1. Selecting candidate antigens based on editing distance

Vujovic M[1] used the distance matrix of TCR for clustering in their paper. Inspired by this, the first part of our model framework also selects candidate antigens through the distance matrix of TCR.

When we calculate the distance between sequences, we take into account the edit distance, which is the minimum number of times a sequence can be changed into another sequence by adding characters, deleting characters, and modifying characters. For each change in the sequence, three cases are considered. Through calculation, we can get the formula for editing distance as follows:

\*(1)(2)(3)(4) and represent the two sequences compared, respectively, and D[i,j] represents the distance from the first character to the i character in and the distance from the first character to the j character in .

And, since the distance from any string to the empty string is the length of the string itself. We have the following formula for initializing D:

From (1.1) (1.2)(1.3)(1.4)(1.5) we can get the edit distance of the two sequences.

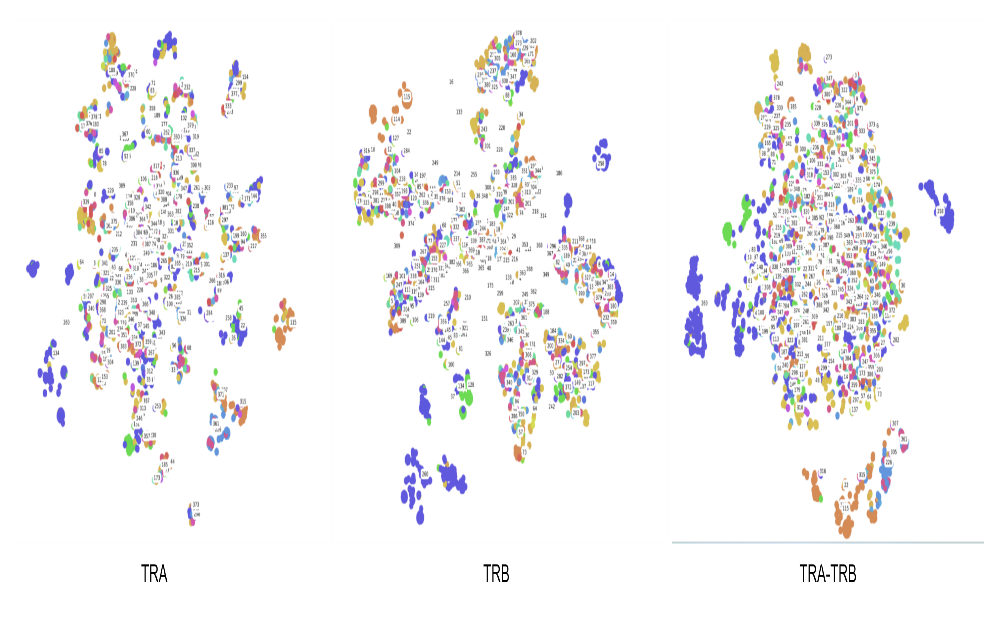
* 1. TCR visualization based on PCA and t-SNE

After calculating TRA, TRB, and TRA-TRB separately, we obtained three different distance matrices based on edit distance. Then we first use PCA to reduce the distance matrix to 50 dimensions and explain the variance as shown in the table:

**Table 1.** PCA explained variance

|  |  |  |  |
| --- | --- | --- | --- |
| Matrix | TRA | TRB | TRA-TRB |
| variance | 0.96041 | 0.965254 | 0.93969 |

It can be seen that after using PCA to reduce dimensionality to 50 dimensions, most of the information was still retained. Then, we used t-SNE for dimensionality reduction, and the results of dimensionality reduction are shown in the following figure:



**Fig. 1.** Visualization after dimensionality reduction

It can be intuitively seen that the clusters of TRB after dimensionality reduction are more prominent, and more TCRs of the same antigen category are assigned to the same cluster.

