T-Cell Receptor Specificity Prediction Based on Machine Learning

Group 3

abstract.

1 Introduction

In this section, we will introduce the background of the problem, such as the importance of predicting TCR specificity. The literature review will also be included in this section, and we will compare the conclusions, models, effectiveness, and advantages and disadvantages of different scholars on the problem.

2 Descriptive statistics and data preprocessing

In this section, we will conduct a descriptive statistical analysis of the data, such as missing value detection, meaning of key data, and correlation analysis. This descriptive analysis will serve as the basis for our data preprocessing. Next, we will introduce our data preprocessing process.

3 Methodology

This section introduces the method we used and the overall framework of our model. We use the method of editing distance to select 5 candidate antigen sequences, which are then input into our subsequent classification model. Finally, we implement a SeqtoSeq model.

4 Different encoding methods for TCR

4.1 Encoding sequences using One-hot representation

In this section, we will introduce how to use one-hot coding to encode amino acid sequences and its limitations.

4.2 Using additional features to improve One-hot representation

In this section, we use length as an additional feature, which is an improved method of one-hot coding.

4.3 Encoding sequences using word bag model

Next, we explore an encoding method that significantly reduces dimensions and computational complexity: the word bag model. Only considering the quantity of each amino acid in the sequence, ignoring the sequence order information.

4.4 Encoding sequences using N-gram model

Next, we will explore using the N-gram model for encoding, which to some extent preserves order and length information, and reduces computational complexity compared to one-hot coding.

5 Selecting candidate antigens based on editing distance

Next, we will explore the first part of our model: selecting candidate antigens, and we will explore the rationality based on editing distance.

5.1 TCR visualization based on PCA and t-SNE

5.2 Clustering TCR based on editing distance

We use different clustering algorithms to directly cluster the edited distance matrix.

6 Specificity Classifier Based on Random Forest

In this section, we will introduce the construction and results of our classifier. We chose traditional machine learning as the baseline model and used different encoding methods for encoding. Analyze the results. Meanwhile, we use deep learning models for modeling. We will use the model with the best performance and encoding method.

7 Conclusion

In this section, we will summarize our conclusions and insights

8 Further Work and Improvement

In this section, we will focus on better data augmentation methods, such as using GANs to enhance data.

9 References