Predicting T-Cell Receptor Specificity

Problem owner: Dr Eszter N. Tóth
Adaptation/Curation: Dr Christos Maniatis, Dr Stephen Preston

T cells (T lymphocytes) are among the most important immune system cells with a vital role in adaptive immunity. T cells recognise cells in the body infected by viruses, bacteria or cells that have undergone cancer transformation. After recognising the infected or cancerous cells, T cells eliminate them from the body thereby preventing the spread of infection or cancer.

T cells recognise their targets through their T Cell Receptors (TCRs) expressed on their cell membrane. A T Cell Receptor consists of an alpha and a beta subunit. The evolutionary arms race between pathogens and the immune system has resulted in a mechanism for generation of a huge number of unique TCRs: and this is essential for a proper immune response against infections and cancer. Although TCR genes are encoded in the genome, their diversity is massively enhanced in several ways: (i) each TCR is composed of a pair of proteins (either alpha + beta chains or gamma + delta chains); (ii) rather than being encoded as a single gene, the DNA encoding the variable region of each of these chains is formed by joining 3 or 4 different stretches of DNA (gene segments) in a process is called VDJ recombination. Each alpha subunit contains a single V and J segment and each beta subunit contains a single V, a D and a J segment. Diversity is provided by the fact that the genome encodes multiple V, D and J segment; (iii) The joining of these segments involves mechanisms which insert and delete nucleotides in a pseudorandom fashion, maximising diversity in the joining region (the CDR3), the region of the TCR chain which contacts the peptide antigen. (ref 1)

T Cell Receptors (TCRs) constitute one of the most promising classes of emerging therapeutics. Whilst TCRs are amongst the most complex facets of immune biology, engineering of an optimum TCR can transform immunotherapies and personalised medicines. The TCR repertoire at any time point reflects on the person's health and contains a memory of all past experiences. However, CRs are highly variable and their specificities aren't easily predictable with traditional empirical methods.

In this project you will analyse TCR repertoire from the VDJdb (link) and use machine learning to predict TCRs that will bind to specific epitopes.

Tasks

1. Data Download and Preprocessing
   1.1 Download the zip file from GitHub and focus on the VDJdb.txt file.
   1.2 Preprocess the dataset. Figure out what each column represents and keep columns that will help you complete the project.

Predicting TCR specificity from sequence alone is the holy grail of immunotherapy. TCRs that are specific to the same target, often have very similar sequences, thereby TCR sequence – target patterns emerge in the data.

A crude approach could be to represent amino acids of the TCR or key regions of it using one-hot representation.

2. What are the limitations of this approach in downstream analysis? Could you describe a way to overcome them (*Hint: Consider the CDR3 length distribution. We are looking for a high level description of the limitation and an approach that would overcome it. No algorithm development is required.*)

A common method to predict specificity from a sequence is described in Vujovic et.al. (1). It creates some kind of distance or similarity score matrix of TCR sequences and uses that representation to train models that can classify TCRs based on specificity (Fig 1.).

