHHHHHHLHLHLLHLHHLH**...

A three state transmembrane HMM:



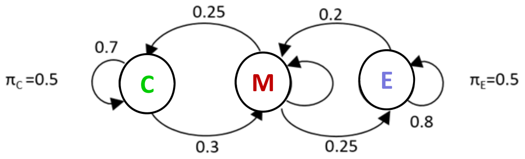
$e_C(H)$	0.3
$e_C(L)$	0.7

$e_M(H)$	0.9
$e_M(L)$	0.1

$e_E(H)$	0.2
$e_E(L)$	0.8

- A state can emit more than one symbol
- Each symbol can be emitted by more than one state
- In this model,
  - State: cellular location
  - Symbol: amino acid class (H or L)

A three state transmembrane HMM:



e_C(H)	0.3
e_C(L)	0.7

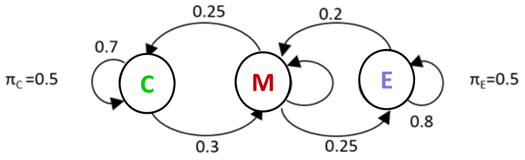
e_M(H)	0.9
e_M(L)	0.1

e_E(H)	0.2
e_E(L)	0.8

An HMM generates *labeled* sequences:

```
LLLHLHLLHLLLHHHLLHHHHLHHHLLHLLHLL...
CCCCCCCCCCCCMMMMMMMMMMMMMEEEEEEEE...
                       LLLHLHHHHHHLLHLLLLLHLHHHLLHLLLHLL...
                       CCCCCMMMMMMMEEEEEEEEMMMMMCCCCCCCC...
               LHLLLHLLHLHLHHHHHHLHLHLLHHLLHHHHHLHLLLLHLL...
               EEEEEEEEEEMMMMMMMMCCCCCCCCCCMMMMMMMEEEEEEE...
  LLLHLHLLHLHHHLLHHHHLHHHLLHLLHLLLLLLLLL...
  CCCCCCCCMMMMMMMMMMMMMMMMMMMMMEEEEEEEEE...
```

A three state transmembrane HMM:



e_C(H)	0.3
e_C(L)	0.7

e_M(H)	0.9
e_M(L)	0.1

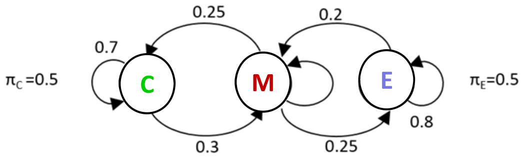
e_E(H)	0.2
e_E(L)	0.8

What is the probability that this model emitted LHHHL via path CMMME?

What is $P(O, Q|\lambda)$, where O = LHHHL and Q = CMMME?

$$P(O, Q|\lambda) = \pi_{q_1} \cdot e_{q_1}(O_1) \prod_{i=2}^{T} a_{q_{i-1}q_i} e_{q_i}(O_i)$$

A three state transmembrane HMM:



$e_C(H)$	0.3
$e_C(L)$	0.7

$e_M(H)$	0.9
$e_M(L)$	0.1

$e_E(H)$	0.2
$e_E(L)$	0.8

What is the probability that this model emitted LHHHL via path CMMME?

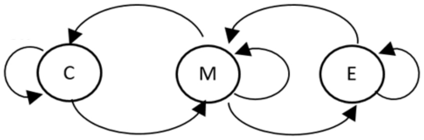
What is $P(O, Q|\lambda)$, where O = LHHHL and Q = CMMME?

$$\pi_C \cdot e_C(L) \cdot a_{CM} \cdot e_M\,(H) \cdot a_{MM} \cdot e_M\,(H) \cdot a_{MM} \cdot e_M\,(H) \cdot a_{ME} \cdot e_E\,(L) =$$
$$0.5 \cdot 0.7 \cdot 0.3 \cdot 0.9 \cdot 0.5 \cdot 0.9 \cdot 0.5 \cdot 0.9 \cdot 0.25 \cdot 0.8$$

