More Motifs
WMM, log odds scores, Neyman-Pearson, background;
Greedy & EM for motif discovery

Neyman-Pearson

• Given a sample \( x_1, x_2, ..., x_n \) from a distribution \( f(...) | \Theta \) with parameter \( \Theta \), want to test hypothesis \( \Theta = \theta_1 \) vs \( \Theta = \theta_2 \).

• Might as well look at likelihood ratio:
\[
\frac{f(x_1, x_2, ..., x_n | \theta_1)}{f(x_1, x_2, ..., x_n | \theta_2)} > \tau
\]

What’s best WMM?

• Given 20 sequences \( s_1, s_2, ..., s_k \) of length 8, assumed to be generated at random according to a WMM defined by \( 8 \times (4-1) \) parameters \( \theta \), what’s the best \( \theta \)?

• E.g., what MLE for \( \theta \) given data \( s_1, s_2, ..., s_k \)?

• Answer: count frequencies per position.

Weight Matrix Models

8 Sequences:

<table>
<thead>
<tr>
<th>Freq.</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATG</td>
<td>A .625</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ATG</td>
<td>C 0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ATG</td>
<td>G .250</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ATG</td>
<td>T .125</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LLR</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATG</td>
<td>A 1.32</td>
<td>-\infty</td>
<td>-\infty</td>
</tr>
<tr>
<td>GTG</td>
<td>C -\infty</td>
<td>-\infty</td>
<td>-\infty</td>
</tr>
<tr>
<td>TTG</td>
<td>G 0</td>
<td>-\infty</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td>T -1.00</td>
<td>2.00</td>
<td>-\infty</td>
</tr>
</tbody>
</table>

Log-Likelihood Ratio:

\[
\log_2 \frac{f_{x_1,i}}{f_{x_i}} , \ f_{x_i} = \frac{1}{4}
\]
Non-uniform Background

- *E. coli* - DNA approximately 25% A, C, G, T
- *M. jannaschi* - 68% A-T, 32% G-C

LLR from previous example, assuming e.g., G in col 3 is 8 x more likely via WMM than background, so \((\log_2)^2\) score = 3 (bits).

<table>
<thead>
<tr>
<th>LLR</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.74</td>
<td>-∞</td>
<td>-∞</td>
</tr>
<tr>
<td>C</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
</tr>
<tr>
<td>G</td>
<td>1.00</td>
<td>-∞</td>
<td>3.00</td>
</tr>
<tr>
<td>T</td>
<td>-1.58</td>
<td>1.42</td>
<td>-∞</td>
</tr>
</tbody>
</table>

For WMM, you can show (based on the assumption of independence between columns), that:

\[
H(P||Q) = \sum_i H(P_i||Q_i)
\]

where \(P_i\) and \(Q_i\) are the WMM/background distributions for column \(i\).

WMM: How “Informative”?
Mean score of site vs bkg?

- For any fixed length sequence \(x\), let
  \(P(x) =\) Prob. of \(x\) according to WMM
  \(Q(x) =\) Prob. of \(x\) according to background

- Recall Relative Entropy:
  \[
  H(P||Q) = \sum_{x \in \Omega} P(x) \log_2 \frac{P(x)}{Q(x)}
  \]

- \(H(P||Q)\) is expected log likelihood score of a sequence randomly chosen from WMM;
  \(-H(Q||P)\) is expected score of Background

WMM Example, cont.

<table>
<thead>
<tr>
<th>Freq.</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.625</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>.250</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>.125</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LLR</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.32</td>
<td>-∞</td>
<td>-∞</td>
</tr>
<tr>
<td>C</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>-∞</td>
<td>2.00</td>
</tr>
<tr>
<td>T</td>
<td>-1.00</td>
<td>2.00</td>
<td>-∞</td>
</tr>
</tbody>
</table>

| RelEnt | 0.70 | 2.00 | 2.00 | 4.70 |

Uniform

<table>
<thead>
<tr>
<th>LLR</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.74</td>
<td>-∞</td>
<td>-∞</td>
</tr>
<tr>
<td>C</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
</tr>
<tr>
<td>G</td>
<td>1.00</td>
<td>-∞</td>
<td>3.00</td>
</tr>
<tr>
<td>T</td>
<td>-1.58</td>
<td>1.42</td>
<td>-∞</td>
</tr>
</tbody>
</table>

| RelEnt | .51  | 1.42 | 3.00 | 4.93 |

Non-uniform
Pseudocounts

- Are the $-\infty$’s a problem?
- Certain that a given residue never occurs in a given position? Then $-\infty$ just right
- Else, it may be a small-sample artifact
- Typical fix: add a pseudocount to each observed count—small constant (e.g., .5, 1)
- Sounds ad hoc; there is a Bayesian justification