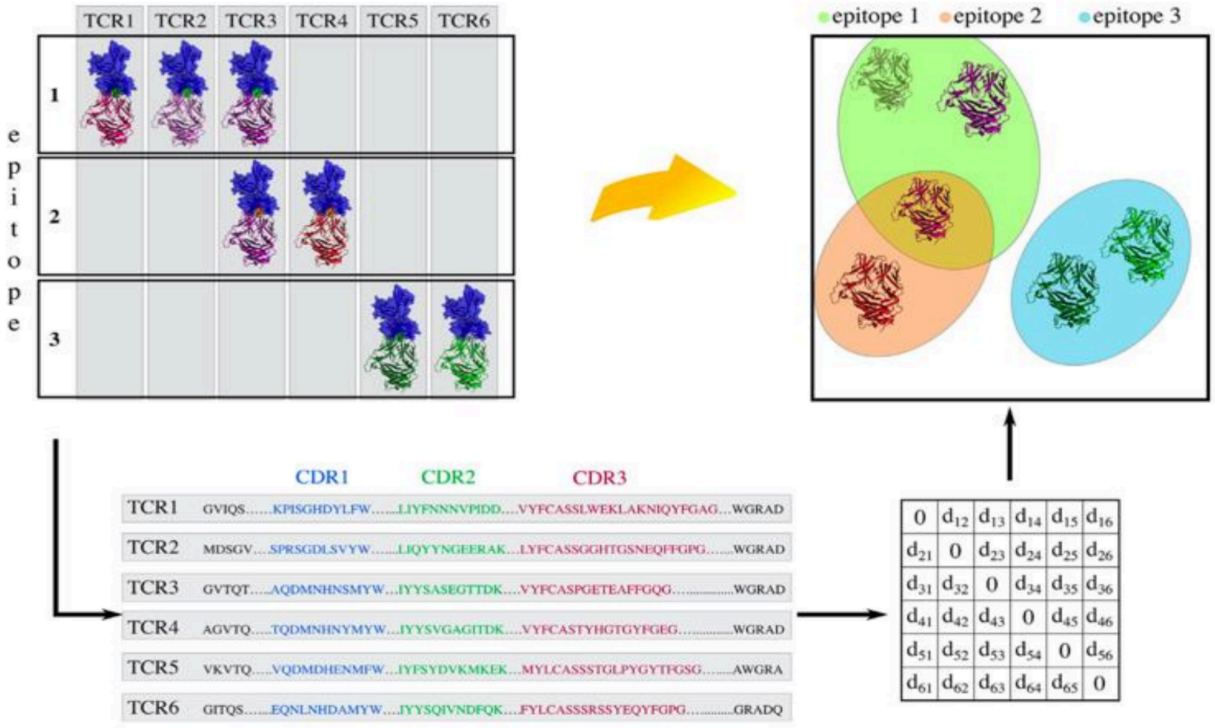

Fig 1. Graphical representation of TCRs to capture specificity. From Vujovic et al (1)

3. Estimate a distance/similarity matrix representation of the data. Calculate these metrics for the alpha and the beta chains separately, then calculate these for the combined alpha and beta chains too. (*Hint: TCRDist, GLIPH or GIANA can be used for this. Alternatively, you can define your own similarity metric.*)
4. Plot the TCRs in 2 dimensions and colour them based on specificity. Compare the plots for the alpha, the beta and the combined alpha-beta chains. Comment on your findings. (*Hint: scikit-learn has a plethora of dimensionality reduction tools. Some examples are PCA, tSNE and UMAP.*)
5. Write code to cluster TCRs. How well do TCRs cluster based on specificity? Can you explain why they do/don't?
6. Write an algorithm that can predict antigen specificity from sequence. You can use any supervised/unsupervised algorithm to predict specificity. Comment on the performance of the model and reason why it performs good or bad. (*Hint: Any reasonable modelling approach is fine. However, keep in mind that simpler models sometimes provide more insights regarding the underlying problem.*)

Bibliography/References

1. Vujovic M, Degn KF, Marin FI, Schaap-Johansen AL, Chain B, Andresen TL, Kaplinsky J, Marcatili P. T cell receptor sequence clustering and antigen specificity. Comput Struct Biotechnol J (2020) 18:2166–2173. doi:10.1016/j.csbj.2020.06.041
2. Mayer-Blackwell. TCR meta-clonotypes for biomarker discovery with tcrdist3: quantification of public, HLA- 2 restricted TCR biomarkers of SARS-CoV-2 infection. bioRxiv (2020) 1:75–94.
3. Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. Nat Biotechnol (2020) 38:1194–1202. doi:10.1038/s41587-020-0505-4
4. Zhang H, Zhan X, Li B. GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation. Nat Commun (2021) 12:1–11.doi:10.1038/s41467-021-25006-7