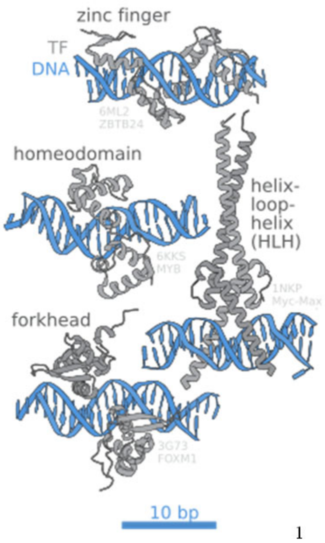
Conserved patterns in biological sequences
Probabilistic Framework

Types of patterns in biological sequences

- Ungapped blocks
  - DNA binding sites



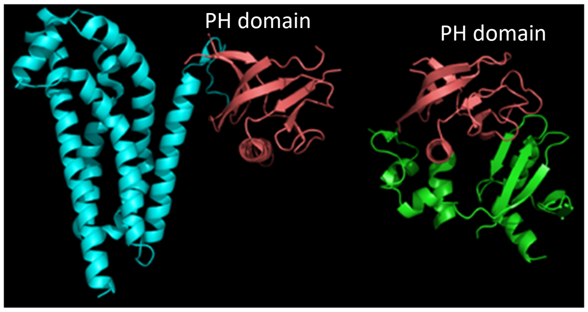
1

Wikipedia.org

Conserved patterns in biological sequences
Probabilistic Framework

Types of patterns in biological sequences

- Ungapped blocks
  - DNA binding sites
- Variable length motifs
  - protein domains



Wikipedia.org

Conserved patterns in biological sequences
Probabilistic Framework

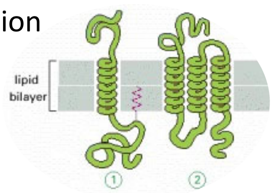
Types of patterns in biological sequences

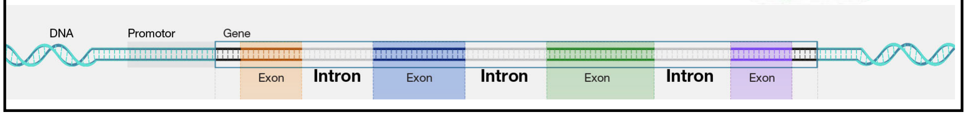
- Ungapped blocks
  - DNA binding sites
- Variable length motifs
  - protein domains
- Regions of characteristic sequence composition
  - transmembrane regions
  - GC rich regions
  - introns versus exons

Conserved patterns in biological sequences
Probabilistic Framework

Types of patterns in biological sequences

- Ungapped blocks
  - DNA binding sites
- Variable length motifs
  - protein domains
- Regions of characteristic sequence composition
  - transmembrane regions
  - GC rich regions
  - introns versus exons





```
... RLSKIISMFQAHIRGYLIRKAYKRGYQARCLLK ...
... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ...
.. GWQKRVRGWIVIVRRNFKKKRNEKLSATAZZZZZYQ ...
 ... MKRSQVVKQEKAARKVQKFWRGHRVQHNQR ...
 ... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...
```

Discovery

```
 ... RLSKIISMIQAHIRGYLIRKAYKRGYQARCLLK ..
 ... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ..
... GWIQKRVRGWIVIRRNFKKKRNEKLSATAZZZZZYQ
   ... MKRSQVVKQEKAARKIQKFWRGHRVQHNQR ...
   ... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...
```