Parameter estimation

- from labeled data
- from unlabaled data

Parameter estimation: transition probabilities



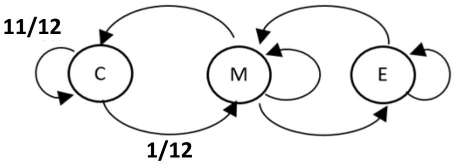
i	C	M	E
π_i			
e_i(H)			
e_i(L)			

LLLHLHLLHLLLHHHLLHHHHLHHHLLHLLHLL. . .
CCCCCCCCCCCCMMMMMMMMMMMMMEEEEEEEE. . .

$$a_{ij} = \frac{A_{ij}}{\sum_h A_{ih}}$$

A_{ij} = # of transitions from i to j in training data

Parameter estimation: transition probabilities



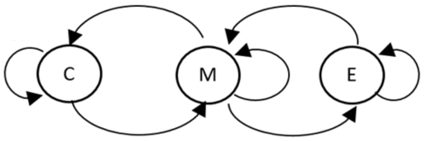
i	C	M	E
π_i			
e_i(H)			
e_i(L)			

LLLHLHLLHLLLHHHLLHHHHLHHHLLHLLHLL. . .
CCCCCCCCCCCCMMMMMMMMMMMMMEEEEEEEE. . .

$$a_{CC} = \frac{A_{CC}}{A_{CC}+A_{CM}} = \frac{11}{11+1}$$

11 CC pairs
1 CM pair

Parameter estimation: emission probabilities



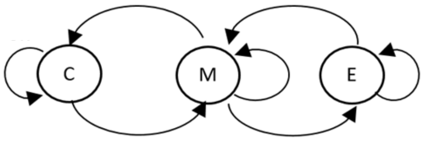
i	C	M	E
π_i			
e_i(H)			
e_i(L)			

LLLHLHLLHLLLHHHLLHHHHLHHHLLHLLHLL. . .
CCCCCCCCCCCCMMMMMMMMMMMMMEEEEEEEE. . .

$$e_i(H) = \frac{\mathcal{E}_i(\sigma)}{\sum_{\alpha\in\Sigma}\mathcal{E}_i(\alpha)}$$

$\mathcal{E}_i(\sigma)$ = # times that state E_i labels σ

Parameter estimation: emission probabilities



i	C	M	E
π_i			
e_i(H)			
e_i(L)			

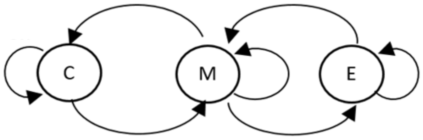
LLLHLHLLHLLLHHHLLHHHHLHHHLLHLLHLL. . .
CCCCCCCCCCCCMMMMMMMMMMMMMEEEEEEEE. . .

$$e_C(H) = \frac{\mathcal{E}_C(H)}{\sum_{\alpha\in\Sigma}\mathcal{E}_C(\alpha)}$$

$\mathcal{E}_C(H)$ = 3, $\mathcal{E}_C(L)$ = 9

$$e_C(H) = \frac{3}{3+9} = \frac{1}{4}$$

Parameter estimation: emission probabilities



i	C	M	E
π$_i$			
e$_i$(H)	0.25		
e$_i$(L)	0.75		

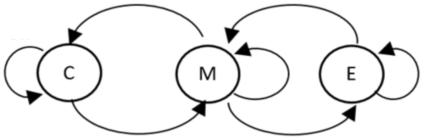
LLLHLHLLHLLLHHHLLHHHHLHHHLLHLLHLL. . .
CCCCCCCCCCCCMMMMMMMMMMMMMEEEEEEEE. . .

$$e_C(H) = \frac{\mathcal{E}_C(H)}{\sum_{\alpha\in\Sigma}\mathcal{E}_C(\alpha)}$$

$\mathcal{E}_C(H) = 3,\ \mathcal{E}_C(L) = 9$

$$e_C(H) = \frac{3}{3+9} = \frac{1}{4}$$

Parameter estimation: initial probabilities



i	C	M	E
π$_i$	3/4		1/4
e$_i$(H)			
e$_i$(L)			

