Motif representation using position weight matrix

Xiaohui Xie

University of California, Irvine

Position weight matrix

- Position weight matrix representation of a motif with width w:

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{21} & \cdots & \theta_{w1} \\ \theta_{12} & \theta_{22} & \cdots & \theta_{w2} \\ \theta_{13} & \theta_{23} & \cdots & \theta_{w3} \\ \theta_{14} & \theta_{24} & \cdots & \theta_{w4} \end{bmatrix} \tag{1}$$

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

$$\sum_{j=1}^{4} \theta_{ij} = 1 \tag{2}$$

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

Likelihood

- Given the position weight matrix θ, the probability of generating a sequence $S = (S_1, S_2, \cdots, S_w)$ from θ is

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

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

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 θ. Then, the likelihood for observing these n sequences is

$$
\begin{aligned}
P(S_1, S_2, \cdots, S_n|\theta) &= \prod_{k=1}^{n} P(S_k|\theta) && (5) \\
&= \prod_{k=1}^{n} \prod_{i=1}^{w} \theta_{i,S_{ki}} && (6)
\end{aligned}
$$

Parameter estimation

- Now suppose we do not know θ. 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 θ,

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

Pick the one with the largest likelihood, that is, to find θ^* that

$$\max_\theta \; P(S_1, S_2, \cdots, S_n|\theta) \tag{8}$$

Estimating θ using maximum likelihood

- The optimal θ^* can be derived by setting

$$\frac{\partial \log L(\theta)}{\theta_{ij}} = 0 \qquad (9)$$

subject to the normalization constraint.

- The maximum likelihood estimate is

$$\theta_{ij} = \frac{n_{ij}}{n} \qquad (10)$$

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 θ, but some of them are not. How to identify θ in this case?
- Let us first define the "*non-motif*" (also called *background*) sequence. Suppose they are generated from a single distribution

$$p^0 = (p_A^0, p_C^0, p_G^0, p_T^0) = (p_1^0, p_2^0, p_3^0, p_4^0) \tag{11}$$

Likelihood for mixture of sequences

- Now the problem is we do not know which sequence is generated from the motif (θ) and which one is generated from the background model (θ^0).
- Suppose we are provided with such label information:

$$z_i = \begin{cases} 1 & \text{if } S_i \text{ is generated by } \theta \\ 0 & \text{if } S_i \text{ is generated by } \theta^0 \end{cases} \quad (12)$$

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

- Then, the likelihood of observing the n sequences

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