Local Multiple Alignment

Ungapped motifs:
- Position Specific Scoring Matrices (PSSMs)
  - Modeling, Recognition
- Gibbs sampler
  - Discovery

Gapped motifs
- 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
- It has been shown 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.

Problems with PSSMs
Do not capture positional dependencies

<table>
<thead>
<tr>
<th>WEIRD</th>
<th>D</th>
<th>1.00</th>
<th>0.60</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEIRD</td>
<td>E</td>
<td>0.6</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>0.6</td>
<td>0.40</td>
</tr>
<tr>
<td>WEIQH</td>
<td>I</td>
<td>1.00</td>
<td>0.40</td>
</tr>
<tr>
<td>WEIRD</td>
<td>Q</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>WEIQH</td>
<td>R</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>5.0</td>
<td>2.7</td>
</tr>
</tbody>
</table>

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

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
Hard to recognize pattern instances that contain indels

<table>
<thead>
<tr>
<th>WEIRD</th>
<th>D</th>
<th>0.8</th>
<th>0.8</th>
<th>0.8</th>
<th>0.8</th>
<th>2.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEIRD</td>
<td>E</td>
<td>0.6</td>
<td>2.9</td>
<td>0.6</td>
<td>0.6</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>WEIQH</td>
<td>I</td>
<td>0.8</td>
<td>0.8</td>
<td>3.1</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>WEIRD</td>
<td>Q</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>2.1</td>
<td>1.1</td>
</tr>
<tr>
<td>WEIQH</td>
<td>R</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>2.8</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>5.0</td>
<td>2.7</td>
<td>2.7</td>
<td>2.7</td>
<td>1.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>W E T I R D</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0 + 2.9 + 1.2 + 1.4 + 1.5 = 11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>W E T I R D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2 + 1.8 + 3.1 + 3.0 + 3.4 = 12.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>W E T I R D</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0 + 2.9 + 3.1 + 3.0 + 3.4 = 18.4</td>
</tr>
</tbody>
</table>
Problems with PSSMs

Variable length motifs

- `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?

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)

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

Plan

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

Markov chains

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

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.

MLVKRFENAKCE... → HHHLLLHHLHLHHL...
Questions to ask:
which subsequences correspond to transmembrane regions?

HHHLLHLHHLHL...

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

HHHLLHLHHLHL...

A Markov chain for recognizing transmembrane sequences

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

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

$P(HHLHH) = 0.7 \times 0.7 \times 0.3 \times 0.7 \times 0.7 = 0.072$
Is it a transmembrane protein?

Problem: need a threshold,
threshold must be length dependent
Transmembrane model:

Null model:

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

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

$$P(HHLHH \mid TM) = 0.7 \times 0.3 \times 0.7 \times 0.7 = 0.072$$

$$P(HHLHH \mid EC) = 0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 = 0.031$$

= 2.3

Transmembrane model:

Null model:

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

How are transition probabilities determined?

From known transmembrane sequences

Transmembrane model:

$$a_{ij} = \frac{A_{ij}}{\sum_i A_{ij}}$$

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

Transmembrane model:

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

$$A_{Hi} = \# \text{ of } H^* \text{ pairs}$$
Transmembrane model:

\[
\begin{align*}
\pi(H) &= 0.7, \pi(L) = 0.3 \\
\text{HHALLAHHLLLALHAHHHLHLLLALHAHHHL} \\
\text{HH...}
\end{align*}
\]

\(\pi(H)\) = # 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

An HMM is a generative model;
gives probability of generating a particular sequence
Markov Chains

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

HMMs

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

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

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 $q_0 = E_M$, $q_1 = E_I$, $q_2 = E_L$...

Unlike Markov chains, in HMMs the states are hidden.
Given an unlabeled sequence, HHHLLLHLL..., we infer the most probable state sequence to obtain the boundaries

A three state transmembrane HMM:

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*