1 TCRen scoring

1.1 Deriving empirical contact scoring matrices

Let $\tilde{n}_{ab}$ denote number of contacts between residues of amino acid a of TCR CDR regions and b of an epitope. We compute the symmetrized matrix $n_{ab} = \frac{1}{2}\left(\tilde{n}_{ab} + \tilde{n}_{ba}\right)$ and marginals $n_a = \sum_b n_{ab}$ and $n = \sum_a n_a$.

We define an observed-to-expected contact frequency matrix as

$$\phi_{ab} = n_{ab}^{obs} / n_{ab}^{exp} = \frac{n_{ab} n}{n_a n_b} \tag{1}$$

and use it to derive scores for amino acid pairs similar to BLOSUM matrix derivation where scores are calculated as $S_{ab} \sim \log \phi_{ab}$.

We decompose ϕ_{ab} into (empirical) baseline interaction score S_a for a given amino acid defined as

$$S_a = \frac{1}{20} \sum_b \log \phi_{ab} \tag{2}$$

and specific interaction score S_{ab} defined based on the formula

$$\log \phi_{ab} = S_{ab} + S_a + S_b \tag{3}$$

effectively decoupling pairwise interactions from raw amino acid frequencies in contacting regions. E.g. by using this correction we account for Glycine residues that are frequent in the central region of CDR3β that is almost always close to an antigen in any TCR:pMHC complex.

In present study we compute a single scoring matrix based on CDR1,2,3 regions of both TCRα and β chains.

1.2 Scoring TCR:pMHC pairs

We score each TCR:pMHC pair having residues $\{a_i\}$ of the TCR and $\{b_i\}$ of the antigen using by subtracting the non-specific binding score from the specific one:

$$S\left(\{a_i\}, \{b_i\}\right) = \sum_{i,j} S_{a_i b_j} \delta_{ij} - \sum_i S_{a_i} - \sum_j S_{b_j} \tag{4}$$

where $\delta_{ij} \in \{0, 1\}$ is 1 for contacting residues and 0 otherwise. We subtract baseline scores in order to account for intrinsic properties of amino acids to bind in a non-specific manner. As four flanking residues from both sides of CDR3 are involved in the Ω-loop formation, are almost never involved in TCR:antigen contacts and are subject to strong bias from V/J segment choice, we would substitute $\{a_i\}$ by $\{a_i\} \,/\, \{a_{1..4}\} \cup \{a_{-4..-1}\}$ for CDR3 regions in our calculations.

1.3 Physical interpretation of scoring

(Experimental, not sure it should be included) The scoring can be decomposed into the following components

$$\begin{aligned} S\left(\{a_i\},\{b_i\}\right) &= \sum_{i,j} S_{a_i b_j}\delta_{ij} \\ &\quad - \sum_i S_{a_i}\left(1-\delta_i\right) - \sum_j S_{b_j}\left(1-\delta_j\right) \\ &\quad - \sum_i S_{a_i}\delta_i - \sum_j S_{b_j}\delta_j \\ &= -\left(E_{bound} - E_{unbound} - E_{bound,nonspec}\right) \\ &= -\left(\Delta E_{bound} - E_{bound,nonspec}\right) = -\Delta\Delta E \end{aligned} \tag{5}$$

where $\delta_i = \max_j \delta_{ij}$ identifying a residue having at least one contact, assuming $E_{unbound,nonspec} = 0$.