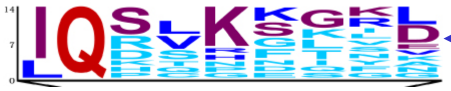
Discovery

Given multiple sequences, often unaligned, find a conserved pattern or *motif*

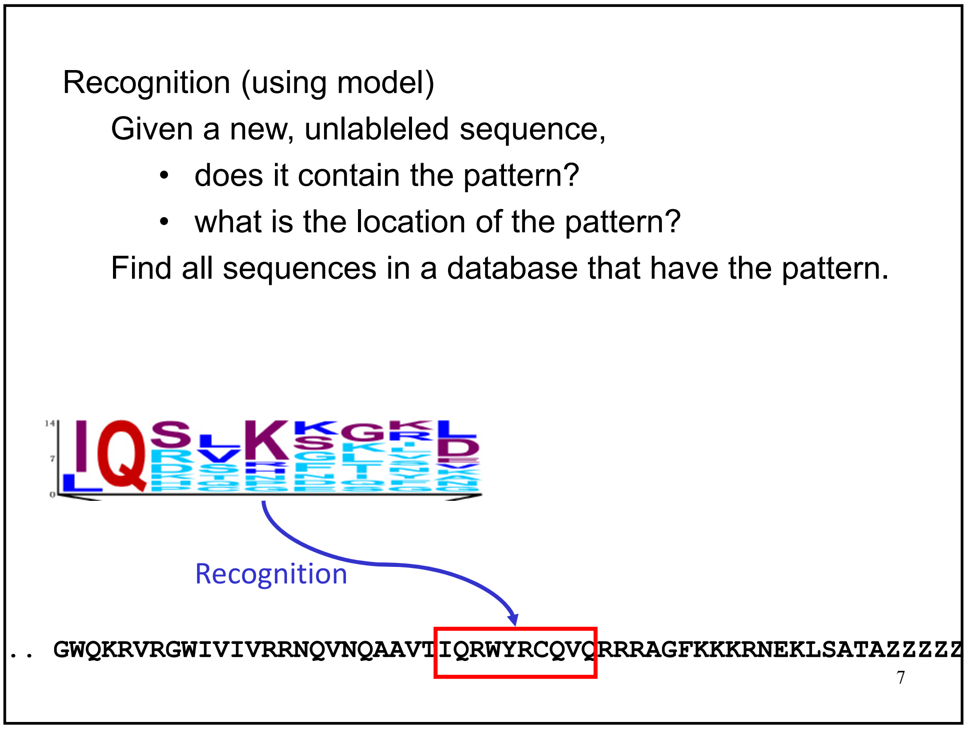
Modelling

Given an alignment of the motif (often ungapped), construct probabilistic model summarizing conserved features

```
 ... RLSKIISMIQAHIRGYLIRKAYKRGYQARCLLK ..
 ... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ..
... GWIQKRVRGWIVIRRNFKKKRNEKLSATAZZZZZYQ
   ... MKRSQVVKQEKAARKIQKFWRGHRVQHNQR ...
   ... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...
```



Modeling



Recognition (using model)

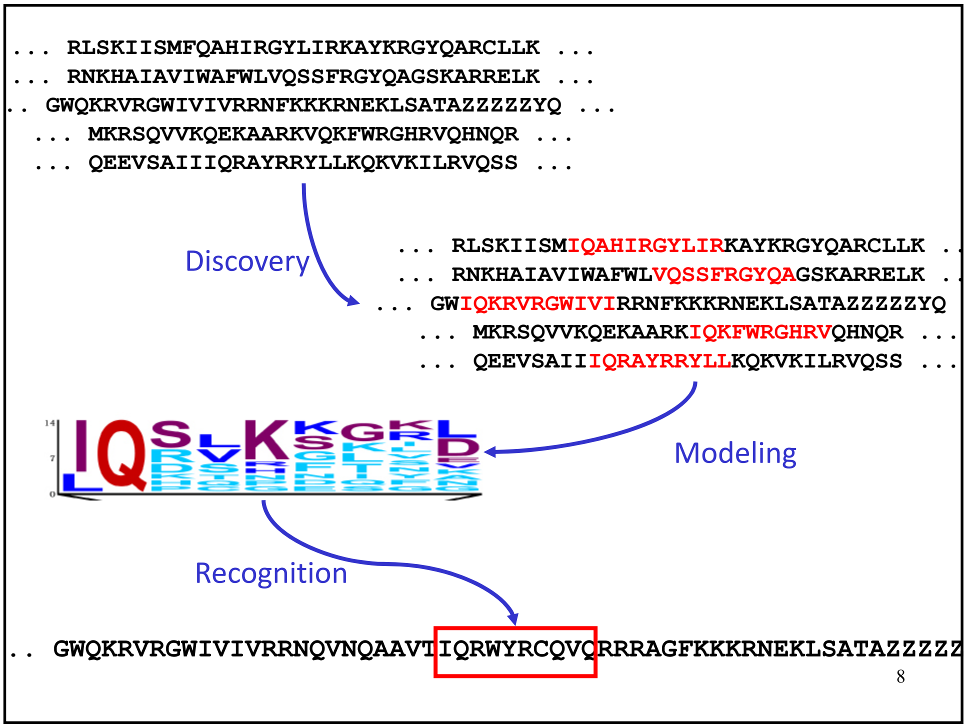
Given a new, unlableled sequence,

- does it contain the pattern?
- what is the location of the pattern?

