1. Different encoding methods for TCR

Using only the cdr3 sequence for specificity prediction is the best approach, so we first discuss the encoding method of the cdr3 sequence. A common encoding method is one-hot representation.

* 1. Encoding sequences using One-hot representation

We first counted all types of amino acids in CDR3, and the one-hot representations are as follows:

**Table 1.** One-hot representation mapping table

|  |  |
| --- | --- |
| Amino acids | One-hot representation |
| A | [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| C | [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| D | [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| E | [0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| F | [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| G | [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| H | [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| I | [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| K | [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| L | [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| M | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0] |
| N | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0] |
| P | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0] |
| Q | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0] |
| R | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0] |
| S | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0] |
| T | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0] |
| V | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0] |
| W | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0] |
| Y | [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1] |

After using mapping code encoding, each cdr3 sequence is encoded as a matrix with a size of length \* 20. Due to the non-uniform length, we use 0 padding backward, and after padding, the cdr3 sequence is encoded as a matrix with a maximum length \* 20. The above is the entire process of cdr3 one-hot representation. The encoded cdr3 is shown in the following formula:

We can choose to use one-hot representation for subsequent calculations, but there are many problems with one-hot representation. Firstly, it ignores the length distribution of cdr3. Secondly, it ignores the relationships between different amino acids. Finally, it ignores the local structural information of cdr3. We have the following improvement methods.

* 1. Using additional features to improve One-hot representation

Given the idea that spam classification can improve classification accuracy by adding additional email length information, we can also incorporate the length information of cdr3 as additional input into the encoding process. The encoded cdr3 is shown in the following formula:

* 1. Encoding sequences using word bag model

We can also use the bag of words model to reduce the sparsity of dimensions in single hot encoding. For a CDR3 sequence, we count the number of each amino acid in the sequence and combine it into a 1 \* 20 vector, with each dimension representing the number of that amino acid. Compared to single hot encoding, it significantly reduces the dimensionality, making it easier for some machine learning models to converge. The formula is as follows:

* 1. Encoding sequences using N-gram model

Although the word bag model significantly reduces dimensions, it only considers length information and loses a lot of information about structure and order. We can encode the cdr3 sequence using N-gram, which preserves local sequence information, length information, and flexibly adjusts the dimension of the feature space. The formula is as follows:

Given the excellent performance of N-gram for TCR specificity, we ultimately chose the N-gram model to encode the cdr3 sequence. However, we can provide a comprehensive summary of the effects of different encoding methods on different encoding methods by discussing (1.1), (1.2), (1.3), and (1.4), and provide experimental basis for subsequent researchers.