PSSM (position specific scoring matrix)

The position specific scoring matrix (PSSM) method (also known as profile) which was derived from a set of aligned peptides had been widely used for the representation and identification of sequence motifs. A simple PSSM was generated to predict the peptides binding to a given HLA molecule by the determining the frequency of each residue at each position of the HLA-binder subsets with a given length.

$$m_{i,r} = f_{i,r} = n_{i,r} / N$$

Where i=1-*l* (*l*, the length of peptide in the subset), $n_{i,r}$ is the number of times that amino acid *r* is present at position *i*, and *N* is the total number of peptides in the given subset. The binding potential of a given peptide to a given HLA molecule was determined by its similarity to a set of known HLA-binders:

$$S = \prod_{i=1}^{l} m_{i,r}$$

To facilitate the computation and determination, the counts in the matrix were transformed to log-odds; the summation of log-odds scores was equivalent to multiplying the corresponding frequency ratios.

ARB (the average relative binding matrix)

Another matrix-based method used here was the average relative binding (ARB) matrix method, where the elements in the matrix were the frequency ratio of a given amino acid in the binder dataset and non-binder dataset. We adopted the multiplicative method (mARB) to calculate the binding potential of a given peptide to a given HLA molecule, which could be easily understood as the concept that a given peptide had the potential to be a binder or non-binder, and when the potential to be a binder was larger than the potential to be a non-binder, the peptide would be a binder to the corresponding HLA molecule:

$$S = \prod_{i=1}^{l} m_{i,r} = \prod_{i=1}^{l} (m_{i,r}^{+} / m_{i,r}^{-})$$

Where $m_{i,r}^{+}$ and $m_{i,r}^{-}$ are counts in the PSSM of the binder dataset and the non-binder dataset, respectively.

Because of the imbalance and diversity of the peptide subsets, occurrences of some residues in some positions were zero, which prevented the logarithmic transformation and division. Therefore, the pseudo-counts based on the substitution probabilities were introduced in accordance with the method described by Henikoff, where the substitution probabilities were generated using the amino acid substitution matrix BLOSUM62 and the number of pseudo-counts was $B_c = \sqrt{N}$.