Find all sequences in a database that have the pattern.

Recognition

.. GWQKRVRGWIVIVRRNQVNQAAVTIQRWYRCQVQRRRAGFKKKRNEKLSATAZZZZZ



... RLSKIISMFQAHIRGYLIRKAYKRGYQARCLLK ...
... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ...
.. GWQKRVRGWIVIVRRNFKKKRNEKLSATAZZZZZYQ ...
... MKRSQVVKQEKAARKVQKFWRGHRVQHNQR ...
... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...

Discovery

... RLSKIISMIQAHIRGYLIRKAYKRGYQARCLLK ..
... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ..
... GWIQKRVRGWIVIRRNFKKKRNEKLSATAZZZZZYQ
... MKRSQVVKQEKAARKIQKFWRGHRVQHNQR ...
... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...

Modeling

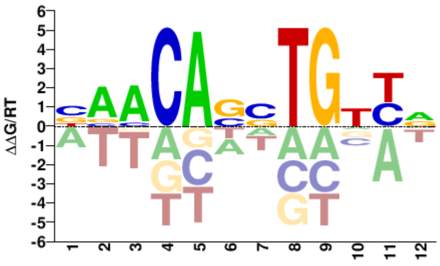
Recognition

.. GWQKRVRGWIVIVRRNQVNQAAVTIQRWYRCQVQRRRAGFKKKRNEKLSATAZZZZZ

Conserved patterns in biological sequences
Probabilistic Framework

- Discovery
  - Given multiple sequences, often unaligned, find a conserved pattern or *motif*
- Modelling
  - Given an alignment of the motif (often ungapped), construct probabilistic model summarizing conserved features
- Recognition (using model)
  - Given a new, unlableled sequence,
    - does it contain the pattern?
    - what is the location of the pattern?
  - Find all sequences in a database that have the pattern.

Last Thursday, we constructed a PSSM to model the binding site of the human MyoD1 transcription factor

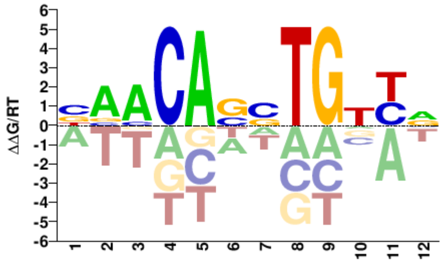


MyoD1 binding site: Homo sapiens
Yin et al, 2017 (https://doi.org/10.1126/science.aaj2239)

http://kstest.med.utoronto.ca/TFreport.php?searchTF=T034283_2.00

PSSMs

- Model fixed width motifs: Given *k* sequence segments of length *w*, derive $w \times \Sigma$ Position Specific Scoring Matrix, *S*, where the *i*'th column represents the distribution of symbols at site *i.*
- $S[a, i]$ is the log odds score associated with observing symbol *a* at site *i* in the motif.



PSSMs

- Model fixed width motifs: Given *k* sequence segments of length *w*, derive $w \times \Sigma$ Position Specific Scoring Matrix, *S*, where the *i*'th column represents the distribution of symbols at site *i.*
- $S[a, i]$ is the log odds score associated with observing symbol *a* at site *i* in the motif.
- Score a segment (or window) of length *w* in a new sequence, *t*:
  - Positive score: the segment is more likely than chance to represent an instance of the motif.
- More generally, score a sliding window to find the highest scoring window or find all candidates with score above some threshold.

PSSMs

- Model fixed width motifs: Given *k* sequence segments of length *w*, derive $w \times \Sigma$ Position Specific Scoring Matrix, *S*, where the

Gibbs Sampler
Motif discovery method that uses the PSSM formalism as its basic data structure.
Given window size *w*, finds ungapped patterns in unlabeled sequences using a Markov Chain Monte Carlo approach. We will not cover this material this year.

