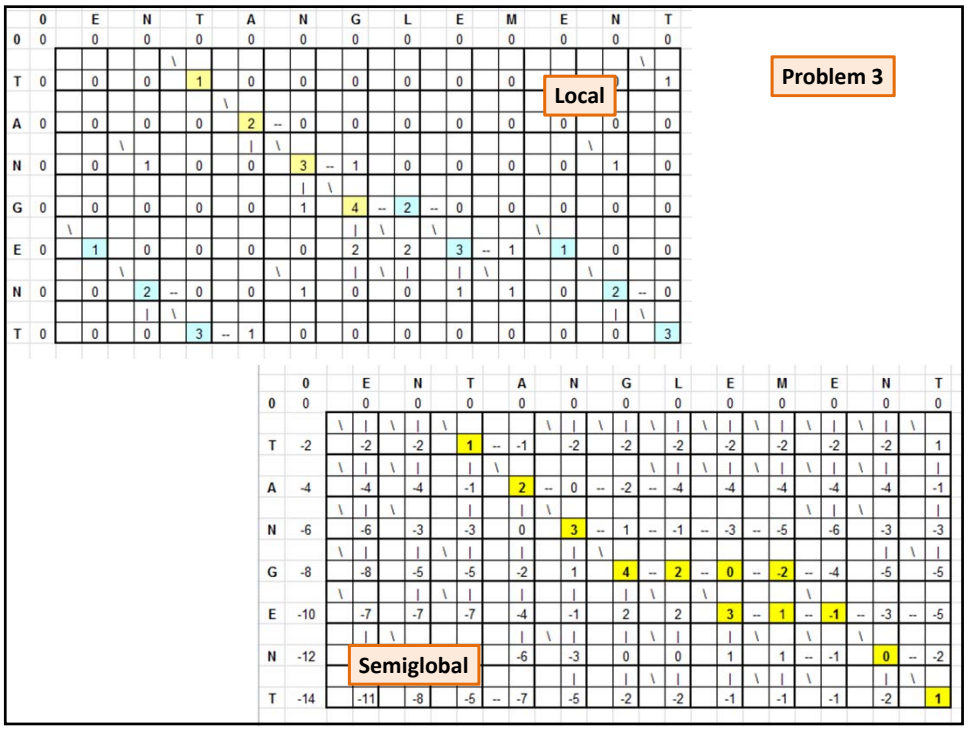
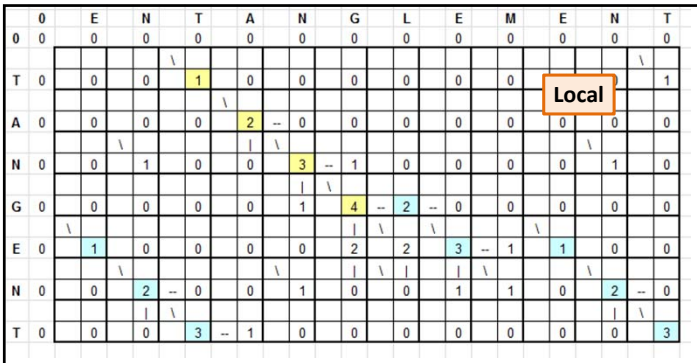
Thursday, September 29th

- Schedule
  - PS 2a online, due Oct 6
  - October 6: Class replaced by Science 2011
  - PS 2b out Oct 6, due Oct 13
  - Literature Assignment 2
  - Midterm Oct 20
- Problem set 1
- Gibbs sampler conclusion
- Motivation for Hidden Markov Models (HMMs)