π$_i$ = # of sequences that begin with E$_i$,
normalized by the total # of training sequences

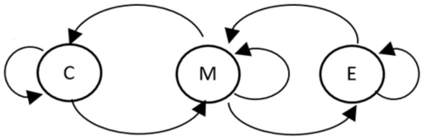
LLLHLHLLHLLLHHHLLHHHHLHHHLLHLLHLL. . .
CCCCCCCCCCCCMMMMMMMMMMMMMEEEEEEEEC. . .

LLLHLHHHHHHHLLHLLLLLLHLHHHLLHLLLHLL. . .
CCCCCMMMMMMMEEEEEEEEMMMMMCCCCCCCC. . .

LHLLLHLLHLHLHHHHHHLHLHLLHHLLHHHHHLHLLLLHLL. . .
EEEEEEEEEEMMMMMMMMCCCCCCCCCCMMMMMMMEEEEEEE. . .

LLLHLHLLHLHHHLLHHHHLHHHLLHLLHLLLLLLLLL. . .
CCCCCCCCMMMMMMMMMMMMMMMMMMMMMMEEEEEEEEE. . .

Parameter estimation: initial probabilities



i	C	M	E
π_i	3/4	0	1/4
$e_i(H)$			
$e_i(L)$			

$\pi_C = \frac{3}{4}$

$\pi_E = \frac{1}{4}$

$\pi_M = 0$

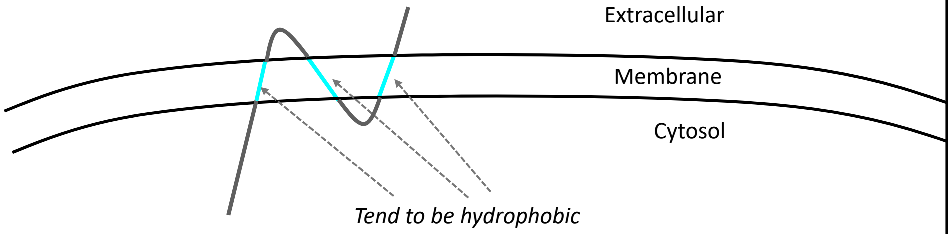
```
LLLHLHLLHLLLHHHLLHHHHLHHHLLHLLHLL...
CCCCCCCCCCCCMMMMMMMMMMMMMEEEEEEEEC...
                 LLLHLHHHHHHHLLHLLLLLHLHHHL
                 CCCCCMMMMMMMEEEEEEEEMMMMMC
           LHLLLHLLHLHLHHHHHHLHLHLLHHLLHHHHH
           EEEEEEEEEEMMMMMMMMCCCCCCCCCCMMMMM
  LLLHLHLLHLHHHLLHHHHLHHHLLHLLHLLLLLLLLL...
  CCCCCCCCMMMMMMMMMMMMMMMMMMMMMEEEEEEEEE...
```

Parameter estimation

- from labeled data
- from unlabeled data
  - learn the pattern and estimate the parameters simultaneously using an *expectation maximization* method called *Baum-Welch*

Next week

Recognition problems



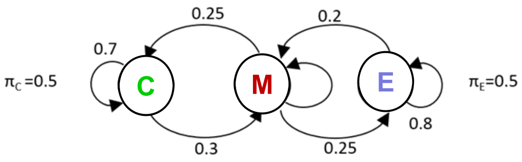
Does a given sequence, O, encode a transmembrane protein?

Boundary detection problem:

Find all transmembrane regions in a given sequence

Requires labeling each residue with its location in the cell

Is a given sequence, O, a transmembrane sequence?