- More generally, score a sliding window to find the highest scoring window or find all candidates with score above some threshold.

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

Some of these problems can be handled by Markov chains….

Problems with PSSMs

Variable length patterns that are not position specific

Patterns characterized by changes in sequence composition, e.g.

- CpG islands
- Transmembrane domains

Transmembrane Regions

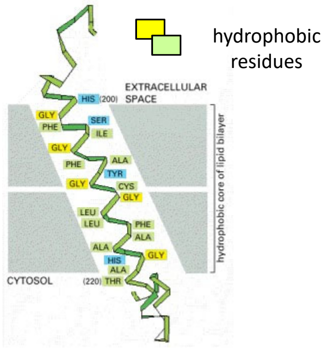


Figure 10-19 A segment of a transmembrane polypeptide chain crossing the lipid bilayer as an α helix

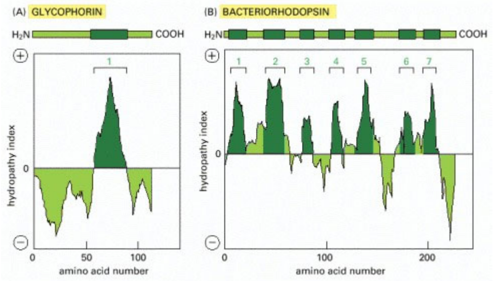
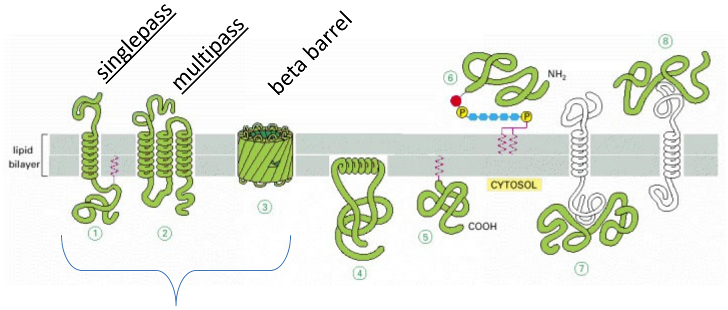


Figure 10-20Using hydropathy plots to localize potential α-helical membrane-spanning segments in a polypeptide chain

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/

Figure 10-17 Various ways in which membrane proteins associate with the lipid bilayer



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/

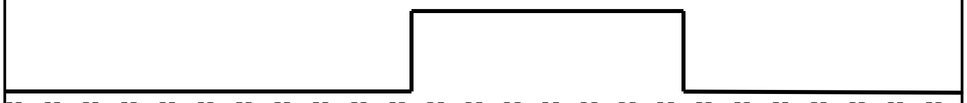
Problems with PSSMs

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

Some of these problems can be handled by Markov chains….

Boundary Detection

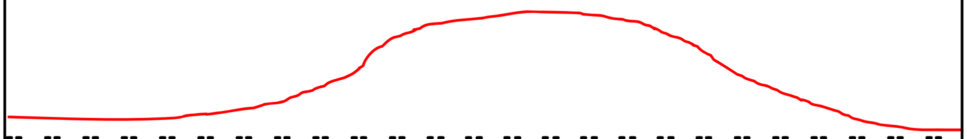
Goal: label every element in the sequence with a zero (not in pattern) or a one (in pattern)



```
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
```

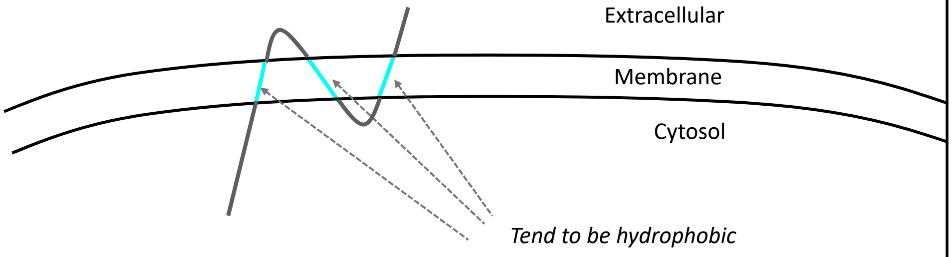
Methods that use sliding window scoring do not handle boundary detection problems well

