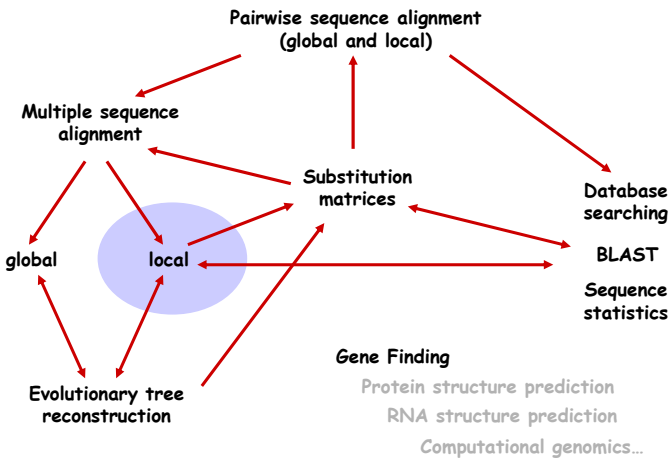
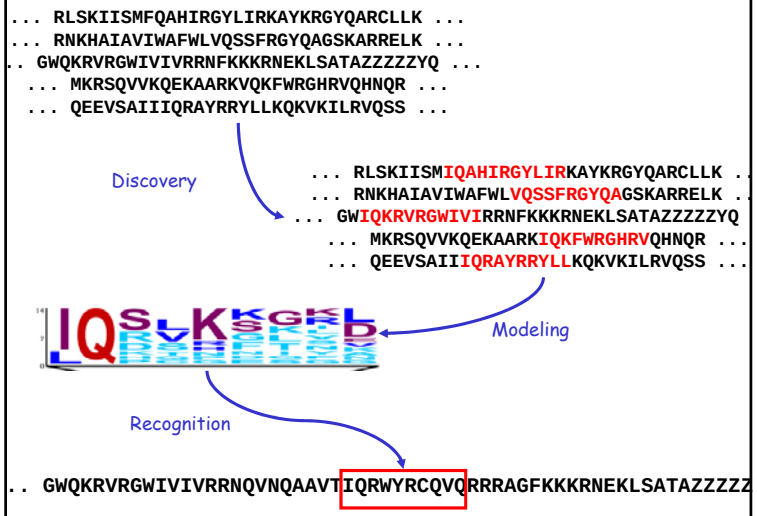




Local Multiple Alignment

- Position Specific Scoring Matrices (PSSMs)
  - Modeling, Recognition
- Gibbs sampler
  - Discovery
- Hidden Markov Models (HMMs)
  - Discovery, Modeling, Recognition
  - Can represent gaps, positional dependencies

Position Specific Scoring Matrices

Given *A[l,k]* (*k* sequences, *l* positions),

- Frequency of aa *i* at position *j* $F[i,j] = \frac{n_{ij}}{k}$
- Propensity of aa *i* at position *j* $P[i,j] = \frac{F[i,j]}{b_i}$
- Log odds scoring matrix $S[i,j] = \log_2 P[i,j]$

$$\sum_{l=0}^{w-1} S[t[i+l],l]$$

Local Multiple Alignment

- Position Specific Scoring Matrices (PSSMs)
  - Modeling, Recognition
- ➢ Gibbs sampler
  - Discovery
- Hidden Markov Models (HMMs)
  - Discovery, Modeling, Recognition
  - Can represent gaps, positional dependencies

Discovery

- Input: *k* sequences containing a common ungapped pattern (e.g., a transcription factor binding site, a domain…)
- Output: A set of *k* subsequences that are "most similar" to each other.
- Approaches
  - Exhaustive enumeration
  - ➢ Gibbs sampler
  - Expectation maximization using HMMs

Gibbs sampler summary

Convergence: (see optional reading for details)

- Model sampling process as a Markov Chain
- Each state is a set of *k* subsequences
- Show that
  - the Markov Chain has a stationary distribution
  - the state corresponding to the most likely pattern has high probability in that distribution

In practice, the sampler can get stuck in local optima

- Randomness helps.
- Run the procedure several times with different starting configurations.

Gibbs sampler summary

Other considerations:

- Problems could arise if a sequence has no copy of the pattern or has more than one copy
- You could find a biologically meaning pattern that is not the pattern you are looking for.
- Use pseudocounts in PSSM to ensure all characters are represented.

Black Magic (see Lawrence et al, optional reading)

- Pseudocounts
- Selecting the window size, *w*
- Selecting the starting configuration
- Termination condition.