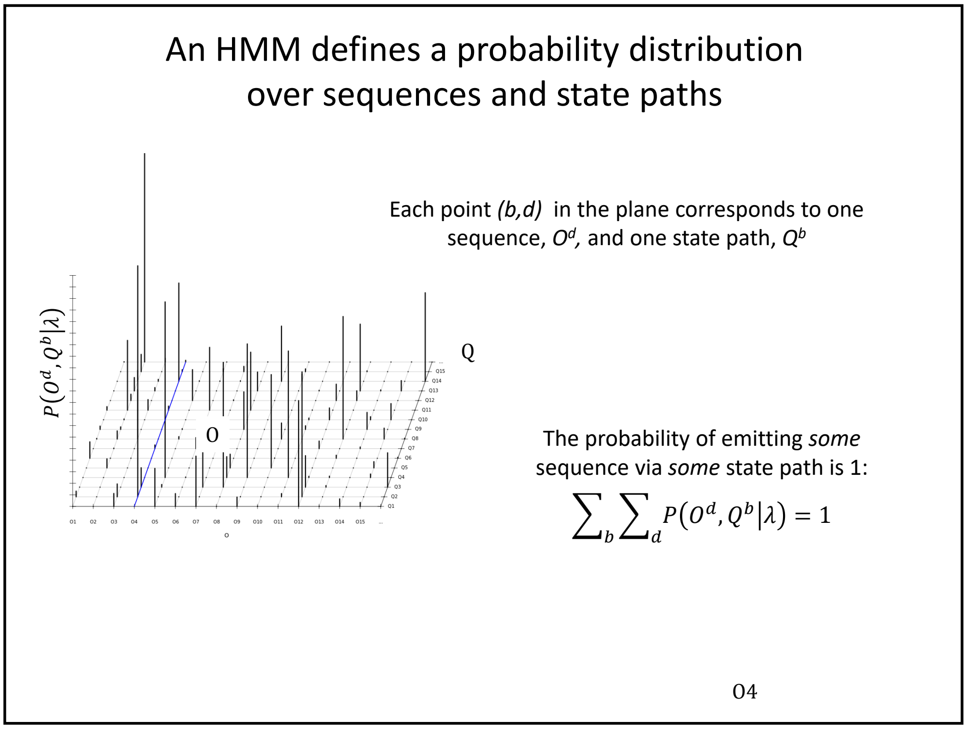
$e_C(H)$	0.3
$e_C(L)$	0.7

$e_M(H)$	0.9
$e_M(L)$	0.1

$e_E(H)$	0.2
$e_E(L)$	0.8

What is $P(O|\lambda_{TM})$, the probability that the TM model emitted O?

