T-Cell Receptor Repertoire Analysis

How to prepare and run the analysis

1. Prepare the required environment
   1. Operating system
      1. Windows 11 with WSL using Ubuntu 22.04.1 LTS
   2. Conda environment
      1. Version 4.13.0
   3. Command-line programs
      1. Clustal Omega
         1. Version 1.2.4
      2. FastTree
         1. Version 2.1.11 SSE3
   4. VS Code
   5. Used to run jupyter notebook TCR\_analysis.ipynb
2. Run analysis using TCR\_analysis.ipynb
   1. Input args
      1. You will need to change these variables appropriately
   2. Output directory
      1. Specify what to name and where to save the resulting files and plots
   3. Process data into fasta file
      1. This will have to be adapted to each use case, however, if the file is already in fasta format then this won’t be needed
   4. Run MSA and create a phylogenic tree
      1. This must be done in the command-line
      2. Use clustalo for MSA to create a fasta aligned file
      3. Use FastTree to create a tree from the aligned fasta
   5. Create and save analysis of the alignment’s pairwise distances between sequences
      1. Heatmap of the distances
      2. Use agglomerative clustering with ward’s metric of the distances to distinguish between the sample types
      3. Visualize ward clustering in a cluster heatmap
      4. PCA plot of the distances

Formal Write-up

We sought to find functional families which might bind the same antigen but differ slightly in sequence between mice infected with TMEV, where about 70% of CD8+ T cells recognize the antigen VP2. We have 4293 samples, of which 1122 were TMEV\_pilot\_CD45, 1651 were TMEV\_pilot\_CD4/CD8, 753 were Inhibitor\_Aak1i, and 767 were Inhibitor\_Ctrl. This is a near prohibitive number of samples for any informative dendrogram with classical methods such as neighbor-joining. However, there are other methods that can quickly approximate and generate phylogenetic trees.

The typical workflow is to first perform multiple sequence alignment for a number of sequences to produce an aligned fasta file. Then, use some sort of agglomerative clustering to create a dendrogram of the distances (or similarities) between the sequences. We first had to create a fasta file of all the TCR sequences that combines the alpha and beta chains into a single string of nucleotides. Then we performed multiple sequence alignment with clustal omega (version 1.2.4 on Ubuntu) with the following command:

./clustalo-1.2.4 -v --full --force --threads 8 --output-order input-order --distmat-out aligned.dm -i input.fa -o aligned.fa

Clustal Omega aligns profile hidden Markov models to each other, rather than using conventional profile alignments to create an aligned fasta file. More information can be found here: <http://www.clustal.org/omega/>.

The aligned fasta file is then used as input into FastTree2 (version 2.1.11 SSE3) to create a phylogenetic tree in the Newick format with the following command:

./FastTree aligned.fa > fast\_tree.nwk

FastTree2 approximates maximum-likelihood phylogenetic trees for either DNA or protein sequences using generalized time-reversible models of nucleotide evolution as well as other models for amino acid evolution. More information can be found here: <http://www.microbesonline.org/fasttree/>.

Using the scaledClonotypeAbundance data, all the TCR sequences appear to be very closely related, therefore, making visualization of over 4000 samples difficult. Since the construction of a phylogenetic tree is the clustering of a distance or identity similarity matrix, we can save this matrix from Clustal Omega and perform visualization using cluster assignment counts, heatmaps, and principal component analysis (PCA), as an alternative to the typical tree visualization.

Using ward distance as our clustering metric, we cluster the sequences into 3 groups and notice that they are almost uniformly distributed across the four types of samples. This suggests that there are distinct attributes that are universally found in the sequences that can be used to distinguish a small subtype (cluster 2) from the other larger 2 clusters for each subtype. This can be visualized with a heatmap that also clusters the samples. It appears that the smallest cluster 2 is very closely related, in contrast to the other two clusters. More analysis is needed to determine the exact similarities and differences between the sequences, of which, the corresponding fasta ID and their cluster assignments have been saved. Using PCA on the similarity matrix and visualizing the top two principal components, there is no possible way to distinguish between any of the four sample groups. These results are comparable to the alignment of the nucleic acids and amino acids.

This analysis is possibly limited, or biased, by concatenating the two alpha and beta chains into one sequence.

Future work is to use this workflow on other sets of TCRs DNA and amino acids.