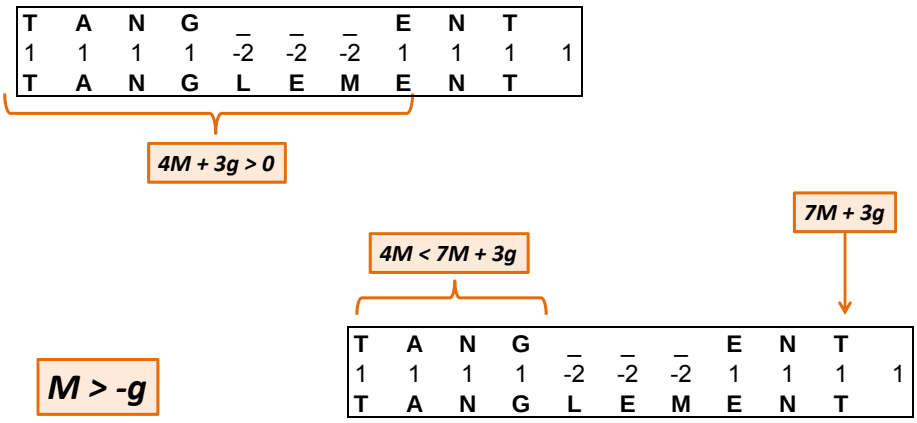


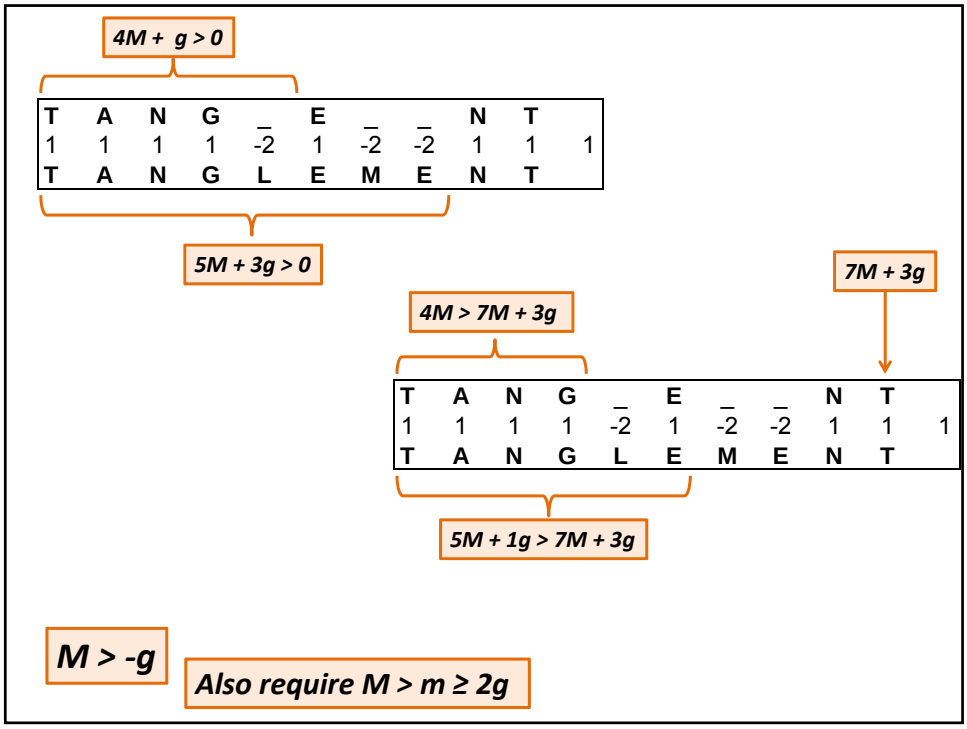


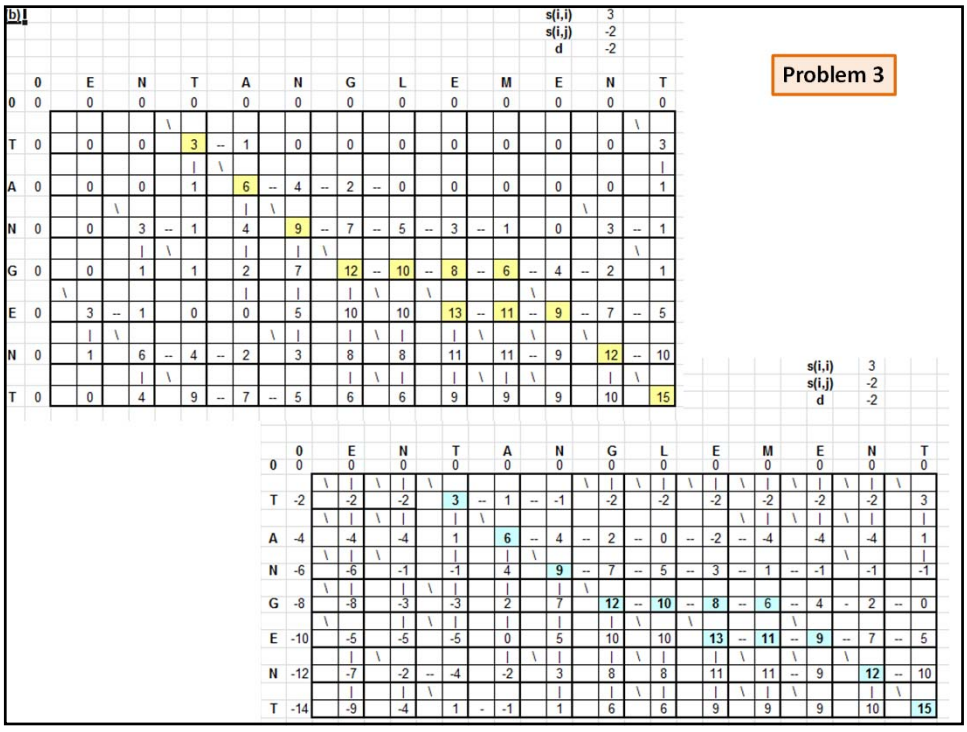
Problem 3

To get "global behavior" with the local alignment, you need

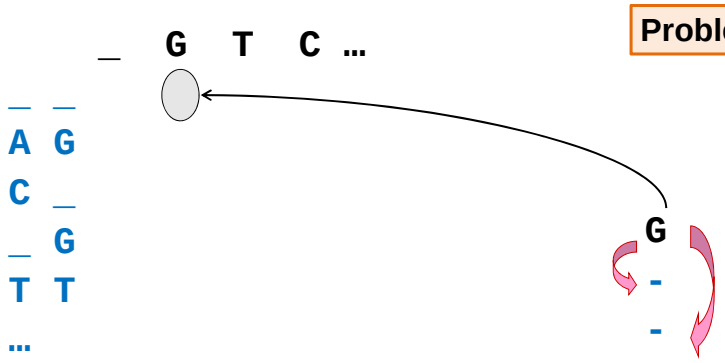
1. *d[m,n] ≥ d[i,j],* for all *1 ≤ i ≤ m* and *1 ≤ j ≤ n*
2. *d[i,j] > 0* on the path to *m,n.*




