Position weight matrix and proteome scanning

Position weight matrices (PWMs) are statistical models for representing sequence motifs. They are real valued $m \times n$ matrices, where m is the size of alphabet (20 amino acids for protein sequences) and n is the motif length. PWMs contain a weight for each alphabet symbol i at each position j in the motif. Weight can be described as a log-odds score of a probabilistic model against a background [Pizzi et al., 2011].

$$M(i,j) = \log \frac{P(i,j)}{B(i)} \tag{1}$$

where $B(i)$ is the background probability of amino acid i in the proteome and $P(i,j)$ is the probability of amino acid i at position j.

$$P(i,j) = \frac{count(i,j)}{N} \tag{2}$$

where $count(i,j)$ is the empirical count of amino acid i at position j and N is the count of all the amino acids at position j. Low information content positions or columns at the edges of PWMs are removed to improve signal of the core motif. The information content of each position in the motif is calculated as [Erill, 2012],

$$IC(j) = \left[-\sum_{i=1}^{m} B(i) \log B(i)\right] - \left[-\sum_{i=1}^{m} P(i,j) \log P(i,j)\right] \tag{3}$$

where $IC(j)$ is the mutual information content of j^{th} position in the motif. Information content ratio is then calculated as,

$$ICR(j) = \frac{IC(j)}{IC_{max}} \tag{4}$$

Amino acid positions on both ends of the motif with $ICR(j) \leq 0.4$ are removed. Trimmed PWMs are used to scan a protein sequence to find matches of the weighted pattern above a threshold score (k). For a protein sequence ($S = s_1 s_2 s_3 ...$) the match score ($W(s)$) of any m amino acid long segment is the sum of individual amino acid weights in the PWM [Pizzi et al., 2011].

$$W(s_j..s_{j+m-1}) = \sum_{j=1}^{m} M(s_j, j) \tag{5}$$

where $M(s_i, j)$ is the log-odds score of amino acid s_i at position j in the PWM. The number of statis-

tically significant matches are controlled by converting match score thresholds to p-values. For a given PWM the relationship between its match scores and p-values is defined such that in the background distribution match scores $W(s) \geq k$ [Pizzi et al., 2011, Wu et al., 2000]. Not all amino acid positions within a motif are significant. For example, in class 1 SH3 binding motif [R/K]xxPxxP, positions 1, 4, and 7 are more significant than others. Amino acid positions with $(IC(j)) \geq 0.5$ within the trimmed PWMs are identified as significant. These significant amino acid positions are used in calculation of disordered region, surface accessibility, and peptide conservation scores.