|  |  |  |  |
| --- | --- | --- | --- |
|  | NMI | Purity Fraction | Purity Retention |
| Human, Levenshtein, Hierarchical Clustering | 0.66 | 0.93 | 0.87 |
| Mouse, Levenshtein, Hierarchical Clustering | 0.50 | 0.91 | 0.81 |
| Human, Levenshtein, DBSCAN | 0.37 | 0.52 | 0.12 |
| Mouse, Levenshtein, DBSCAN | 0.49 | 0.62 | 0.20 |
| Human, TCRDist, Hierarchical Clustering | 0.63 | 0.91 | 0.80 |
| Mouse, TCRDist, Hierarchical Clustering | 0.58 | 0.92 | 0.78 |
| Human, TCRDist, DBSCAN | 0.36 | 0.45 | 0.13 |
| Mouse, TCRDist, DBSCAN | 0.39 | 0.64 | 0.22 |

This form can be split into two.

From the clustering results, it is clear that the different ways of calculating the distances have very little effect on the results. With the same species and clustering method, the difference between the two distance calculation methods is below 0.1 for any metric. When using hierarchical clustering, TCRDist produced slightly higher NMIs for both human and mouse data. Results were mixed when using DBSCAN clustering, with TCRDist producing lower purity scores for the human data but slightly higher purity scores for the mouse data. Both TCRDist and Levenshtein distances are widely used tools for TCR distance calculations, although the principles of the two are different, with the former referring to the properties of the CDR3 sequence in relation to different amino acids and the latter focuses on the differences between sequences, they have no significant impact on the task of clustering.

When hierarchical clustering was performed, the clusters of human TCRs performed better than those of mouses in general, but DBSCAN obtained the exact opposite result. This may imply that the human TCR data distribution may have a more pronounced hierarchical structure that is more consistent with hierarchical clustering's assumptions about the data distribution, thus making it easier to generate clusters that conform to the intrinsic organisational structure of the dataset. In contrast, the low performance of DBSCAN is likely due to the algorithm's difficulty in forming coherent clusters with that data distribution. Due to its principle of operation, DBSCAN is difficult to handle data with large differences in density. **[Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density‐based clustering. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(3), 231-240.]** This suggests that the human TCR is likely to have a more dispersed or variable density distribution compared to the mouse data, which is consistent with the previous analysis of the dimensionality reduction results.

Regardless of the distance calculation method for any species, the performance of DBSCAN is much weaker than that of hierarchical clustering, the purity fraction of DBSCAN results may even be half of that of hierarchical clustering, and the purity retention is often only a quarter of that of hierarchical clustering, and the NMI also has a relatively small difference. The purity fraction of DBSCAN results may even be half of that of hierarchical clustering, the Purity Retention is often only a quarter of that of hierarchical clustering, and the NMI has a relatively small gap. Even if we choose the parameters that can make all of metrics reach a relatively high level by comparing the parameters before conducting the experiments, the performance of DBSCAN is still far inferior to that of hierarchical clustering. As demonstrated in the dimensionality reduction section, the data is denser in some regions of the space and sparser in others, with huge gaps in density within different epitopes. it is very difficult for DBSCAN to form effective clusters in regions where the data is too sparse, or where the density varies too much. This problem makes it impossible to find suitable parameters, if the radius is set too large it will lead to too few clustering results with 0 purity, while setting the radius too small will lead to a large number of points being classified as noise. We believe that this is the main reason for the poor performance of DBSCAN in clustering against TCR. On the other hand, hierarchical clustering is robust and can handle datasets with different density of clusters by iteratively merging different levels of clusters. Therefore, it is more suitable for dealing with complex and variable data distributions like TCR. Overall, the results of hierarchical clustering are good, with NMIs of 0.58-0.66 indicating moderate to high similarity between the clustering results and the true labels. This suggests that the model captures the structure of the dataset better, but there may be some mismatches or discrepancies. The purity fraction of 0.91-0.93 and the purity retention of 0.78-0.87 suggest that the clustering results retain most of the structure of the dataset, and that in a very large majority of the clusters the points belong to the same class, while most of the data points of the same class points were able to be grouped into the same clusters.