General ideas on models used here

Local GLM-based pairwise contact model for TCR:pMHC complexes

Local model doesn't account for CDR and antigen chain conformation, relying only on the residue index l, k in CDR and antigen sequences and their lengths L, K to compute CA atom distances.

CA distances between TCR and antigen residues can be modelled with mean distances when centering CDR/antigen regions ($l \leftarrow l - L/2, k \leftarrow k - K/2$), $R = 0.86$.

Distances between TCR and antigen residues and amino acid types of the residues (a_l, a_k) allow predicting contact residues with high accuracy, $AUC = 0.81$.

Formally the model can be written as:

$$P_C\left(a_l, a_k, l, k, L, K\right) = \left(1 + \exp(\alpha d_{CA}\left(l - L/2, k - K/2\right) + \beta_{a_l, a_k})\right)^{-1}$$

i.e. contact probability P_C is modelled with CA atom distance d_{CA} and amino acid types a_l, a_k using a generalized linear model (GLM) with a binomial link and formula $C \sim d_{CA} + a_l : a_k$. Note that here we actually reduce the number of β_{a_l, a_k} coefficients (and protect from overfitting) by requiring $\beta_{a_l, a_k} = \beta_{a_k, a_l}$.

$$d^M_{CA}\left(l, k, L, K\right) = \langle d_{CA}\left(l - L/2, k - K/2\right)\rangle$$

i.e. distance is modelled by mean of centered distance matrices.

The working model is obtained by another round of binomial-link GLM on d^M_{CA} and β_{a_l, a_k} variables: $C \sim d^M_{CA} + \beta_{a_l, a_k}$.

Global model

Global model first predicts CDR x^C_i and antigen x^A_i coordinates in some pre-specified coordinate set.

- CDR coordinate prediction is done using its amino acid sequence as follows:
- Entire PDB is scanned for k-mers similar to CDR loops with e.g. $RMSD = 1.5A$
- Obtained profiles are clustered.
- A amino acid sequence PWM is built for each cluster and used to predict the most likely x^C_i for an unknown CDR
- Antigens are treated in similar way, although here we can get more data by looking at pMHC complexes.

A series of rotations and transitions with pre-fitted parameters (fitting performed on known structural data) are then applied to coordinates $x' \leftarrow \prod_n R_n x$ and d_{CA} are computed from this data. Global model here continues in the same way as local (GLM, etc).