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



```
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 .1.2.3.8.9.9.9.8.7.7.5.2.1 0 0 0 0
```

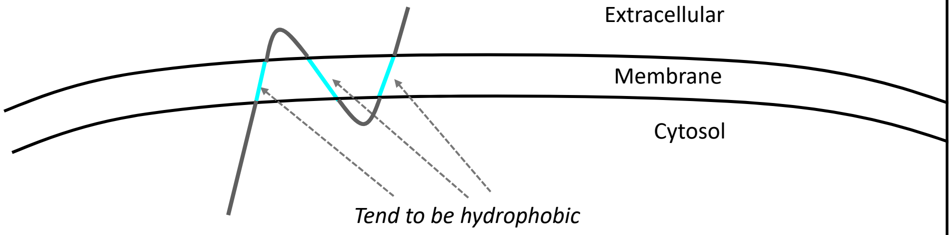
An example: transmembrane regions



Boundary detection problem:

Given sequence of amino acids, find all transmembrane regions

An example: transmembrane regions



Is a given sequence, O, a transmembrane sequence? Is $\frac{P(O|TM)}{P(O|H_0)}$ >>1?

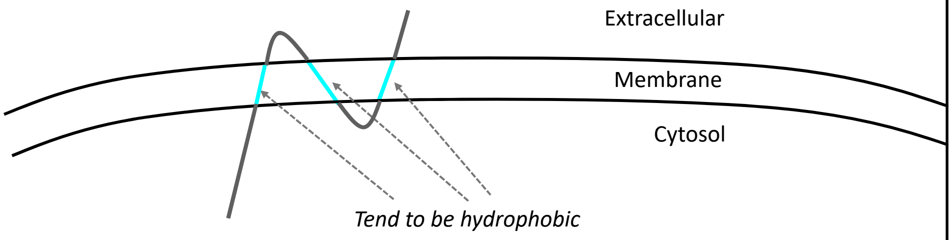
Solve this with a Markov chain or an HMM

Boundary detection problem:

Given sequence O, find all transmembrane regions

Requires labeling each residue with its location in the cell

An example: transmembrane regions



Is a given sequence, O, a transmembrane sequence? Is $\frac{P(O|TM)}{P(O|H_0)}$ >>1?

Solve this with a Markov chain or an HMM

Boundary detection problem:

Given sequence O, find all transmembrane regions

Requires labeling each residue with its location in the cell

This requires an HMM

Examples of boundary detection problems

- Recognition of regulatory motifs
- Recognition of protein domains
- Intron/exon boundaries
- Gene boundaries
- Transmembrane regions
- Secondary structures (α helices, β sheets)

Markov chains

- States: E_1, E_2, ... E_s
- States visited: q_0, q_1, …., q_t, q_{t+1}, …
- Initial distribution of states: $\pi(i) = P(q_0 = E_i)$
- Transition probabilities: $P_{ij} = P(q_t = E_j \mid q_{t-1} = E_i)$

Markov chains can model the probability of observing O, but cannot not label individual residues in O.

Modeling transmembrane proteins with...

Markov chains

States: E_1, E_2, ... E_N

Initial state probabilities: $\pi(i)$

Transition probabilities: P_{ij}

- Each state represents a symbol
- A sequence corresponds to a single state path.
- State path associates a probability with a sequence:

$$\text{e.g., } \frac{P(O|TM)}{P(O|H_0)}$$

Hidden Markov models

States: E_1, E_2, ... E_N

Initial state probabilities: $\pi(i)$

Transition probabilities: a_{ij}

Alphabet, Σ

Emission probabilities: e_i

- Each state represents a cellular compartment
- A sequence corresponds to potentially many state paths
- A state path labels residues with inferred localization

HMMs

States: E_1, E_2, ... 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”).

Parameter estimation

- from labeled data
- from unlabaled data

Read the section on calculating parameters
from labeled sequences before the next lecture
(see first half of 7.6.2).

