Methodology:

\subsection{Gaussian Mixture Models(GMM)}

Gaussian mixture model(GMM) is a parametric probability density function characterized by a weighted sum of multiple Gaussian densities\cite{Gaussian}. The given equation could describe a Gaussian mixture model,

\begin{center}

$p(\mathbf{x}|\lambda) = \sum\_{i=1}^{M} w\_i g(\mathbf{x}|\mu\_i, \Sigma\_i),$\\[1em]

$g(\mathbf{x}|\mu\_i, \Sigma\_i) = \frac{1}{(2\pi)^{D/2}|\Sigma\_i|^{1/2}} \exp\left( -\frac{1}{2} (\mathbf{x} - \mu\_i)^T \Sigma\_i^{-1} (\mathbf{x} - \mu\_i) \right)$

\end{center}

where $\mathbf{x}$ is a D-dimensional vector, $g(\mathbf{x}|\mu\_i, \Sigma\_i)$ are the components Gaussian densities with their own mean vector $\mu\_i$ and covariance matrix $\sigma\_i$ and each Gaussian density is assigned a weight $\omega\_i$.

The parameters of a GMM are estimated by the Expectation-Maximization(EM) algorithm. EM algorithm is an iterative algorithm consisting of two steps: the Expectation step (E-step) and the Maximization step (M-step). In the E-step, this algorithm calculates the expectation of the log-likelihood function of variables, and in the M-step, the algorithm will find the parameters that maximize the expectation computed in the E-step. This process continues until the algorithm converges. In this case,

E-step:

\begin{center}

Calculate $r\_k(\mathbf{x}\_n)=\frac{\omega\_kg(\mathbf{x}|\mu\_, \Sigma\_k)}{\sum\_{j=1}^k\omega\_jg(\mathbf{x}|\mu\_j, \Sigma\_j)}$

\end{center}

M-step:

\begin{center}

$N\_k=\sum\_{n=1}^Nr\_k(\mathbf{x}\_n)$\\

$\mu\_{k}^{new}=\frac{1}{N\_k}\sum\_{n=1}^Nr\_k(\mathbf{x}\_n)\mathbf{x}\_n$\\

$\Sigma\_{k}^{new}=\frac{1}{N\_k}\sum\_{n=1}^Nr\_k(\mathbf{x}\_n)(\mathbf{x}\_n-\mu\_{k}^{new})(\mathbf{x}\_n-\mu\_{k}^{new})^T$\\

$\omega\_{k}^{new}=\frac{N\_k}{N}$

\end{center}

In this report, antigen epitopes are assumed to be in Gaussian distributions or in multimodal distributions where simple clusters like k-means are usually unable to capture varied and complicate features. Moreover, GMM provides high flexibility as it allows clusters with different shapes, sizes, and directions,, which is important when dealing with biological data.

\subsection{UMAP}

UMAP is an algorithm for dimension reduction based on Riemannian geometry and algebraic topology. This algorithm can retain the global structure of datasets while considering the local structure in low-dimensional space\cite{Umap}.

UMAP takes a comprehensive approach to dimension reduction. Before mapping data onto a low dimensional space, UMAP thoroughly learns the data's pattern in high dimension. It starts by using Nearest-Neighbour-Descent to find the data points' nearest neighbors. Then, for each point, it considers their local structures, creating a weighted graph where edges represent the distance between each point. UMAP also incorporates topological data analysis. Once the graph is created, UMAP uses manifold learning methods to capture the geometry and topological structure of the dataset, maintaining the distances between data points. While preserving the data structures, UMAP begins the task of mapping data points onto a low-dimensional space. It identifies a low dimension that best approximates the topological structure of the high-dimensional space. In this process, UMAP defines a cross entropy between the high dimensional and low dimensional spaces and then it minimizes this cross entropy using optimization methods such as stochastic gradient descent.

\begin{center}

$CE = \sum\_{e \in E} w\_l(e) \log\left(\frac{w\_h(e)}{w\_l(e)}\right) + (1 - w\_l(e)) \log\left(\frac{1 - w\_h(e)}{1 - w\_l(e)}\right)$

\end{center}

This is the cross entropy. In this equation, e represents the edges, $w\_h(e)$ are the known weights of edges from high-dimensional manifold approximation, and $w\_l(e)$ are the weights to be discovered for low-dimensional representation.

\subsection{K Neighbors Classifier(knn)}

knn is a simple but effective classification method that keeps all the training data for classification \cite{knn}. The basic idea behind this algorithm is to learn based on the k nearest neighbors of each data point where k is an integer value decided by users\cite{sklearn}. After determining k's value, the algorithm computes the distances between the test points to be classified into those points in known categories. Usually, the distance is calculated using Euclidean distance or Manhattan distance. Based on the computed distance, the algorithm will find the k nearest neighbors to the point from the training dataset. Then, the category with the majority of votes will become the category of the test point.