* 1. Clustering TCR based on editing distance

We can directly use clustering algorithms to cluster the distance matrix and obtain clustering results, as shown in the Table 2,Table 3:

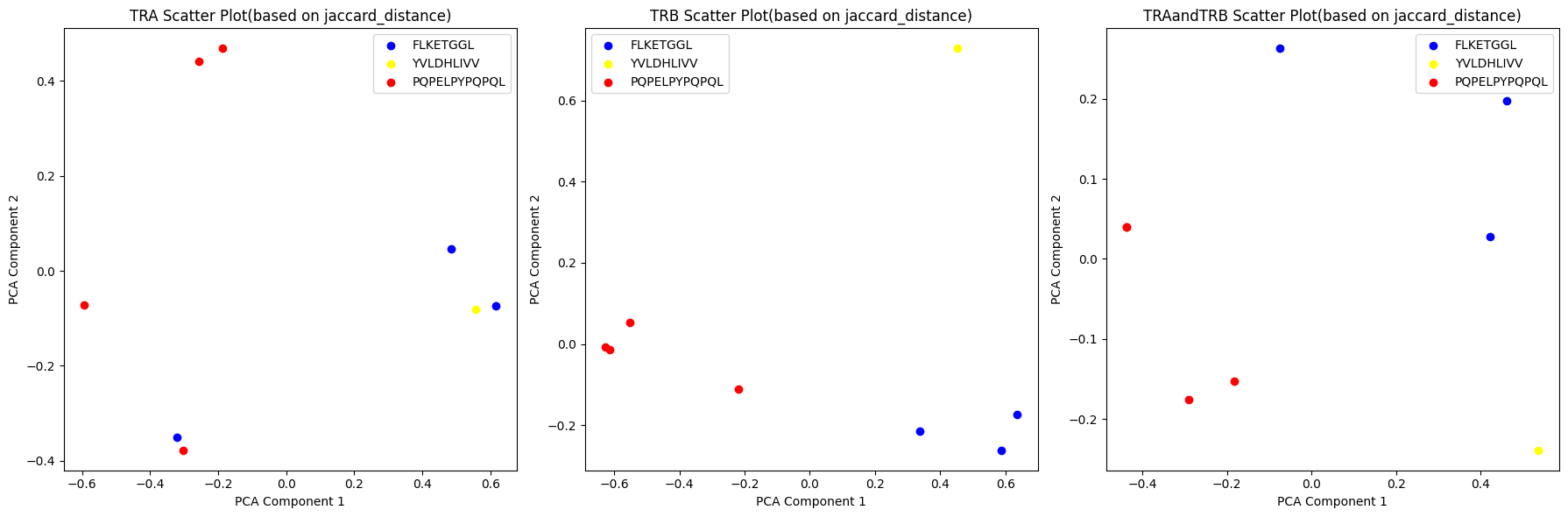
**Table 2.** Cluster performance based on k-means

|  |  |  |  |
| --- | --- | --- | --- |
|  | TRA | TRB | TRA-TRB |
| NMI | 0.01507 | 0.02621 | 0.10711 |
| ARI | 0.112103 | 0.113624 | 0.022367 |

**Table 3.** Cluster performance based on DBSCAN

|  |  |  |  |
| --- | --- | --- | --- |
|  | TRA | TRB | TRA-TRB |
| NMI | 0.41187 | 0.405358 | 0.352910 |
| ARI | 0.01912 | 0.014908 | 0.00949 |

From the table, it can be seen that directly using clustering algorithms has poor clustering performance on the entire distance matrix. This is because TCRs with specificity for similar antigens will be grouped into the same cluster, and there are many such TCRs in the database. If clustering is only performed on TCRs with dissimilar antigens, the effect will be relatively ideal, as shown in the following figure:



**Fig. 2.** Dimensionality reduction display of minority points with dissimilar antigens

For a small number of TCRs with dissimilar antigens, the clustering effect will be very good, but for databases with a large number of TCRs, it is not appropriate to directly use clustering to predict specificity. However, we can still conclude that TCRs with close editing distances have similar antigens.

So the idea of the first part of our model is to generate the target antigen through a distance matrix: calculate the edit distance matrix of the TRB between the target TCR and other TCRs in the training database, calculate the top five closest training data, and use our subsequent classifier for further more accurate discrimination.

1. Reference
2. Vujovic, M., Degn, K. F., Marin, F. I., Schaap-Johansen, A. L., Chain, B., Andresen, T. L., ... & Marcatili, P. (2020). T cell receptor sequence clustering and antigen specificity. *Computational and Structural Biotechnology Journal*, *18*, 2166-2173.