$$P(O|\lambda_{TM}) = \sum_{q} P\left(O, Q^b \middle| \lambda_{TM}\right)$$

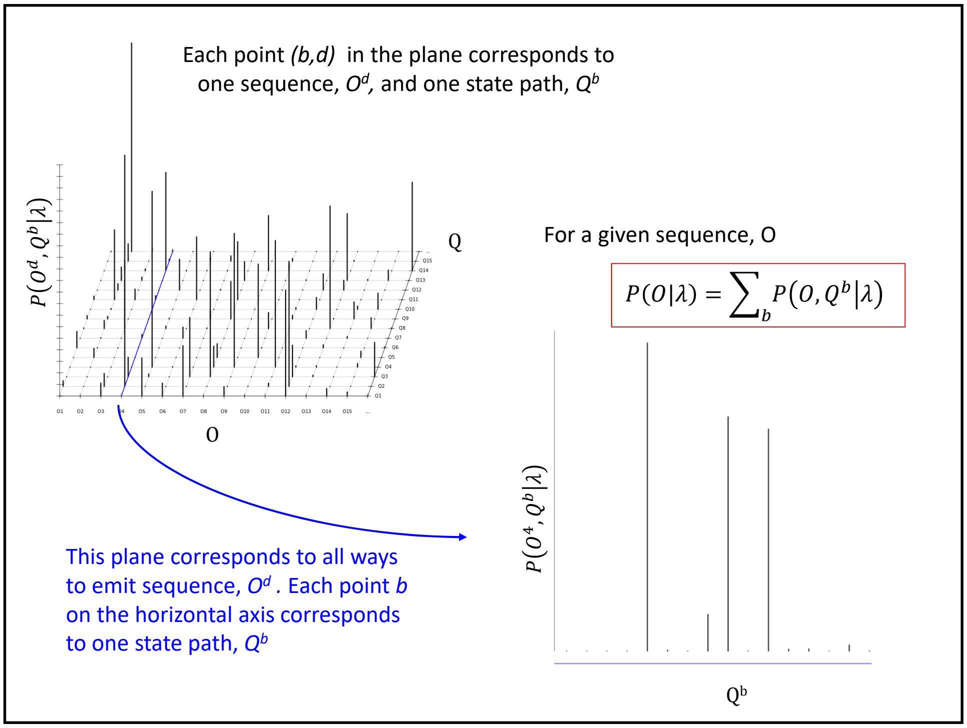


An HMM defines a probability distribution over sequences and state paths

Each point *(b,d)* in the plane corresponds to one sequence, *O*^{d}*,* and one state path, *Q*^{b}

The probability of emitting *some* sequence via *some* state path is 1:

$$\sum_b \sum_d P(O^d, Q^b|\lambda) = 1$$



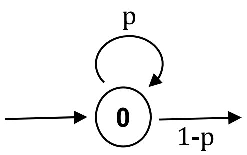
Each point *(b,d)* in the plane corresponds to one sequence, *O*^{d}*,* and one state path, *Q*^{b}

For a given sequence, O

$$\boxed{P(O|\lambda) = \sum_b P(O, Q^b|\lambda)}$$

This plane corresponds to all ways to emit sequence, *O*^{d} . Each point *b* on the horizontal axis corresponds to one state path, *Q*^{b}

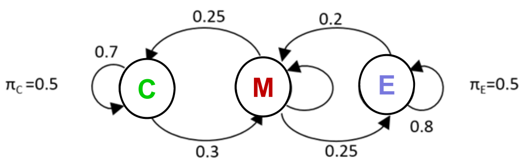
A null model



e_0(H)	0.25
e_0(L)	0.75

What is $P(O|\lambda_0)$, the probability that the null model emitted O?

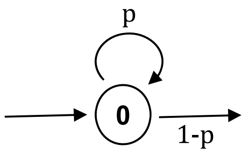
Is a given sequence, O, a transmembrane sequence?



e_C(H)	0.3
e_C(L)	0.7

e_M(H)	0.9
e_M(L)	0.1

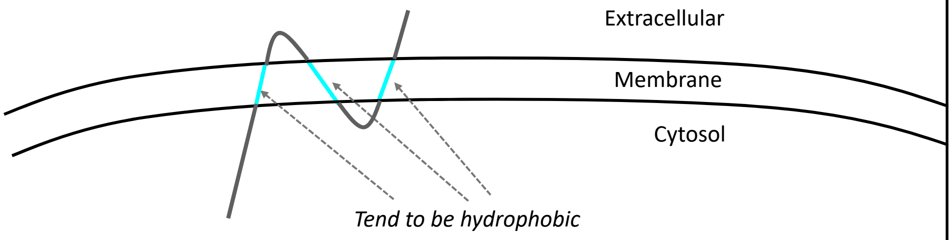
e_E(H)	0.2
e_E(L)	0.8



e_0(H)	0.25
e_0(L)	0.75

Is $\frac{P(O|\lambda_{TM})}{P(O|\lambda_0))}$ >>1?