Local Multiple Alignment

- Position Specific Scoring Matrices (PSSMs)
  - Modeling, Recognition
- Gibbs sampler
  - Discovery
- ➢ Hidden Markov Models (HMMs)
  - Discovery, Modeling, Recognition
  - Can represent gaps, positional dependencies

Problems with PSSMs

Do not capture positional dependencies

WEIRD
WEIRD
WEIQH
WEIRD
WEIQH

→

D					0.60
E		1.00			
H					0.40
I			1.00		
Q				0.40	
R				0.60	
W	1.00				

Note: We never see QD or RH, only RD and QH.
But, P(RH) = P(QD) = 0.24, while P(QH) = 0.16

Problems with PSSMs

Hard to recognize pattern instances that contain indels

D	0.8	0.8	0.8	0.8	2.4
E	0.6	2.9	0.6	0.6	1.6
H	2.0	2.0	2.0	2.0	3.0
I	0.8	0.8	3.1	0.8	0.8
Q	1.1	1.1	1.1	2.1	1.1
R	0.8	0.8	0.8	2.8	0.8
W	5.0	2.7	2.7	2.7	1.8

W E T I R D

5.0+2.9+1.2+1.4+1.5 = 11

W E T I R D

1.2+1.8+3.1+3.0+3.4 = 12.5

W E T I R D

5.0+2.9+3.1+3.0+3.4= 18.4

Problems with PSSMs

Variable length motifs

WETIRD
WE_IRD
WETIQH
WE_IRD
WETIQH

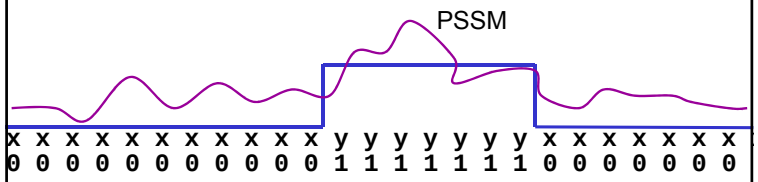
Gaps can be represented by expanding Σ, but what size window should be used to score new instances of the motif?

x x x W E T I R D x x x x x x W E I Q H x x x x x

Problems with PSSMs

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)



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)

Plan

- Review Markov chains
- Extend to Hidden Markov Models
  - Boundary detection
  - Scoring sequences
- HMM construction
- Biological applications: revisit gaps and dependencies.

Markov chains

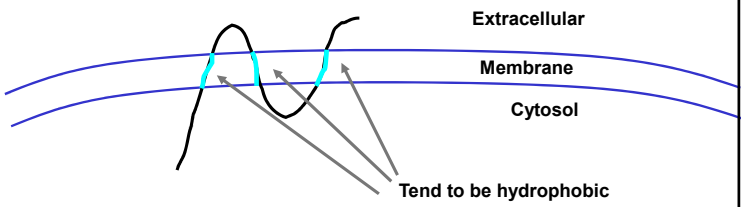
- States: $S_1, S_2, \ldots S_N$
- States visited: $q_0, q_1, \ldots, q_t, q_{t+1}, \ldots$
- Initial distribution of states: $\pi(i) = P(q_0 = S_i)$
- Transition probabilities: $a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)$

Questions we can ask:

What is the probability of being in a particular state at a particular time?

What is the probability of seeing a particular sequence of states?

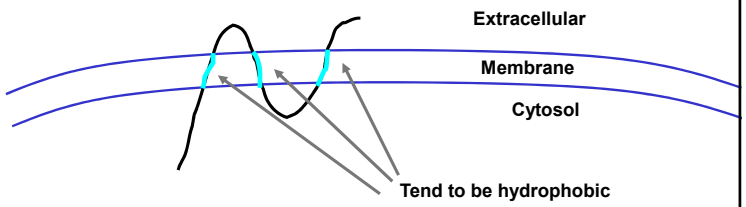
An example: transmembrane regions



Model each amino acid as hydrophobic (H) or hydrophilic (L)
→ A peptide sequence can be represented as a sequence of H's and Ls.

MLVKRFWKCE…. ➔ HHHLLHLHHLHL...

An example: transmembrane regions

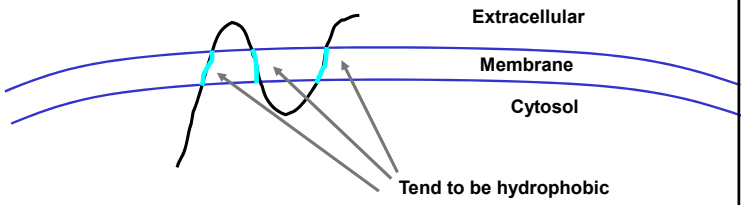


Questions to ask:

which subsequences correspond to transmembrane regions?