4M + g > 0
T A N G _ E _ _ N T
1 1 1 1 -2 1 -2 -2 1 1 1
T A N G L E M E N T
5M + 3g > 0
7M + 3g
4M > 7M + 3g
T A N G _ E _ _ N T
1 1 1 1 -2 1 -2 -2 1 1 1
T A N G L E M E N T
5M + 1g > 7M + 3g
M > -g
Also require M > m ≥ 2g

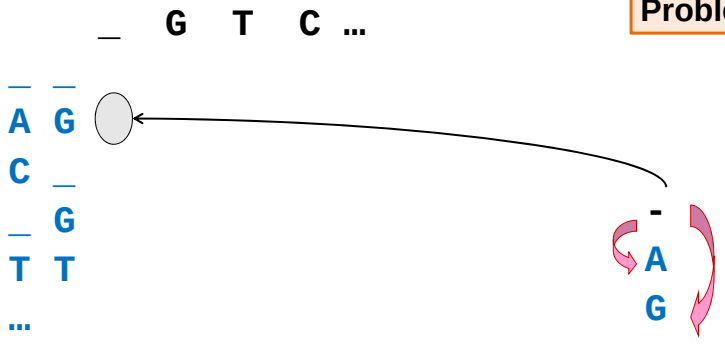

b)
s(i,i) 3
s(i,j) -2
d -2
Problem 3
s(i,i) 3
s(i,j) -2
d -2

Problem 4(a)



