Documentation on epiTCR Paper

Problems:

On using the dataset used by epiTCR for training the model, the result was it was predicting all 1’s for all cdr3b & epitope pairs on lab sample dataset which was wrong.

Why it happened?

We noticed that the train dataset of epiTCR had cdr3b & epitopes which are very unrealistic, thus leading to poor performance on test data.

Solution:

We figured that if we use realistic cdr3b & epitope pairs to train, the model prediction will drastically improve.

Used CDR3B from <https://tcr3d.ibbr.umd.edu/tcrb> stored in cdr3btest

Used peptides from <https://tcr3d.ibbr.umd.edu/class1> stored in epitopetest

Merge both the files (Inner join) on PDB ID stored in merged\_data

All the pairs in the merge files are binding i.e 1

Randomly generated new pairs of cdr3b & epitope using random function. (Assuming that this randomly generated pairs will not be binding i.e 0). Used random.py script to do that.

Removed epitopes that have X in their sequence. Also remove the length of cdr3b >19 and epitope > 11.

Training epiTCR model using:

python3 epiTCR.py --trainfile merged\_data2.csv --testfile testing.csv --chain cem -o result\_merged.csv

The result is stored in result\_merged

For 0.txt:

Using peptide file given and 0.txt to generate all the pairs

Run 0.py script to generate all the combinations of pairs => This generates 0combine.csv

Removed \* & \_ from the cdr3b column

Test file 0combine.csv

Result file 0Result.csv

python3 epiTCR.py --trainfile merged\_data2.csv --testfile 0combine.csv --chain cem -o 0Result.csv