HMMs

States: $E_1, E_2, \ldots E_N$

Initial state probabilities: $\pi(i)$

Transition probabilities: a_{ij}

Alphabet, Σ

Emission probabilities: e_i

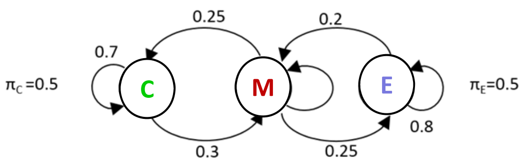
The parameters of the HMM $\lambda=(a_{ij}, e_i(\sigma), \pi)$

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:



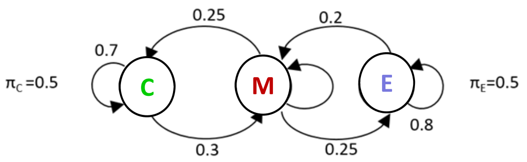
$e_C(H)$	0.3
$e_C(L)$	0.7

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

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

- 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
    - hydrophobic (H)
    - hydrophilic (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

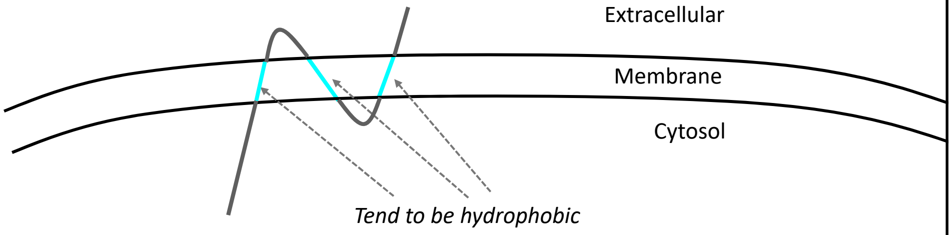
e_E(H)	0.2
e_E(L)	0.8

- A state can emit more than one symbol
- Each symbol can be emitted by more than one state

An HMM generates *labeled* sequences:

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

Recognition problems



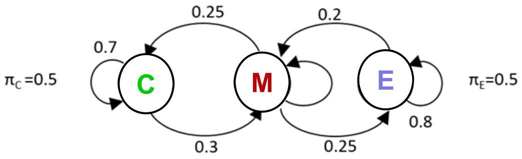
Does a given sequence, O, encode a transmembrane protein?

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

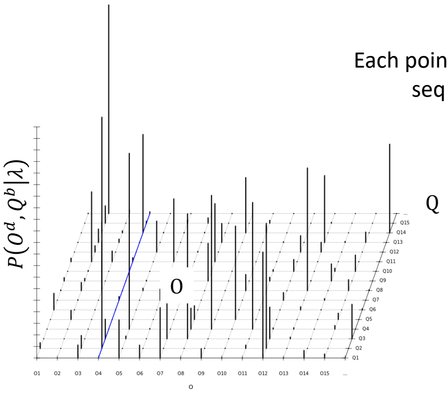
$e_M(H)$	0.9
$e_M(L)$	0.1

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

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

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

An HMM defines a probability distribution over sequences and state paths



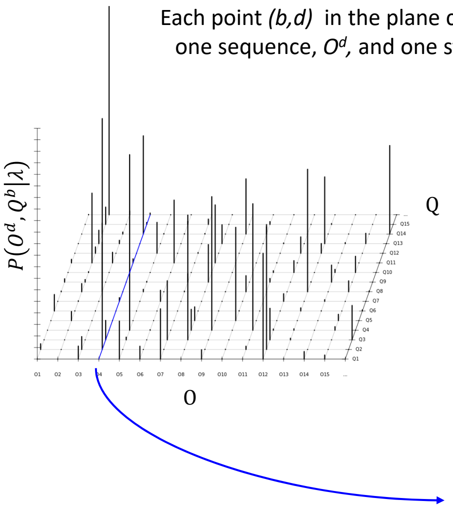
Each point *(b,d)* in the plane corresponds to one sequence, O^d*,* and one state path, Q^b

