Motif representation using position weight matrix

Xiaohui Xie
University of California, Irvine
Position weight matrix representation of a motif with width $w$:

$$\theta = \begin{bmatrix}
\theta_{11} & \theta_{12} & \cdots & \theta_{1w} \\
\theta_{21} & \theta_{22} & \cdots & \theta_{2w} \\
\theta_{31} & \theta_{32} & \cdots & \theta_{3w} \\
\theta_{41} & \theta_{42} & \cdots & \theta_{4w}
\end{bmatrix}$$  

(1)

where each column represents one position of the motif, and is normalized:

$$\sum_{j=1}^{4} \theta_{ij} = 1$$  

(2)

for all $i = 1, 2, \cdots, w$. 
Given the position weight matrix $\theta$, the probability of generating a sequence $S = (S_1, S_2, \cdots, S_w)$ from $\theta$ is

$$P(S|\theta) = \prod_{i=1}^{w} P(S_i|\theta_i)$$  \hspace{1cm} (3)$$

$$= \prod_{i=1}^{w} \theta_{i,S_i}$$  \hspace{1cm} (4)$$

For convenience, we have converted $S$ from a string of $\{A, C, G, T\}$ to a string of $\{1, 2, 3, 4\}$. 
Likelihood

Suppose we observe not just one, but a set of sequences $S_1, S_2, \cdots, S_n$. Assume each of them is generated independently from $\theta$. Then, the likelihood for observing these $n$ sequences is

$$P(S_1, S_2, \cdots, S_n | \theta) = \prod_{k=1}^{n} P(S_k | \theta)$$

$$= \prod_{k=1}^{n} \prod_{i=1}^{w} \theta_{i,S_{ki}}$$
Parameter estimation

Now suppose we do not know $\theta$. How to estimate it from the observed sequence data $S_1, S_2, \cdots, S_n$?

One solution: calculate the likelihood of observing the provided $n$ sequences for different values of $\theta$,

$$L(\theta) = P(S_1, S_2, \cdots, S_n | \theta) = \prod_{k=1}^{n} \prod_{i=1}^{w} \theta_{i,S_{ki}}$$ \hspace{1cm} (7)

Pick the one with the largest likelihood, that is, to find $\theta^*$ that

$$\max_{\theta} P(S_1, S_2, \cdots, S_n | \theta)$$ \hspace{1cm} (8)
Estimating $\theta$ using maximum likelihood

The optimal $\theta^*$ can be derived by setting

$$\frac{\partial \log L(\theta)}{\theta_{ij}} = 0$$

subject to the normalization constraint.

The maximum likelihood estimate is

$$\theta_{ij} = \frac{n_{ij}}{n}$$

which is simply the frequency of different letters at each position. ($n_{ij}$ is the number of letter $j$ at position $i$).
Mixture of sequences

Suppose we have a more difficult situation. Among the set of $n$ given sequences, $S_1, S_2, \cdots, S_n$, some of them are generated by a weight matrix $\theta$, but some of them are not. How to identify $\theta$ in this case?

Let us first define the "non-motif" (also called background) sequence. Suppose they are generated from a single distribution

$$p^0 = (p_A^0, p_C^0, p_G^0, p_T^0) = (p_1^0, p_2^0, p_3^0, p_4^0)$$

(11)
Now the problem is we do not know which sequence is generated from the motif ($\theta$) and which one is generated from the background model ($\theta^0$).

Suppose we are provided with such label information:

$$z_i = \begin{cases} 
1 & \text{if } S_i \text{ is generated by } \theta \\
0 & \text{if } S_i \text{ is generated by } \theta^0 
\end{cases}$$

for all $i = 1, 2, \cdots, n$.

Then, the likelihood of observing the $n$ sequences

$$P(S_1, S_2, \cdots, S_n | z, \theta, \theta^0) = \prod_{i=1}^{n} [z_i P(S_i | \theta) + (1 - z_i) P(S_i | \theta^0)]$$