How-to Questions

- Given aligned motif instances, build model?
  - Frequency counts (above, maybe with pseudocounts)
- Given a model, find (probable) instances?
  - Scanning, as above
- Given unaligned strings thought to contain a motif, find it? (e.g., upstream regions for co-expressed genes from a microarray experiment)
  - Hard... next few lectures.

Motif Discovery: 3 example approaches

- Greedy search
- Expectation Maximization
- Gibbs sampler

Note: finding a site of max relative entropy in a set of unaligned sequences is NP-hard (Akutsu)

Greedy Best-First Approach [Hertz & Stormo]

Input:
- Sequence $s_1, s_2, ..., s_k$; motif length $I$; “breadth” $d$

Algorithm:
- create singleton set with each length $I$ subsequence of each $s_1, s_2, ..., s_k$
- for each set, add each possible length $I$ subsequence not already present
- compute relative entropy of each
- discard all but $d$ best
- repeat until all have $k$ sequences
Expectation Maximization
[MEME, Bailey & Elkan, 1995]

Input (as above):
- Sequence $s_1, s_2, ..., s_l$: motif length $l$; background model; again assume one instance per sequence (variants possible)

Algorithm: EM
- Visible data: the sequences
- Hidden data: where’s the motif
- Parameters: The WMM

Typical EM algorithm:
- Given parameters $\theta^t$ at $t^{th}$ iteration, use them to estimate where the motif instances are (the hidden variables)
- Use those estimates to re-estimate the parameters $\theta$ to maximize likelihood of observed data, giving $\theta^{t+1}$
- Repeat

Expectation Step
(where are the motif instances?)

$$
\hat{Y}_{i,j} = E(Y_{i,j} | s_i, \theta^t) = P(Y_{i,j} = 1 | s_i, \theta^t)
= P(s_i | Y_{i,j} = 1, \theta^t) P(Y_{i,j} = 1 | \theta^t)
P(s_i | \theta^t)
= c P(s_i | Y_{i,j} = 1, \theta^t)
= \cdots \hat{Y}_{i,j} = 1.
$$

Bayes

1 3 5 7 9 11 ...
Sequence $i$

$\hat{Y}_{i,j}$

Expectation Maximization

Maximization Step
(what is the motif?)

Find $\theta$ maximizing expected value:

$$
Q(\theta | \theta^t) = E_{Y \sim \theta^t}[\log P(s, Y | \theta)]
= E_{Y \sim \theta^t}[\log \prod_{i=1}^{k} P(s_i, Y_{i,j} | \theta)]
= E_{Y \sim \theta^t}[\sum_{i=1}^{k} \sum_{j=1}^{l} \log P(s_i, Y_{i,j})]
= E_{Y \sim \theta^t}[\sum_{i=1}^{k} \sum_{j=1}^{l} \log P(s_i | Y_{i,j} = 1, \theta) P(Y_{i,j} = 1 | \theta)]
= \sum_{i=1}^{k} \sum_{j=1}^{l} \log P(s_i | Y_{i,j} = 1, \theta) + C
= \sum_{i=1}^{k} \sum_{j=1}^{l} \hat{Y}_{i,j} \log P(s_i | Y_{i,j} = 1, \theta) + C
$$
**M-Step (cont.)**

\[ Q(\theta \mid \theta^t) = \sum_{i=1}^{k} \sum_{j=1}^{[s_i] - l + 1} \hat{Y}_{i,j} \log P(s_i \mid Y_{i,j} = 1, \theta) + C \]

Exercise: Show this is maximized by “counting” letter frequencies over all possible motif instances, with counts weighted by \( \hat{Y}_{i,j} \), again the “obvious” thing.

---

**Initialization**

1. Try every motif-length substring, and use as initial \( \theta \) a WMM with, say 80% of weight on that sequence, rest uniform
2. Run a few iterations of each
3. Run best few to convergence
   (Having a supercomputer helps)