Second in-class exam scores

![Bar chart showing exam scores with a mean of 89.1.]

Problems with PSSMs
Do not capture positional dependencies

<table>
<thead>
<tr>
<th></th>
<th>D</th>
<th>E</th>
<th>H</th>
<th>I</th>
<th>Q</th>
<th>R</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEIRD</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WEIRD</td>
<td></td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WEIQH</td>
<td></td>
<td></td>
<td>0.4</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WEIRD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WEIQH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

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

- **WEIRD**: 5.0 + 2.9 + 1.2 + 1.4 + 1.5 = 11
- **WEIRD**: 5.0 + 2.9 + 3.1 + 3.0 + 3.4 = 18.4

Note: Discovery and Recognition models focus on different aspects of the data.
Problems with PSSMs
Motifs with indels

WETIRD
WE_IRD
WETIQH
WE_IRD
WETIQH

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

Problems with PSSMs

Variable length patterns that are not position specific

Patterns characterized by changes in sequence composition, e.g.
• Isochores
• Transmembrane domains

Amino acids recoded in a two letter alphabet representing hydrophobic and hydrophilic residues: $\Sigma = \{H,L\}$

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 (\(\alpha\) helices, \(\beta\) sheets)
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....

Variable length patterns
An example: transmembrane regions

A simpler question:
Is a given sequence a transmembrane sequence?

HHHLHLLLHHLHL . . .

An example: transmembrane regions

Boundary detection problem:
Given sequence of H’s & L’s, find all transmembrane regions

A Markov chain for recognizing transmembrane sequences

• States: E_H, E_L
• π(H) = 0.7, π(L) = 0.3

Is a given sequence, say HHLHH, a transmembrane sequence?
Transmembrane model:  

- $\pi(H) = 0.7$, $\pi(L) = 0.3$
- $P(H|HHL|TM) = 0.7 	imes 0.7 	imes 0.3 	imes 0.7 = 0.072$

Null model:  

- $\pi(H) = 0.5$, $\pi(L) = 0.5$
- $P(H|HHL|EC) = 0.5 	imes 0.5 	imes 0.5 	imes 0.5 = 0.031

$P(HHLHH|TM) = 0.072$, $P(HHLHH|EC) = 0.031$, $\frac{0.072}{0.031} = 2.3$

### 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*

### Boundary detection problem:

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

- $\pi(H) = 0.5$, $\pi(L) = 0.5$
- $P(HHLHH|TM) = 0.072$, $P(HHLHH|EC) = 0.031$
- $\frac{0.072}{0.031} = 2.3$

### An example: transmembrane regions

- Extracellular
- Membrane
- Cytosol
- Tend to be hydrophobic

Boundaries correspond to transitions between $M$ states and $E/C$ states

- $\pi(H_M) = 0.4$, $\pi(L_M) = 0.1$
- $\pi(H_{EC}) = 0.2$, $\pi(L_{EC}) = 0.3$
Markov Chains

States: $E_1, E_2, \ldots, E_N$
Initial state probabilities: $\pi(i)$
Transition probabilities: $P_{xy}$
Alphabet, $\Sigma$
Emission probabilities: $e_i$

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

An HMM is a generative model:

The parameters are “learned” from known examples (“labeled data”).

HMMs

States: $E_1, E_2, \ldots, E_N$
Initial state probabilities: $\pi(i)$
Transition probabilities: $a_{ij}$
Alphabet, $\Sigma$
Emission probabilities: $e_i$

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

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 $Q=q_1 q_2 q_3 \ldots q_T$”

For the TM and E/C Markov chains, sequence of symbols = sequence of states

Given HLLLHLLL, know $q_0 = E_H, q_1 = E_L, q_2 = E_L, \ldots$

In an HMM model, there are many possible state paths. The true sequence of states is hidden.
A three state transmembrane HMM:

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>M</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_i$</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>$e_i(H)$</td>
<td>0.3</td>
<td>0.9</td>
<td>0.2</td>
</tr>
<tr>
<td>$e_i(L)$</td>
<td>0.7</td>
<td>0.1</td>
<td>0.8</td>
</tr>
</tbody>
</table>

A state can emit more than one symbol

HMMs

States: $E_1, E_2, \ldots, E_N$

Initial state probabilities: $\pi(i)$

Transition probabilities: $a_{ij}$

Alphabet, $\Sigma$

Emission probabilities: $e_i$

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 $Q = q_1 \ q_2 \ q_3 \ \ldots \ q_T$”

The parameters are “learned” from known examples (“labeled data”).

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 HHHLLHLL\ldots, where is the TM region?
- What state emitted the symbol $O_t$?
  
  Example: is the isoleucine at position 32 localized to the membrane?
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*