HHHLLHLHHLHL...

An example: transmembrane regions

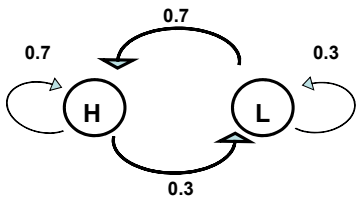


A simpler question:

is a given sequence a transmembrane sequence?

HHHLLHLHHLHL...

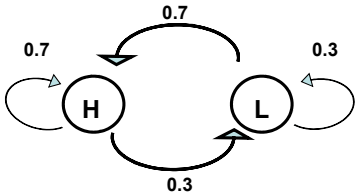
A Markov chain for recognizing transmembrane sequences



- States: S_H, S_L
- $\Sigma = \{H,L\}$
- $\pi(H) = 0.7$, $\pi(L) = 0.3$

Is a given sequence, say HHLHH,
a transmembrane sequence?

Transmembrane model:



π(H) = 0.7, π(L) = 0.3

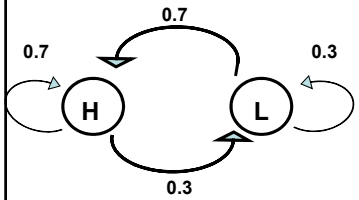
P(HHLHH) = 0.7 x 0.7 x 0.3 x 0.7 x 0.7 = 0.072

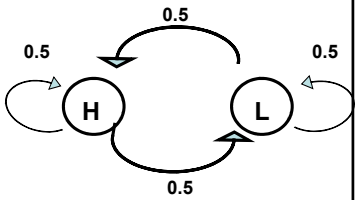
Is it a transmembrane protein?

Problem: need a threshold,

threshold must be length dependent

Transmembrane model: **Null model:**

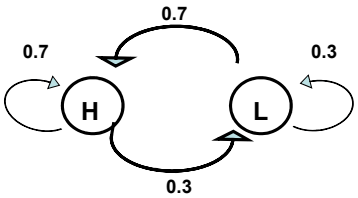




π(H) = 0.7, π(L) = 0.3 π(H) = 0.5, π(L) = 0.5

$$\frac{P(HHLHH \mid TM)}{P(HHLHH \mid EC)} = \frac{0.7 \times 0.7 \times 0.3 \times 0.7 \times 0.7}{0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5} = \frac{0.072}{0.0.31} = 2.3$$

Transmembrane model:

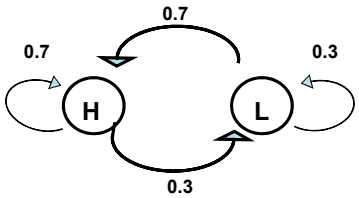


π(H) = 0.7, π(L) = 0.3

How are transition probabilities determined?

From known transmembrane sequences

Transmembrane model:



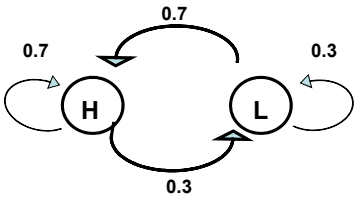
π(H) = 0.7, π(L) = 0.3

```
HHHLLHHHLLLHLHLLHLLLHLHHHL
HHHLLHHHLLLHLHLLHLLLHLHHHL
HL...
```

$$a_{ij} = \frac{A_{ij}}{\sum_l A_{il}}$$

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

Transmembrane model:



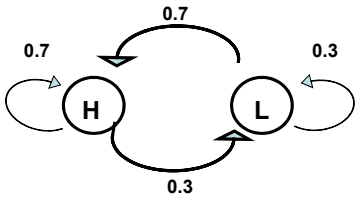
π(H) = 0.7, π(L) = 0.3

```
HHHLLHHHLLLHLHLLHLLLHLHHHL
HHHLHHLHLLLLLHHHHLLLHHHHHL
HH...
```

$$a_{HL} = \frac{A_{HL}}{\sum_i A_{Hi}}$$

$\frac{12}{\text{\# of H* pairs}}$

Transmembrane model:

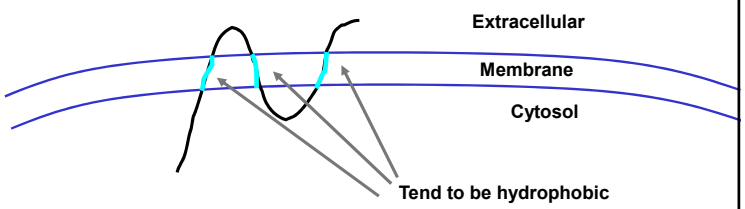


π(H) = 0.7, π(L) = 0.3

```
HHHLLHHHLLLHLHLLHLLLHLHHHL
HHHLHHLHLLLLLHHHHLLLHHHHHL
HH...
```

π(H) = # of sequences that begin with H,
normalized by the total # of training sequences