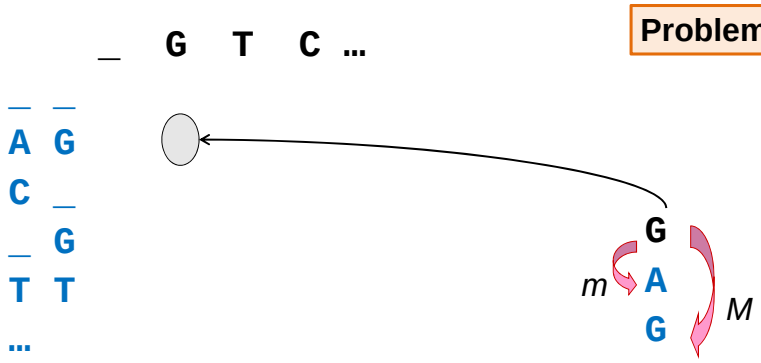
$$a[i,j] = \begin{cases} \underline{a[i,j-1] + 2g} \\ a[i-1,j-1] + p(x,t[j]) + p(y,t[j]) \\ a[i-1,j] + 2g \end{cases}$$

Problem 4(a)



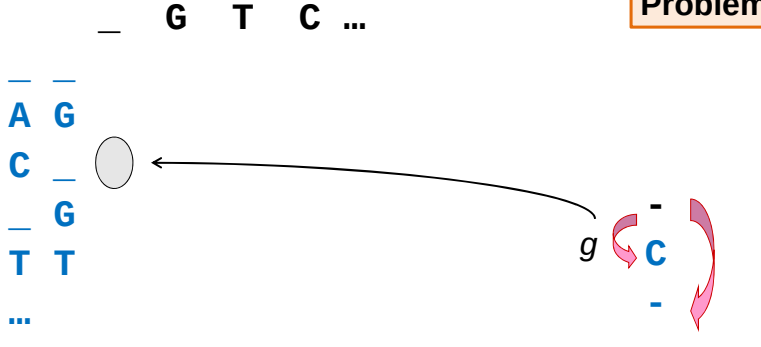
$$a[i,j] = \begin{cases} a[i,j-1] + 2g \\ a[i-1,j-1] + p(x,t[j]) + p(y,t[j]) \\ \underline{a[i-1,j] + 2g} \end{cases}$$

Problem 4(a)



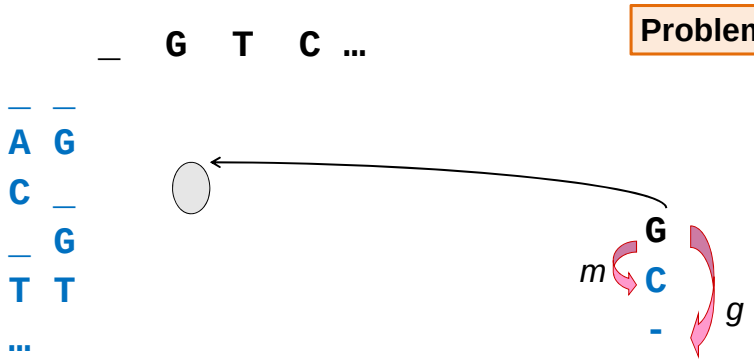
$$a[i,j] = \begin{cases} a[i,j-1] + 2g \\ \underline{a[i-1,j-1] + p(x,t[j]) + p(y,t[j])} \\ a[i-1,j] + 2g \end{cases}$$

Problem 4(b)



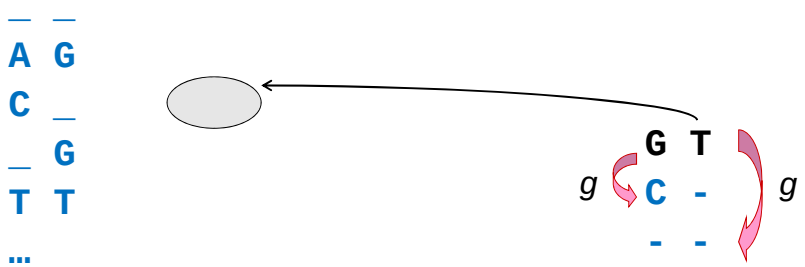
$$a[i,j] = \begin{cases} a[i,j-1] + 2g \\ a[i-1,j-1] + p(x,t[j]) + g \\ \underline{a[i-1,j] + g} \end{cases}$$