Recognition problems

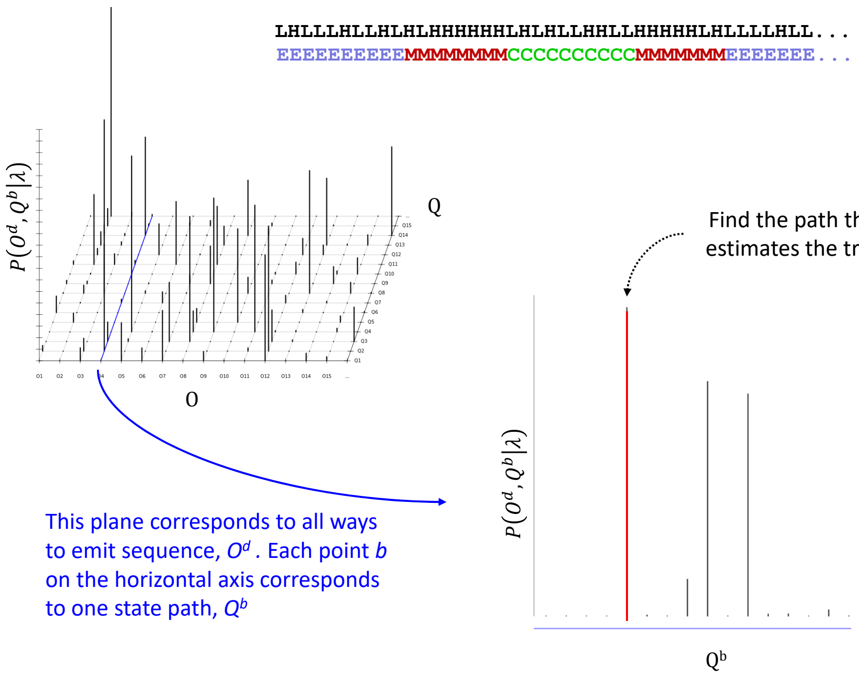


Boundary detection problem:

Find all transmembrane regions in a given sequence

Requires labeling each residue with its location in the cell

```
LHLLLHLLHLHLHHHHHHLHLHLLHHLLHHHHHLHLLLLHLL...
EEEEEEEEEEMMMMMMMMCCCCCCCCCCMMMMMMMEEEEEEE...
```



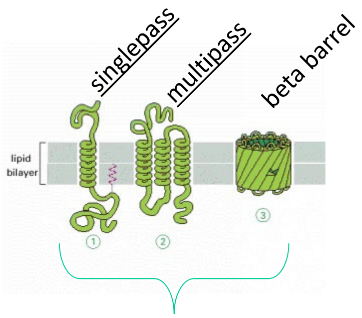
Find the path that best estimates the true path

This plane corresponds to all ways to emit sequence, O^d . Each point b on the horizontal axis corresponds to one state path, Q^b

HMM Design

- How many states?
- Which pairs of states have non-zero transitions?
- Alphabet
- Positional dependence
- Length distribution

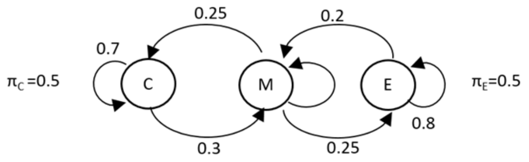
Various types of transmembrane proteins



Transmembrane proteins

Molecular Biology of the Cell. 4th edition.
Alberts B, Johnson A, Lewis J, et al.
New York: Garland Science; 2002.
https://www.ncbi.nlm.nih.gov/books/NBK26878/

HMM design



$e_C(H)$	0.3
$e_C(L)$	0.7

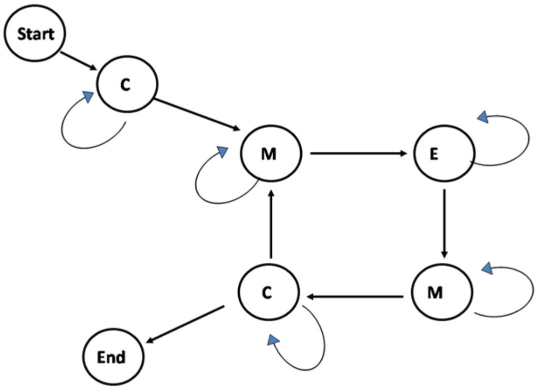
$e_M(H)$	0.9
$e_M(L)$	0.1

$e_E(H)$	0.2
$e_E(L)$	0.8

Can this model emit

1. Singlepass TM proteins that starts in the ECM? In the cytosol?
2. Multipass TM proteins that starts in the ECM? In the cytosol?
3. Extracellular proteins? Intracellular proteins?
4. Proteins that start or end in the membrane?

