1. Data Download and Preprocessing
2. Encoding CDR sequences using one-hot representation
3. The limitations of one-hot representation and how to improve them
4. Calculate the distance and similarity matrix of CDR sequences
   1. Edit distance calculation based on Dynamic Programming

When we calculate the distance between sequences, we take into account the edit distance, which is the minimum number of times a sequence can be changed into another sequence by adding characters, deleting characters, and modifying characters. For each change in the sequence, three cases are considered. Through calculation, we can get the formula for editing distance as follows:

\*(1)(2)(3)(4) and represent the two sequences compared, respectively, and D[i,j] represents the distance from the first character to the i character in and the distance from the first character to the j character in .

And, since the distance from any string to the empty string is the length of the string itself (i deletions). We have the following formula for initializing D:

From (1) (2)(3)(4)(5) we can get the edit distance of the two sequences

The formula for calculating similarity by editing distance is as follows:

1. Reduce the dimensionality of CDR sequence and draw TCR as a 2D graph
2. Cluster different TCRs and determine if specificity can be reasonably determined
3. Different algorithms and discussions for predicting specificity based on CDR sequences