Problem 4(b)



$$a[i,j] = \begin{cases} a[i,j-1] + 2g \\ \underline{a[i-1,j-1] + p(x,t[j]) + g} \\ a[i-1,j] + g \end{cases}$$

Problem 4(b)



$$a[i,j] = \begin{cases} \underline{a[i,j-1] + 2g} \\ a[i-1,j-1] + p(x,t[j]) + g \\ a[i-1,j] + g \end{cases}$$

Problem5(e)

		T	C	-	G	T	C	-	T
	0	T	C	A	G	T	C	G	T
0	0	6	12	15	21	27	33	36	42
T	6	0	6	9					
A	12	6	4	?					
G	18								
T	24								
G	30								
T	36								

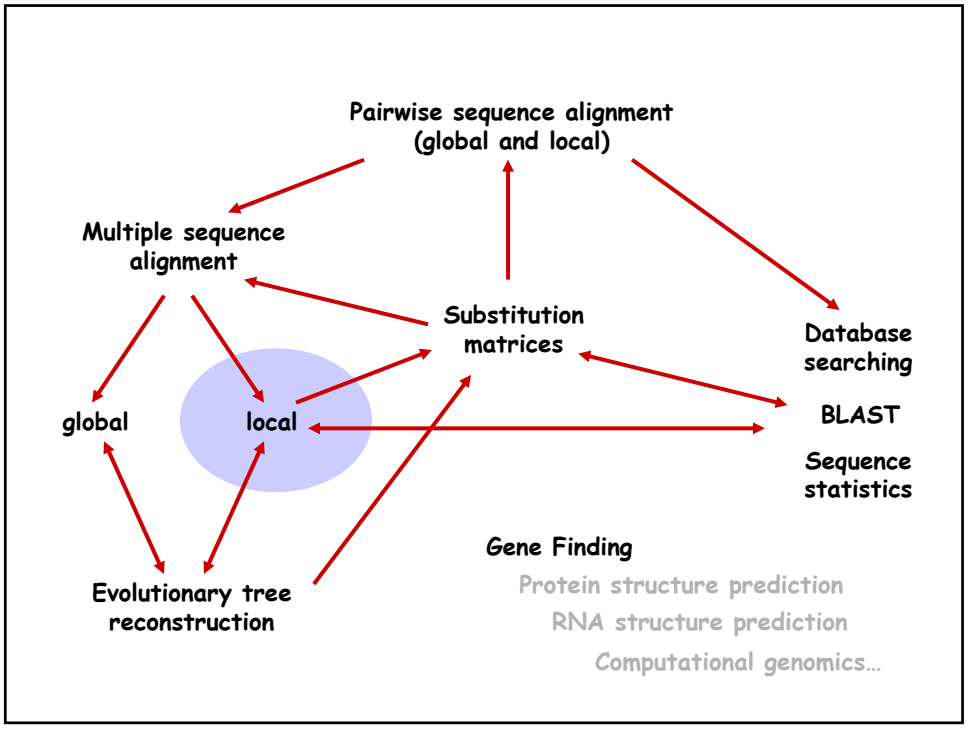
s(a,a)	0
s(a,b)	2
g	3

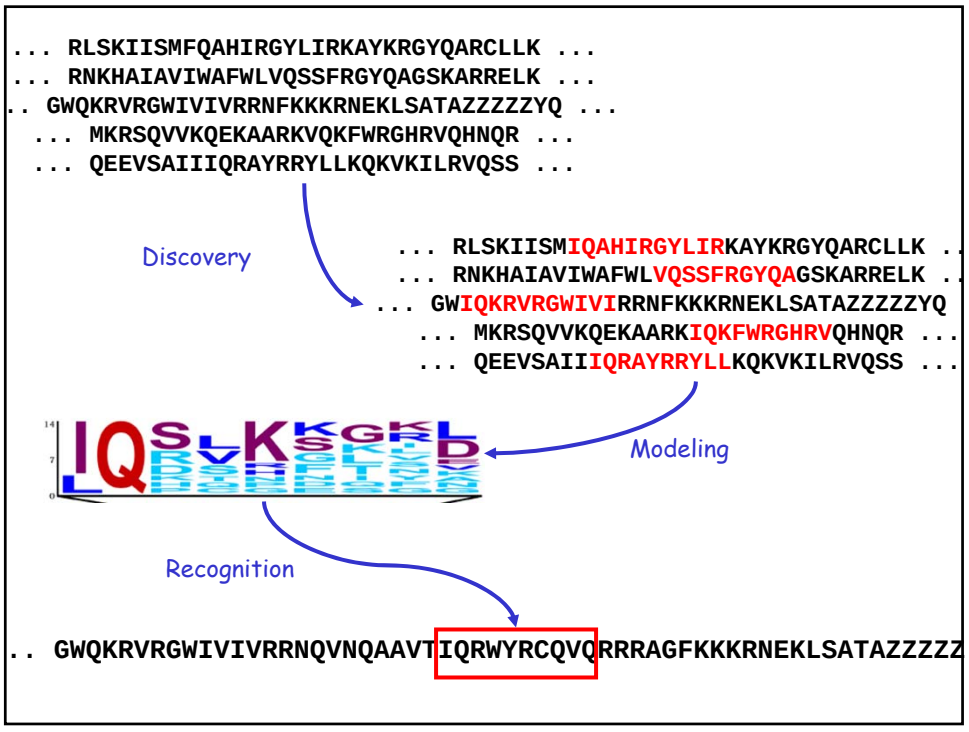
Problem5(e)

		T	C	-	G	T	C	-	T
	0	T	C	A	G	T	C	G	T
0	0	6	12	15	21	27	33	36	42
T	6	0	6	9					
A	12	6	4 → 7						
G	18								
T	24								
G	30								
T	36								

Add 1g

s(a,a)	0
s(a,b)	2
g	3


Pairwise sequence alignment
(global and local)
Multiple sequence alignment
Substitution matrices
Database searching
global
local
BLAST
Sequence statistics
Gene Finding
Protein structure prediction
RNA structure prediction
Computational genomics…
Evolutionary tree reconstruction