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

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

O4

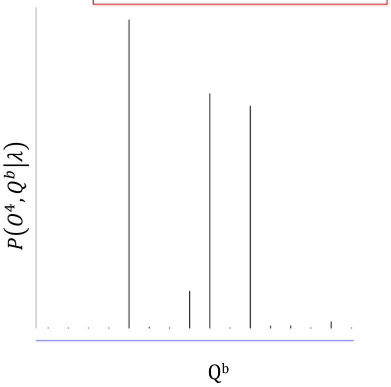
Each point *(b,d)* in the plane corresponds to one sequence, O^d, and one state path, Q^b



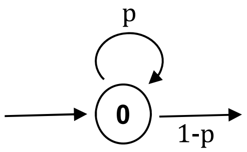
This plane corresponds to all ways to emit sequence, O^d . Each point *b* on the horizontal axis corresponds to one state path, Q^b

For a given sequence, O

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



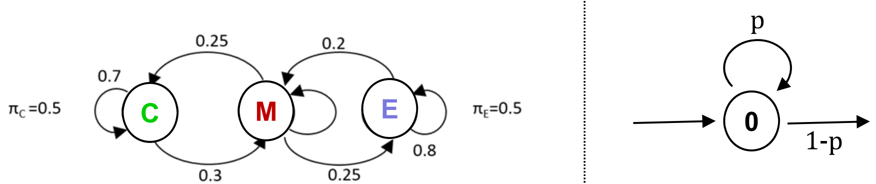
A null model



$e_0(H)$	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

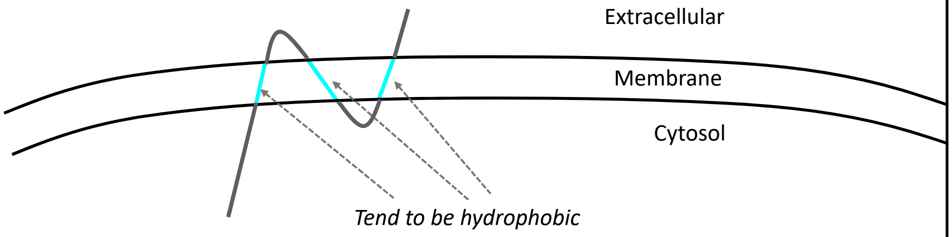
$e_M(H)$	0.9
$e_M(L)$	0.1

$e_E(H)$	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

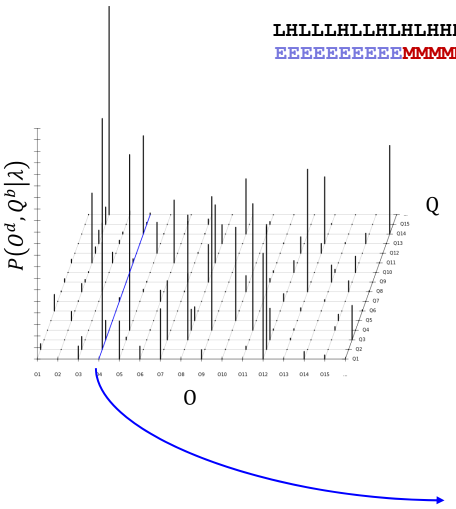


Boundary detection problem:

Find all transmembrane regions in a given sequence

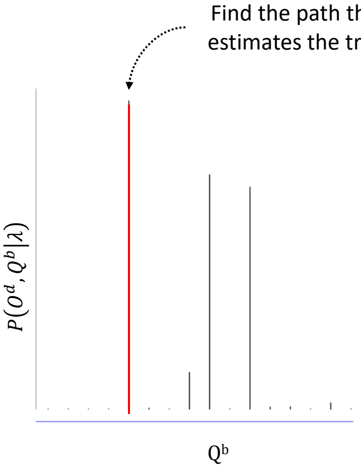
Requires labeling each residue with its location in the cell

LHLLLHLLHLHLHHHHHHLHLHLLHHLLHHHHHLHLLLLLHLL...
EEEEEEEEEEMMMMMMMMCCCCCCCCCCMMMMMMMEEEEEEE...



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

Find the path that best estimates the true path



Recognition problems

- What is the probability of a given sequence?

  Example: given HHLHH, is it a TM sequence or not?

- Given a sequence of symbols, what is the “true” sequence of states?

  Example: given HHHLLHL…, where is the TM region?

- What state emitted the symbol O_t?

  Example: is the isoleucine at position 32 localized to the membrane?