How would you modify the model topology so that it only emits one class of TM protein and no other sequences?



Note: the Start and End states are silent

This HMM models multipass sequences that start and end in the cytosol.

How would you modify the model topology so that it only emits

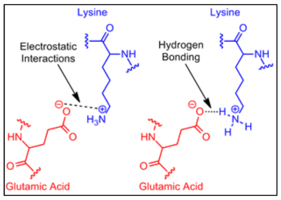
- Singlepass TM proteins that start in the ECM? In the cytosol?
- Multipass TM proteins that start in the ECM and end in the cytosol?
- Multipass TM proteins that start in the cytosol and end in the ECM?
- ...

HMM Design

- How many states?
- Which pairs of states have non-zero transitions?
- Alphabet
- Positional dependence
- Length distribution

More parameters: more precise, harder to estimate parameters

Positional dependence



Salt bridge

```
      12345678910
seq1  LIVKSMDGAL        +...-
seq2  STMECARLIT        -...+
seq3  LITDNSHQLI        -...+
seq4  LIMKVVDGYA        +...-
```

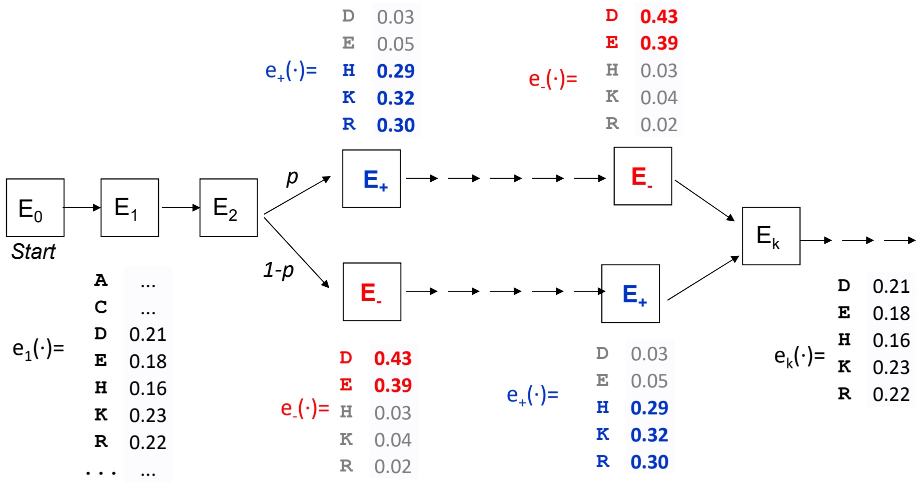
Residues that participate in a salt bridge

Afonnikov, Kolchanov, https://doi.org/10.1093/nar/gkh451

By Chem540f09grp6 https://commons.wikimedia.org/w/index.php?curid=8686974

Branching topologies can model positional dependencies

Emission frequencies shown only for amino acids that partipate in salt bridges



$e_1(\cdot)$=

A	…
C	…
D	0.21
E	0.18
H	0.16
K	0.23
R	0.22
. . .	…

$e_+(\cdot)$=

D	0.03
E	0.05
H	**0.29**
K	**0.32**
R	**0.30**

$e_-(\cdot)$=

D	**0.43**
E	**0.39**
H	0.03
K	0.04
R	0.02

$e_-(\cdot)$=

D	**0.43**
E	**0.39**
H	0.03
K	0.04
R	0.02

$e_+(\cdot)$=

D	0.03
E	0.05
H	**0.29**
K	**0.32**
R	**0.30**

$e_k(\cdot)$=

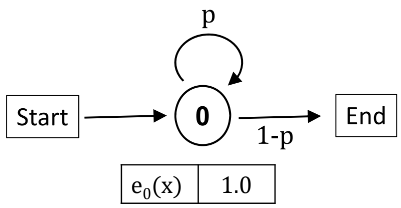
D	0.21
E	0.18
H	0.16
K	0.23
R	0.22

HMM Design

- How many states?
- Which pairs of states have non-zero transitions?
- Alphabet
- Positional dependence
- <u>Length distribution</u>