... RLSKIISMFQAHIRGYLIRKAYKRGYQARCLLK ...
... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ...
.. GWQKRVRGWIVIVRRNFKKKRNEKLSATAZZZZZYQ ...
... MKRSQVVKQEKAARKVQKFWRGHRVQHNQR ...
... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...
Discovery
... RLSKIISMIQAHIRGYLIRKAYKRGYQARCLLK ..
... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ..
... GWIQKRVRGWIVIRRNFKKKRNEKLSATAZZZZZYQ
... MKRSQVVKQEKAARKIQKFWRGHRVQHNQR ...
... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...
Modeling
Recognition
.. GWQKRVRGWIVIVRRNQVNQAAVTIQRWYRCQVQRRRAGFKKKRNEKLSATAZZZZZ

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

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 meaningful pattern that is not the pattern you are looking for.
- Use pseudocounts when building the 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

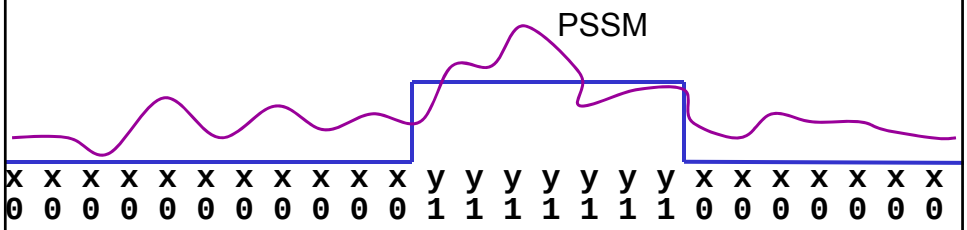
Gaps can be represented by expanding Σ, but what size window should be used to score new instances of the motif?

