**1.2 Preprocess the dataset.**

**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) Figure out what each column represents**

By searching for some TCR-related information and reading the journal "T cell receptor sequence clustering and antigen specificity (Vujovic et al., 2020)," we became more familiar with the TCR. Below are descriptions of each column:

1. **complex.id:** A unique identifier for the complex composed of the TCR-peptide and the MHC.

**2. gene:** This value indicates whether the TCR is beta (TRB) or alpha (TRA).

**3. cdr3:** A crucial feature of the TCR is its Complementarity-Determining Region 3, which plays an important role in its ability to recognize antigens.

**4. “v.segm” and “j.segm” :** TCR gene segments include variable segments (V) and joining segments (J) that encode for TCR.

**5. species:** Species from which the TCR sequence was derived, for example, Homo Sapiens might be used.

**6. “mhc.a”, “mhc.b”, and “mhc.class” :** The MHC, which presents the antigen, provides information about that antigen's major histocompatibility complex (MHC).

**7. antigen.epitope:** The epitope (peptide sequence) recognized by the TCR.

**8. “antigen.gene”, “antigen.species”:** An indication of the gene and species from which the antigen has been derived.

**9. “reference.id”:** Reference identifier, often a PubMed ID, for the source of the information.

**10. method:** A description of the method used to determine TCR specificity is provided.

**11. meta:** Information about the cell subset, donor information, etc., contained in the metadata.

**12. cdr3fix:** The sequences of CDR3 have been corrected or confirmed.

**13. vdjdb.score:** Scores indicating how reliable the data provided by TCR are.

**14. “web.method”, “web.method.seq”, “web.cdr3fix.nc”, “web.cdr3fix.unmp”:** The details related to the how the data was processed or validated before being presented on the web.

The dataset can be used for studying the specificity of TCR sequences in response to particular antigens as well as variability in their response. As a result, this has important implications for the study of immune responses and the development of immunotherapies.

**(b) Data Cleaning**

In order to perform the task of predicting TCR specificity using sequence data, we should choose columns that provide information directly related to the sequences of TCR antigens, their targets antigens, and possibly the contexts where these antigens interacted with MHC molecules.

The following operations need to be performed on the data:

1. **Keep Relevant Columns:** In order to accomplish the project's goal and the provided dataset, the columns that are listed below are very important.

**gene:** The purpose of this analysis is to distinguish between the beta chain and the alpha chain of TCRs.

**cdr3:** This is the part of the TCR that is the most variable. It is crucial for the specificity of the antigen.

**“v.segm” and “j.segm”** : A better understanding of TCR diversity and specificity can be gained by studying the characteristics of variable segments (V) and joining segments (J).

**mhc.class:** This is due to the fact that interactions with other classes of MHC molecules might affect the specificity of TCR molecules.

**antigen.epitope:** For an understanding of the specificity of the TCR, this is the sequence of antigens that are recognized by it.

2. **Remove Unnecessary Columns:** Columns that are not directly related to predicting TCR specificity based on sequence information, such as `complex.id`, `species`, `mhc.a`, `mhc.b`, `antigen.gene`, `antigen.species`, `reference.id`, metadata columns like `method`, `meta`, `cdr3fix`, and web-related columns, can be excluded from further analysis.

**(c) 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.**