The goal: find common motifs within co-regulated genes

Why is it important: understanding functionality of proteins and regulatory gene networks; aids in systems biology approach to drug discovery

Why is it hard:
(1) motifs are short (6-8 bp), sometimes degenerate
(2) can contain any set of nucleotides (no specific rules)
(3) their distances from genes vary

Problem statement:
Input: A set of sequences $S^{(1)}, S^{(2)}, \ldots, S^{(m)}$, and an integer $w$.  
Task: For each string $S^{(i)}$, find a substring of length $w$ so that the similarity between them is maximal.

$S^{(1)}$  
$S^{(2)}$  
\vdots  
$S^{(m)}$  

Essentially, need to find starting indices of the substrings - denote them by $a_1, a_2, \ldots, a_n$

So, the substrings are $\{S_{a_1}, S_{a_1+1}, \ldots, S_{a_1+w-1}\}, \{S_{a_2}, S_{a_2+1}, \ldots, S_{a_2+w-1}\}, \ldots, \{S_{a_n}, S_{a_n+1}, \ldots, S_{a_n+w-1}\}$

Solutions:  
(1) Combinatorial (exhaustive search, ILP)  
(2) Probabilistic (EM, Gibbs sampling, SMC)

Probabilistic description of motifs - position weight matrix (PWM)

$S^{(1)}$  
$S^{(2)}$  
\vdots  
$S^{(m)}$  

$\begin{array}{c|ccccc}
   & 1 & 2 & \cdots & w \\
\hline
A & 0.1 & 0.3 & \cdots & 0.1 \\
C & 0.5 & 0.2 & \cdots & 0.7 \\
G & 0.2 & 0.2 & \cdots & 0.4 \\
T & 0.2 & 0.3 & \cdots & 0.1 \\
\end{array}$
The EM approach to motif finding

$q_k(b)$ - the probability of symbol $b$ in the $k^{th}$ position in the motif; clearly, $\{q_k(b)\}$ defines the PWM, $1 < k < w$

$q_0(b)$ - the probability of symbol $b$ outside the motif

$z_{ij}$ - the probability that, in the $i^{th}$ seq., motif starts from $j^{th}$ position.

Initialize: Choose initial $Q = \{q_k(b), q_0(b)\}$ (e.g., uniform)

Repeat: (1) Re-estimate $Z$ from $Q$ (E-step)
(2) Re-estimate $Q$ from $Z$ (M-step)

Termination: Stop when change smaller than $\varepsilon$

**Step (1):** $Z_{ij} = \Pr(a_i = j \mid S^{(i)}, Q) = \frac{\Pr(a_i = j, S^{(i)} \mid Q)}{\Pr(S^{(i)} \mid Q)} = \frac{\sum_{j=1}^{L-w+1} \Pr(S^{(i)}, a_i = j \mid Q) \cdot \Pr(a_i = j \mid Q)}{\sum_{j=1}^{L-w+1} \Pr(S^{(i)}, a_i = j \mid Q) \cdot \Pr(a_i = j \mid Q)}$

All starting positions equally likely $\Rightarrow \Pr(a_i = j \mid Q)$ - uniform

Moreover, $\Pr(S^{(i)} \mid a_i = j, Q) = q_0(s_1^{(i)}) \cdots q_0(s_{j-1}^{(i)}) q_k(s_j^{(i)}) \cdots q_k(s_{j+w-1}^{(i)}) q_0(s_{j+w}^{(i)}) \cdots q_0(s_L^{(i)})$

Note: we may need to re-normalize $\sum_j Z_{ij} = 1$.

**Step (2):** $q_k(b) = \frac{A_k(b)}{\sum_{b \in B} A_k(b)}$, $A_k(b) = \sum_i \sum_{j: S_j^{(i)} = b} Z_{ij}$

Usually converges after a small # of iterations.
Gibbs sampling

A way of sampling from \( p(x_1, \ldots, x_n) \) by iteratively sampling from \( p(x_j | x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \) for each \( j \) - a type of MCMC (Markov Chain Monte Carlo) algo.

Ex.: Say \( p(x, y) \) is wanted. The procedure:

1. Begin with \( y_0 \), draw \( x \) from \( p(x | y = y_0) \) \( \rightarrow x_1 \)
2. Proceed with drawing:
   \[
   x_i \sim p(x_1, \ldots, x_{i-1}, y_i | y_{i+1}, \ldots, y_n) \\
   y_i \sim p(y_i | x = x_i)
   \]
   after a burn-in period, treat \( (x_i, y_i) \sim p(x, y) \)

- Application to motif discovery:

1. Given: \( \mathcal{Y} = \{ S^{(1)}, S^{(2)}, \ldots, S^{(m)} \} \), \( \mathcal{W} \).

2. Initialize: randomly choose \( \{a_1, a_2, \ldots, a_n\} \)

A probabilistic model of the motif given the above choice:

\[
q_k(b) = \frac{1}{m} \sum_{i=1}^{m} I(S_{a_i+k-1}^{(i)} = b), \ 1 \leq k \leq \mathcal{W}, \quad \text{(PWM)}
\]

where \( I(\cdot) \) - indicator function, i.e., \( I(x = a) := \mathbb{1}(x = a) = \begin{cases} 1, & x = a \\ 0, & \text{else} \end{cases} \)

\( q_k(b) \) - the prob. symbol \( b \) is in the \( k \)th position in motif

3. Sampling iterations:

   (a) Remove one sequence, \( S^{(i)} \), from \( \mathcal{Y} \) (round-robin)

   (b) Recalculate \( q_k(b) \) based on \( \mathcal{Y} \setminus S^{(i)} \):

\[
q_k(b) = \frac{1}{m-1} \sum_{j \neq i} I(S_{a_j+k-1}^{(j)} = b), \ 1 \leq k \leq \mathcal{W}.
\]
(c) Find a probabilistic description of $a_i$

For every $w$-long word $S_j^{(i)} S_{j+1}^{(i)} \ldots S_{j+w-1}^{(i)}$ in $S^{(i)}$, evaluate

\[ P_j = \Pr \text{ [word is motif]} = q_1(S_j^{(i)}) \cdot \ldots \cdot q_w(S_{j+w-1}^{(i)}) \]

\[ R_j = \Pr \text{ [word is background]} = q_1(S_j^{(i)}) \cdot \ldots \cdot q_w(S_{j+w-1}^{(i)}) \]

Compute \[ A_j^{(i)} = \frac{P_j / R_j}{\sum_{j=1}^{L-w+1} P_j / R_j} \]

Note: \[ \sum_{j=1}^{L-w+1} A_j^{(i)} = 1 \]

$a_i$ is described by the distribution parameterized via \{ $A_j^{(i)}$ \} $\Rightarrow$ draw $a_i$ from this distribution!

\[
\begin{array}{cccccc}
& A_1^{(i)} & A_2^{(i)} & \ldots & A_{L-w+1}^{(i)} \\
1 & 2 & 3 & \ldots & L-w+1 & j
\end{array}
\]

(4) Run (3) until convergence

The Gibbs sampler is not much different from EM

Advantages: easier to implement, more robust w.r.t. initial conditions, allows modifications

Disadvantages: less systematic, problems if not all sequences contain a common motif