An example: transmembrane regions



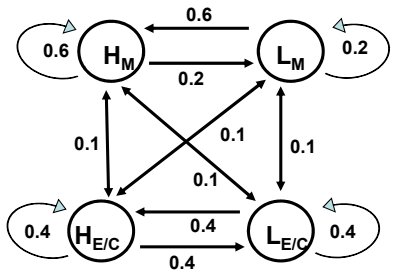
Boundary detection problem:

Given sequence of H's & L's, find all transmembrane regions

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

Markov chains can handle positional dependencies, indels and variable length motifs, but boundary detection is still a problem



$\pi(H_M) = 0.4$, $\pi(L_M) = 0.1$

$e_H(H_M) = 1.0$,
$e_L(L_M) = 1.0$,
$e_H(H_{EC}) = 1.0$,
$e_L(L_{EC}) = 1.0$

$\pi(H_{EC}) = 0.2$, $\pi(L_{EC}) = 0.3$

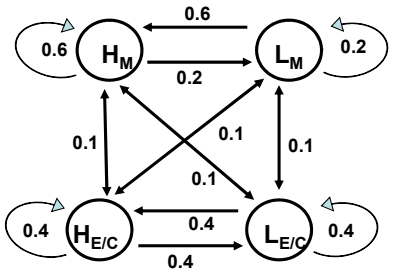
- *An **HMM** is a generative model; gives probability of generating a particular sequence*

Markov Chains	**HMMs**
States: $S_1, S_2, \ldots S_N$	States: $S_1, S_2, \ldots S_N$
Initial state probabilities: $\pi(i)$	Initial state probabilities: $\pi(i)$
Transition probabilities: a_{ij}	Transition probabilities: a_{ij}
	Alphabet, Σ
	Emission probabilities: e_i

We refer to the initial state, transition and emission probabilities as the parameters of the HMM: $\lambda=(a_{ij}, e_i(a), \pi)$

As before, the parameters are "learned" from known examples ("labeled data").

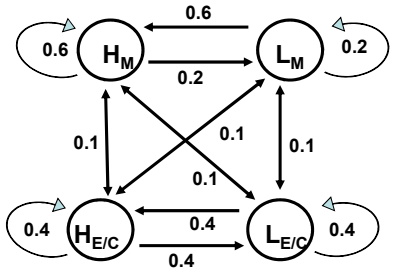
For the TM and E/C *Markov chains*,
sequence of symbols = sequence of states
Given HLLLHLLL, know $q_0 = S_H$, $q_1 = S_L$, $q_2 = S_L$...



$\pi(H_M) = 0.4$, $\pi(L_M) = 0.1$

$\pi(H_{EC}) = 0.2$, $\pi(L_{EC}) = 0.3$

In this HMM model, what is q_0, q_1, q_2 ... ?



$\pi(H_M) = 0.4$, $\pi(L_M) = 0.1$

$\pi(H_{EC}) = 0.2$, $\pi(L_{EC}) = 0.3$

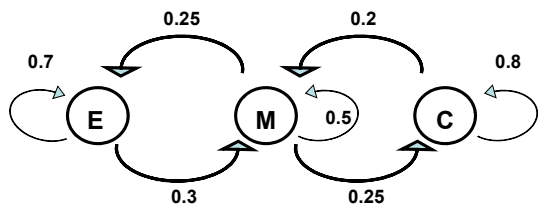
$e_H(H_M) = 1.0$, $e_L(L_M) = 1.0$

$e_H(H_{EC}) = 1.0$, $e_L(L_{EC}) = 1.0$

Unlike Markov chains, in HMMs the states are *hidden*.

Given an unlabeled sequence, HHHLLHL...,
we infer the most probable state sequence to obtain the boundaries

A three state transmembrane HMM:



i	E	M	C
π_i	0	0	1
$e_i(H)$	0.2	0.9	0.3
$e_i(L)$	0.8	0.1	0.7

A state can emit more than one symbol

Questions to ask

- Given a sequence of symbols, what is the most probable set of states?

  Example: given HHHLLHL…, where is the TM region?
- What is the probability of a given sequence?

  Example: given HHLHH, is it a TM sequence or not?
- Given an HMM, generate sequences according the model

  Example: simulate transmembrane sequences