# 3. 1 Calculate metrics for the alpha and the beta chains separately.

Various distance metrics were represented by using the default parameter pwseqdist >= 0.2.0 of tcrdist3 (Dash et al., 2017), which typically weights CDR3 more heavily as it is in direct contact with antigenic peptides. We can calculate the distance matrix for alpha and beta chain metrics separately as shown in the screenshot below:

表格

描述已自动生成图片包含 日历

描述已自动生成

These two graphs show the distance matrices of the alpha and beta chains for the human TCR alone, the results are similar for other species, only the human results are shown here. We can find that the distance matrices of the TCR for humans as well as other species have their diagonal lines at 0, and each value in each row represents the distance between the TCR and all other rows, respectively, and the distance matrices of the alpha and beta chains are not one-to-one correspondences to be able to compute their binding chains directly. In order to further analyse and calculate their alpha and beta binding chains, we need to use deeptcr's method.

# 3.2 calculate for the combined alpha and beta chains.

Because apes (MacacaMulatta) do not have paired α and β chains, they were not considered in the preparation of the data. We first categorised the paired α and β chains according to the antigenic species, the files were named after the antigenic species, and the paired α and β chains in each file were saved as a tsv file. According to the approach 'DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires' (Sidhom et al., 2021), we use a variational autoencoder (VAE) to take the CDR3 sequences from both the α- and β-chains along with their corresponding V, D, and J gene usage and learn a joint representation of these inputs.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

表格

低可信度描述已自动生成

As shown above, the final feature matrix shows that there's only 26249 rows left. So there are around 1000 rows missing. The reasons can be a little bit complex. I think the main reasons are that some features share same cdr3 representations and v, j segments information or these tcrs may also be found having similar information when the neutral network processed them. Overall, these reasons are just our assumptions and the ture causes need further development.

For other species, the procedures are same with human.

文本

描述已自动生成

For further analysis, we next perform clustering such as Multidimensional Scaling (MDS) to perform cluster analysis to find similar groups of TCR sequences to visualise the distance relationship between sequences.

# Reference:

1. Dash, P. et al. (2017) ‘Quantifiable predictive features define epitope-specific T cell receptor repertoires’, Nature, 547(7661), pp. 89–93. doi:10.1038/nature22383.

2. Sidhom, J.-W. et al. (2021) ‘DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires’, Nature Communications, 12(1). doi:10.1038/s41467-021-21879-w.