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

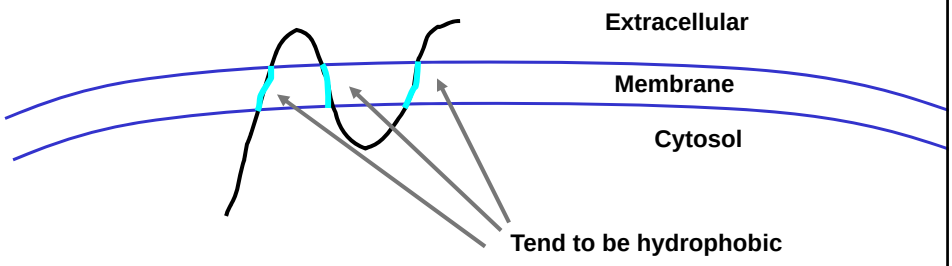
- States: S_1, S_2, ... S_N
- States visited: q_0, $_{q1}$, …., q_t, q_{t+1}, …
- Initial distribution of states: $\pi(i) = P(q_0 = S_i)$
- Transition probabilities: $a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)$

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?

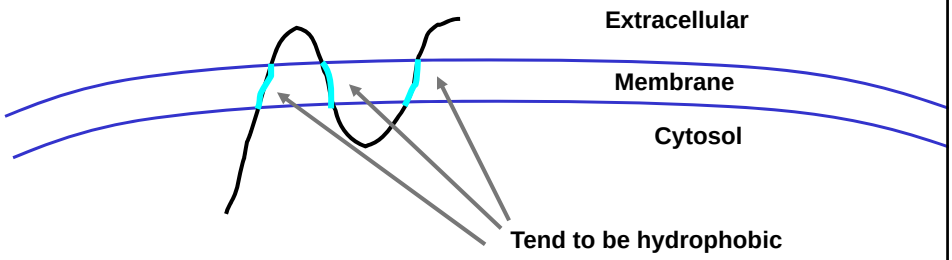
An example: transmembrane regions



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

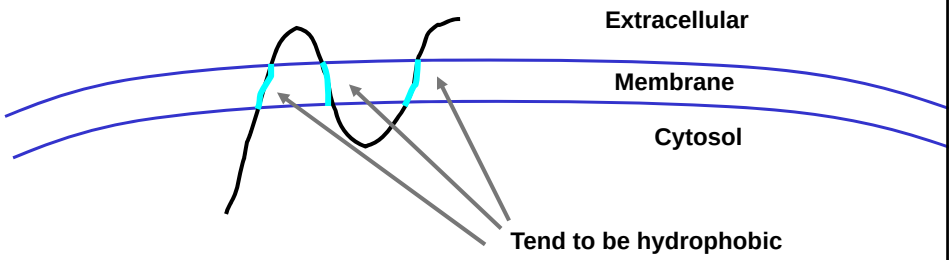


Questions to ask:

which subsequences correspond to transmembrane regions?

HHHLLHLHHLHL...

An example: transmembrane regions

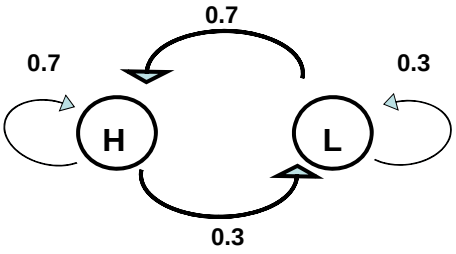


A simpler question:

is a given sequence a transmembrane sequence?

HHHLLHLHHLHL...

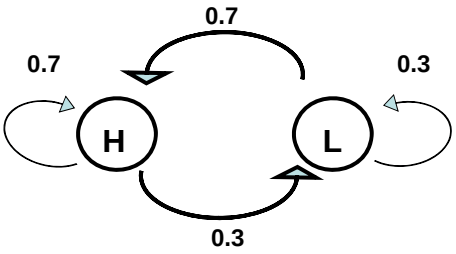
A Markov chain for recognizing transmembrane sequences



- 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:

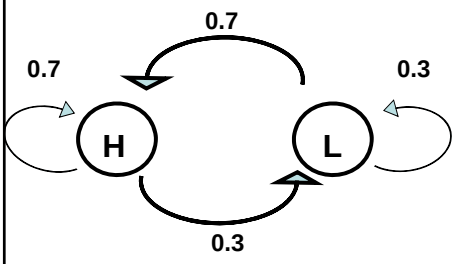


$\pi(H) = 0.7$, $\pi(L) = 0.3$

P(HHLHH) = 0.7 x 0.7 x 0.3 x 0.7 x 0.7 = 0.072
Is it a transmembrane protein?

