Standardizing and Simplifying Analysis of Peptide Library Data Supporting Information

Andrew D White, Andrew J Keefe, Qing Shao,
Ann K Nowinski, Kyle Caldwell, Shaoyi Jiang

1 MEME Details

The equation to calculate the pseudocount, a sort of "guess" for the motifs which becomes less important as the amount of data grows, is given below:

$$Q = \min(N, \mathcal{A}) \tag{S1}$$

where Q is the pseudocount, N is the number of sequences, and $\mathcal{A}$ is the size of the alphabet. Here the alphabet size is the number of amino acids (20) plus a gap character and unknown residue character. The equation to update a motif (M-step) is given below:

$$m_{kja} = \frac{\sum_i^N z_{ik}\mathbf{1}_{\left\{a=s_{i(j+k)}\right\}} + \dfrac{Q}{\mathcal{A}}}{\sum_i^N \sum_a^{\mathcal{A}} z_{ik}\mathbf{1}_{\left\{a=s_{i(j+k)}\right\}} + Q} \tag{S2}$$

where m_{kja} is the estimated probability of amino acid a occurring at position j in the kth motif. $j \in [1, w]$. w is the motif width. The number of starting positions is $L - w$, where L is the sequence lengths. $k \in [1, L-w]$. z_{ik} is the estimated probability for the ith sequence to start the motif in the kth position. $\mathbf{1}_{\{x\}}$ is the indicator function, which is 1 if the condition x is true. $s_{i(j+k)}$ is amino acid at the $(j+k)$th position in the ith sequence. The other unknown parameter, z_{ik}, is updated (E-step) according to:

$$z_{ik} = \frac{\sum_{j=k}^{k+w} m_{k(j-k+1)(s_{ij})}}{\sum_{k=1}^{L-w} \sum_{j=k}^{k+w} m_{k(j-k+1)(s_{ij})}} \tag{S3}$$

where $m_{k(j-k+1)(s_{ij})}$ is the estimated probability of the amino acid belonging to the ith sequence at the jth position occurring at the $(j-k+1)$th position in the kth motif. The initial guesses for z_{ij} and m_{kja} are uniform. The background distribution, as mentioned in text, is not updated as described in Bailey[1]. Instead, it is known to be uniform for solid-phase peptide libraries and is constant $\frac{1}{\mathcal{A}}$.

2 Comparison of methods

A comparison of the choice of substitution matrix and clustering methods are given in the tables below. A hamming distance is a substitution matrix where all off-diagonal elements are 1 and the diagonal is 0. This provides none of the chemical similarity information encoded into a BLOSUM substitution matrix. Based on these results, the K-means clustering method was selected and the BLOSUM85 substitution matrix was selected.

Matrix	Agglomerative	K-means
Hamming	193	264
BLOSUM50	276	283
BLOUSM62	279	282
BLOSUM85	280	**283**
BLOSUM90	266	274

Table S1: Comparison of different clustering and substitution matrix types for clustering the SHP2 Dataset. The table entries are the number of peptides which match the clustering done by experts in[2], which contains 331 peptide sequences. The version used in the main text is bolded.

Matrix	Agglomerative	K-means
Hamming	86	151
BLOSUM50	108	101
BLOSUM62	109	102
BLOSUM85	86	**102**
BLOSUM90	93	107

Table S2: Comparison of different clustering and substitution matrix types for clustering the TULA-Pre Dataset. The table entries are the number of peptides which match the clustering done by experts in[3], which contains 151 peptide sequences. The version used in the main text is bolded.

References

[1] Bailey, T. L. Ph.D. thesis, University of California at San Diego, 1995.

[2] Sweeney, M. C.; Wavreille, A.-S. S.; Park, J.; Butchar, J. P.; Tridandapani, S.; Pei, D. *Biochem* **2005**, *44*, 14932–14947.

[3] Chen, X.; Ren, L.; Kim, S.; Carpino, N.; Daniel, J. L.; Kunapuli, S. P.; Tsygankov, A. Y.; Pei, D. *J Biol Chem* **2010**, *285*, 31268–31276.