Task5

// It may be similar to the evaluation indicators of tasks 3 and 4. If it is repeated, just delete it.

1. Data Acquisition and Preprocessing:

The Python script first loads the comprehensive dataset of TCR sequences from an Excel file. Preprocessing consists of selecting relevant columns (gene, cdr3, v.segm, j.segm) and generating a unique identifier (clone\_id) based on the complex.id of each TCR. This identifier distinguishes unique TCR clones and populations, which is critical for subsequent cluster analysis.

2. Clustering Approach:

In this project, we utilize the TCRrep class from the tcrdist library for TCR clustering analysis. The TCRrep[1] [tcrdist3 — tcrdist3 0.1.0 documentation](https://tcrdist3.readthedocs.io/) class is specifically designed to analyze and compare TCR repertoires, with its core functionality being the computation of biochemical distances between TCR sequences, allowing for clustering based on their functional characteristics.

Below, we detail the clustering process:

2.1 Initialization of TCRrep Object:

First we define a function called calculate\_distance\_matrix\_and\_cluster, which is used to calculate the distance matrix and perform cluster analysis. The function first checks whether the given subset of data is empty, and if so, skips subsequent calculations. Next, we create a TCRrep instance, which requires a DataFrame containing the TCR sequence data. The data frame must contain specific columns, such as the amino acid sequence of the CDR3 region, and other relevant information that defines the TCR variable region. In this project, the CDR3 region received special attention due to its central role in TCR recognition.

2.2 Distance Matrix Computation:

TCRrep employs a predefined biochemical distance metric to calculate the distance between each pair of TCRs. This metric is based on the amino acid composition of the CDR3 region and its physicochemical properties. A distance matrix was calculated based on the TCR chain selected and whether a CDR3 sequence was selected. If alpha and beta strands are selected, and a CDR3 sequence is selected, the CDR3 sequence distances of the two strands are added. Otherwise, the corresponding distance matrix is calculated based on the selected chain and CDR3 sequence. The distance matrix is generated by an in-house algorithm that takes into account similarities and differences between TCRs. This matrix provides basic data for subsequent cluster analysis.

2.3 Clustering Algorithm Execution:

We employ hierarchical clustering, spectral clustering, and density-based clustering to capture relationships among TCR sequences. Hierarchical clustering builds a dendrogram illustrating cluster composition. Spectral clustering reduces dimensionality to detect non-globular clusters. DBSCAN identifies clusters based on density, robust to outliers. Integrating results from these methods offers a comprehensive view of TCR specificity. This approach accommodates TCR sequence complexity, enhancing analysis reliability and depth.



Figure For the human sample data set human\_23, the β chain is analyzed, the distance matrix is calculated using the CDR3 sequence, and a spectral clustering image is generated. Based on the comprehensive clustering evaluation indicators, the clustering effect at this time is relatively better.

3. Clustering result analysis and biological conclusions

Upon applying hierarchical, spectral, and density-based clustering algorithms to the TCR data, it is crucial to quantitatively evaluate the quality of the resulting clusters. For this purpose, we have employed three well-established metrics: the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index.

// It may be similar to the evaluation indicators of tasks 3 and 4. If it is repeated, just delete it.

Silhouette Score:

The Silhouette Score is a measure of how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, the clustering configuration is appropriate. In the context of TCR clustering, a high Silhouette Score for each TCR sequence would suggest that the sequence is appropriately placed in its cluster with a high degree of specificity.

Calinski-Harabasz Index:

Also known as the Variance Ratio Criterion, the Calinski-Harabasz Index is the ratio of the sum of between-clusters dispersion and of within-cluster dispersion for all clusters. Essentially, a higher Calinski-Harabasz score relates to a model with better-defined clusters. For TCR clustering, this means the model has effectively separated different specificities into distinct groups.

Davies-Bouldin Index:

The Davies-Bouldin Index is a function of the ratio of within-cluster scatter to between-cluster separation. Unlike the previous two metrics, for the Davies-Bouldin Index, the lowest score represents the optimal clustering solution. A lower Davies-Bouldin score indicates that the clusters are well separated and compact, which, for TCR clustering, would imply clear delineation of specificity groups.

By examining these metrics in tandem, we can obtain a comprehensive view of the clustering performance. These scores collectively provide insights into the cohesion and separation of the TCR clusters, informing us about the distinctness of TCR specificities within each cluster. The use of multiple metrics ensures a balanced evaluation, mitigating the bias that might come from relying on a single metric.

For example, in spectral clustering, data points of the same color represent TCR sequences clustered with each other in a reduced dimension space, and these sequences may have similarities in their functional properties. This similarity may manifest itself in their ability to recognize the same pathogen antigens. Therefore, by observing clusters of the same color, we can identify potentially reactive TCR clusters against a specific antigen. This helps reveal how the immune system responds to specific pathogens and provides important clues for research and development of immunotherapy or vaccine design against these pathogens. By identifying these potentially reactive TCR clusters, we can better understand the immune system's immune response to different pathogens, thereby providing more precise and personalized methods for preventing and treating disease.

4. Specificity Analysis

The specificity of the clusters was analyzed by comparing the known antigen specificities of TCRs within the same cluster. Experimental results show that TCRs within clusters exhibit minimal differences in antigen recognition, indicating that the tcrdist method successfully groups TCRs based on their functional properties.

//Delete if there is not enough space

5. Limitations and Potential Improvements:

While the tcrdist provides a robust framework for TCR analysis, the clustering quality heavily depends on the accuracy of the distance metrics and the chosen thresholds. Variations in TCR sequence alignment and the biochemical models used for distance calculations could introduce biases. Future improvements could include integrating more comprehensive biochemical models and exploring alternative clustering algorithms that might capture more complex relationships within the data.