Assignment 3

Kastala Murali Krishna
MT19132

Kaggle Username : MT19132
Kaggle Teamname : MT19132

1. In Prediction of B-cell epitopes, I have used two preprocessing methodologies :
   (i) Composition of residue (AAC)
   (ii) Dipeptide (DPC)

In **Composition of residue,** we use the 20 amino acids as data vectors as per their amino acid sequence.
For each training sequence, we will convert them into a vector of 20 length with help of formula :

$$\mathbf{AAC(i) = R(i) / L}$$

Where, R(i) is the number of residue
L is the length of the current seq.

In **Dipeptide**, we take all the possible combinations from amino acids of length "2", i.e we get a vector of size 400. Now for each training sequence, we will convert them into a vector of length 400 with the formula :

$$\mathbf{DPC(i) = D(i) / (L - 1)}$$

where,D(i) is the number of dipeptides.

After Data Preprocessing, Tensorflow Keras Deep Learning model is used to predict To predict the final output. The code is attached.

Results in next page

By using **Composition of residue :**

(i) Model has got 71.51 as accuracy on validation data

(ii) Model has got 70.167 as accuracy on Kaggle Public dataset

By using **Dipeptide** :

(i) Model has got 76.6519 as accuracy on validation data

(ii) Model has got 77.511 as accuracy on Kaggle Public dataset

Both the results are uploaded.