Motif Discovery

Gibbs Sampling

Position-specific scoring matrix

\[
| (\Theta, \Theta_0) | = y \times L + y = y(\text{L+1})
\]

\[
S_m = x_1, x_2, \ldots, x_L
\]

\[
\mathcal{O} = \log \left( \frac{P(\Theta | \theta)}{P(\Theta_0 | \theta_0)} \right)
\]

\[
\mathcal{O} = \sum_{i=1}^{L} \log (P(X_i | \Theta)) - \sum_{i=1}^{L} \log (P(X_i | \Theta_0))
\]

\[
\mathcal{O} = \log (P_{\text{PWM}}) - \log (P_{\text{background}})
\]
\[
(p_i) = \frac{c_i}{\sum_j c_j}
\]
for all \(i = 1, \ldots, |s_i|-L+1\)

Fix a position \(i^*\) by selecting proportionally to the probabilities \((p_1, p_2, \ldots, p_{|s_i|})\)

\[
\begin{pmatrix}
0.5 & 0.3 & 0.2 \\
p_1 & p_2 & p_3 \\
0.5 & 0.5 & 1
\end{pmatrix}
\]

**Meme:** Expectation Maximization

Every starting position will contribute to both \(\Theta_1\) and \(\Theta_0\)

1. Random guess of starting positions

\[L \leftarrow \text{mohj len} \quad \text{across all sequence } s_i\]

\[|s_i| = n\]

\[s_1, \ldots, s_k\]

It's sequence

\[L \leftarrow \text{mohj len} \quad \text{across all sequence } s_i\]

\[\sum_{\text{starting positions}} (n-L+1)k = s\]

Total position to scan for the initial motif

\[(n-L+1)k = s\]
EM iteration

1) Given $\lambda, \theta, \theta_0$

Re-estimate the starting probabilities for each position in each sequence:

$$
\pi_i = \frac{\lambda \frac{1}{\theta} (X_i | \theta)}{\lambda \frac{1}{\theta} (X_i | \theta) + (1-\lambda) \frac{1}{\theta_0} (X_i | \theta_0)}
$$

$$
\ell = \frac{\sum_{i=1}^{L} \pi_i (X_i | \theta)}{L}
$$

2) Maximization step: Given $\pi_i$ for all $i$ and for all $\ell$

Re-estimate $\lambda, \theta, \theta_0$

$$
\lambda = \frac{1}{N} \sum_{i=1}^{N} \pi_i
$$

$$
\pi_i = \frac{\frac{1}{\theta} (X_i | \theta_i)}{\frac{1}{\theta} (X_i | \theta) + (1-\lambda) \frac{1}{\theta_0} (X_i | \theta_0)}
$$
$$\sum = 5$$

Normalize to obtain new $\theta$, $\theta_0$

$$\text{Stop: if } (6 - \theta_0)^2 < \varepsilon$$

Suffix Trees:

$$((l, d), \text{motif})$$

motif length

mismatch count

$$(3, 1)$$

$$(l, d) \text{ motif with frequency } \geq q$$

Quorum support frequency

$q = 1$

$q = 0.5 \times k$

$\varepsilon = 0.5 \times k$

$$l = 3$$

$$l, 0 \leftarrow \text{ motifs}$$
$$(y, 0) \leftarrow \text{modify}$$

$$\text{exact} \iff d = 0 \Rightarrow$$

$$s_1: \begin{array}{c} \text{ACAGTA} \#
\text{ACAGTC} \#
\end{array}$$

$$s_2: \begin{array}{c} \text{ACAGTC} \#
\text{ACAGTA} \#
\end{array}$$

Generalized suffix tree (Sid, pers)

exact match
$$(3, 0) \rightarrow$$

$$q = 2$$

$$\text{AGT} \rightarrow (1, 3)(2, 4)$$

$$(1, 0) \rightarrow$$

$$(3, 1) \rightarrow$$

$$d = 1 \quad \text{allow up to 1 mismatch}$$

$$A \ll \ll \ll l$$

$$A = \{ A, C, G, T \}$$

$$A(A) = \{ \frac{A\{G\}}{A}, \frac{AG}{1}, \frac{CA}{1}, \frac{TA}{1} \}$$
\[ l = 100 \]
\[ d = 5 \]

\[ \text{AG} = \{ \text{AGC, ACG, CAG, TAG} \} \]

\[ \text{AG} \text{ is a (3,1) motif with } l = 2 \]

Graph-based approach\[(l, d) \text{ motif, quorum-} \]

\[ k - \text{path} \text{ graph} \]

\[ s_1, \text{ACTAG} \]
\[ s_2, \text{TCGCG} \]
\[ s_3, \text{ACTCG} \]

Put an edge between two motifs if they differ by at most \(2d\).

\[ (l \geq d) \text{ motif with } q \text{ quorum} \]

\[ \text{clique of size at least } q \]

\[ O(N) \text{ graph} \]

\[ N = (n_1 + n_2 + \ldots + n_k) \]
Exam I syllabus: Open book/nots...

- sequence alignment
  - global
  - local
  - semi-local (gaps only allowed at beginning or end)

 recurrence

$V(i,j) = \max \left\{ V(i-1,j-1) + \text{Score} \right.$

$\left. V(i-1,j) + \text{gap} \right\}$

Exponential time need use

Maximal

Clique finding

Total size

Use right tree

Diagram: Forward DP

Diagram: Backward DP?
2. gap models
   - linear gap
   - affine gap: open, extension
   - arbitrary gaps
     \[ V(g) = g^2 \]

3. Scoring matrix
   \[ \lambda = 0.01 \]
   \[ PM1 \rightarrow given a tree \]

857. similarity clustering
1. FASTA/Blast

2. Suffix Trees / Anvaco

3. $GST$ =

4. $\text{match} : \text{Gill, Mem, } (g,d) \text{ and }$