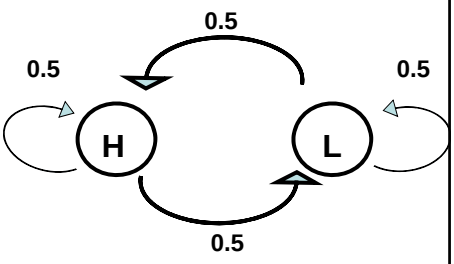
Problem: need a threshold,
threshold must be length dependent

Transmembrane model:



$\pi(H) = 0.7, \pi(L) = 0.3$

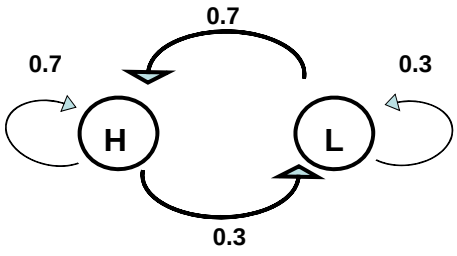
Null model:



$\pi(H) = 0.5, \pi(L) = 0.5$

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

Transmembrane model:

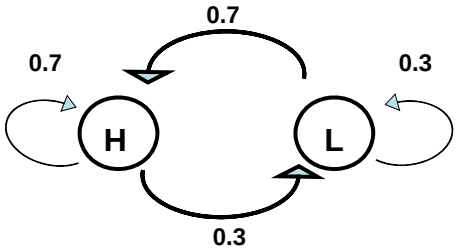


$\pi(H) = 0.7, \pi(L) = 0.3$

How are transition probabilities determined?

From known transmembrane sequences

Transmembrane model:



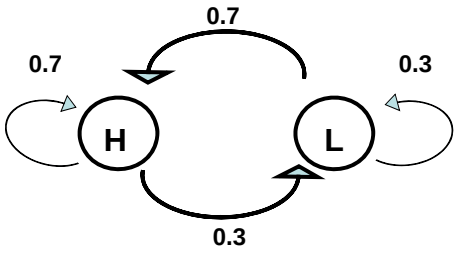
π(H) = 0.7, π(L) = 0.3

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:



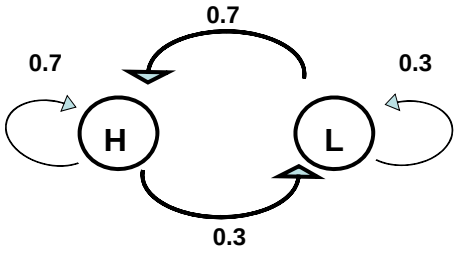
π(H) = 0.7, π(L) = 0.3

HHHLLHHHLLLHLHLLHLLLHLHHHL
HHHLHHLHLLLLLHHHHLLLHHHHHL
HH...

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

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

Transmembrane model:

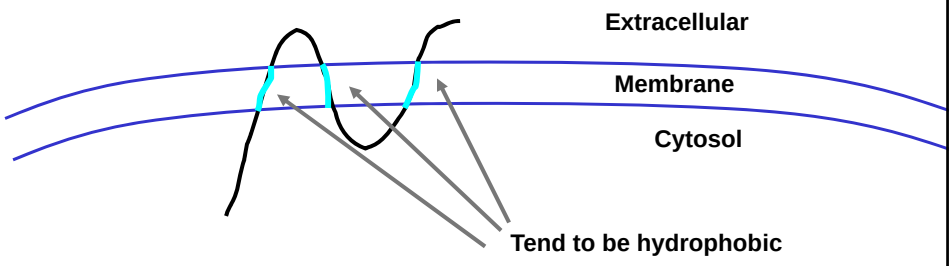


π(H) = 0.7, π(L) = 0.3

```
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