Results

This report uses Gaussian Mixture Models to cluster TCR sequences based on their specificity. This strategy is applied to human and mouse data separately across alpha and beta chains.

Considering the only known thing is the distance matrix, MDS is applied to this distance matrix to map it to a specific space. MDS can reveal data points' relative positions and yield valuable insights \cite{mds}. Moreover, MDS allows users to use a precomputed distance matrix as input, which is suitable for this case. The output of MDS represents the relative positions of each TCR, which is equivalent to mapping the distance matrix onto a feature space about TCR. However, before applying MDS, like PCA, the number of principal components must be determined to represent critical structures of the dataset while avoiding overfitting and unnecessary complexity. From Kruskal, J. B., $\&$ Wish, M. (1978) \cite{mds}, a method called Scree plot is introduced where in each step 'stress' which is the square root of a normalized "residual sum of squares" is recorded. Like the elbow plot, the turning point is usually chosen as the number of components that best balance the information and complexity.

挑两幅scree plot

In this case, the number of components is all chosen as eight from the plots for both human and mouse. After using MDS to map the distance matrix onto a feature space, GMM is employed to do clustering. From the dimension reduction results discussed in the last chapter, it is clear that some antigen epitopes cluster together. Still, many antigen epitopes are scattered in the space, which means that some TCR sequences may not have a specific boundary, and they are likely to overlap in feature space. Therefore, in contrast to other cluster methods like K-means, GMM allows a data point belonging to multiple clusters, so it is suitable for dealing with fuzzy classification boundaries and overlapping distribution, which is possible for TCR sequences. When implementing GMM, the initial mean state is set to be the mean of each kind of antigen species as their distributions are assumed to be in Gaussian. Moreover, consistent with the number of antigen species decided to do analysis, the numbers of components for humans and mice in this experiment are set to be 8 and 5 individually.

挑几幅图就行

The cluster results show that for both chains and species, antigen epitopes in some clusters show strong clustering patterns, but for some clusters, the data points seem to scatter everywhere. The reasons for such a phenomenon are complicated. Firstly, from a technique perspective, to do visualization, data have to be mapped to a 2-dimensional space, so in this process, a lot of information is lost, which may lead to the condition that data is clustered in high dimensions but scattered in low dimensions. From a biological aspect, the complexity of TCRs may become a reason. Some epitopes may have a diverse array of TCR responses, which could result in a dispersed pattern in each cluster. Then, conservative antigen epitopes usually behave more densely in space, which could lead to good cluster results. In contrast, those with polymorphism or frequent mutations may tend to scatter in the space. Moreover, TCRs exhibit cross-reactivity and can bind to multiple antigens, each with different degrees of affinity, which may also contribute to some scattering phenomena.

Future work

About the model, this report uses Gaussian mixture models and K neighbors to do the clustering and classification. However, the experiment results show that these models could only display some TCR features related to TCRs' differences. TCR is complicated, so it is hard to describe its features using a simple model. Therefore, in the future, to explore TCR with more information, some complicated models can be considered to capture TCR structures. Deep learning is a good direction, and it has already created some excellent works, such as DeepTCR. Moreover, Madi et al.(2017)\cite{networkanalysis} use network analysis, showing that CDR3 sequences with a similar annotation tend to be linked with the same cluster, and their experiments show that this method has good performance, So Network analysis can also become a future study direction.

Reference

\bibitem{Gaussian}

D. A. Reynolds et al., ``Gaussian mixture models,'' \textit{Encyclopedia of Biometrics}, vol. 741, no. 659-663, 2009.

\bibitem{Umap}

Leland McInnes, John Healy, and James Melville. \textit{Umap: Uniform manifold approximation and projection for dimension reduction}. arXiv preprint arXiv:1802.03426, 2018.

\bibitem{knn}

Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer.

\textit{KNN model-based approach in classification}.

In \textit{On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings},

pages 986--996, Springer, 2003.

\bibitem{sklearn}

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.

\textit{Scikit-learn: Machine Learning in Python}.

Journal of Machine Learning Research, 12:2825--2830, 2011.

\bibitem{networkanalysis}

Asaf Madi, Asaf Poran, Eric Shifrut, Shlomit Reich-Zeliger, Erez Greenstein, Irena Zaretsky, Tomer Arnon, Francois Van Laethem, Alfred Singer, Jinghua Lu, et al.

\textit{T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences}.

Elife, 6:e22057, 2017.

\bibitem{mds}

Joseph B. Kruskal and Myron Wish.

\textit{Multidimensional scaling sage university papers series. Quantitative applications in the social sciences}.

Sage Publications, Inc., 1978.