See also Durbin, 3.4

Topology implicitly defines the length distribution



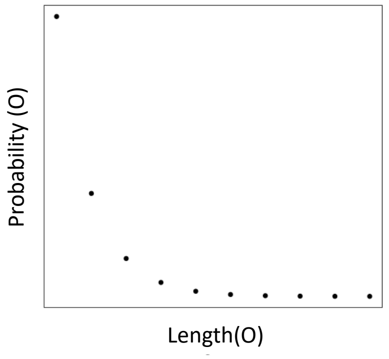
$p(x)=1\text{-}p$

$p(xx)=p\cdot(1\text{-}p)$

$p(xxx)=p^2\cdot(1\text{-}p)$

…

$p(\underbrace{x\ldots x}_{\text{length } l})=p^l\cdot(1\text{-}p)$



See Durbin, 3.4 for model topologies that avoid this decreasing exponential

Next: Recognition problems

- What is the probability of a given sequence, O?

  Forward algorithm

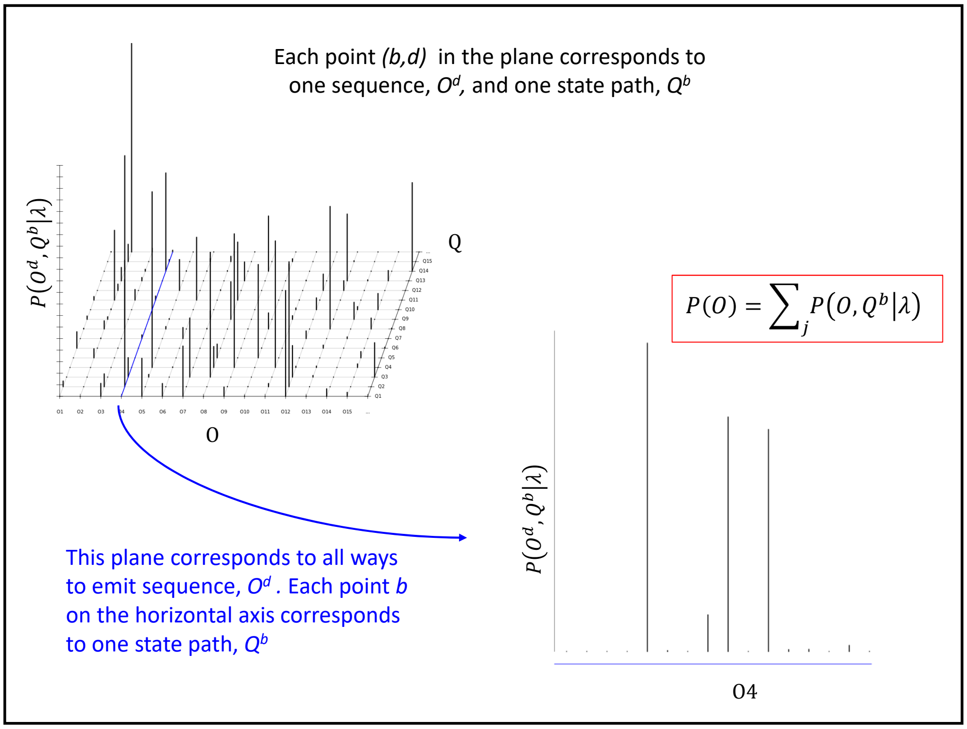
- Given a sequence O, what is the "true" sequence of states?

  Viterbi decoding: Viterbi algorithm

  Posterior decoding: Forward and Backward algorithms

- What state emitted the symbol O_t?

  Posterior decoding: Forward and Backward algorithms



Each point *(b,d)* in the plane corresponds to one sequence, O^d, and one state path, Q^b

This plane corresponds to all ways to emit sequence, O^d . Each point *b* on the horizontal axis corresponds to one state path, Q^b

$$\boxed{P(O) = \sum_j P(O, Q^b|\lambda)}$$