BDMH Assignment-2

-Subhani Shaik (MT18117)

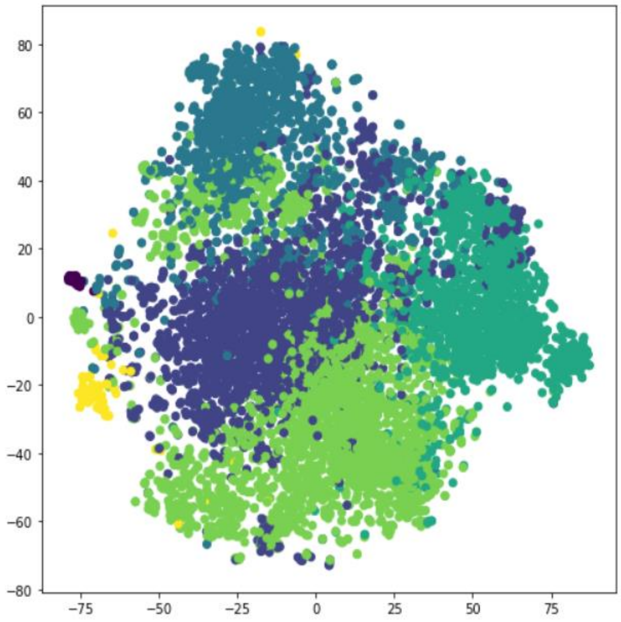
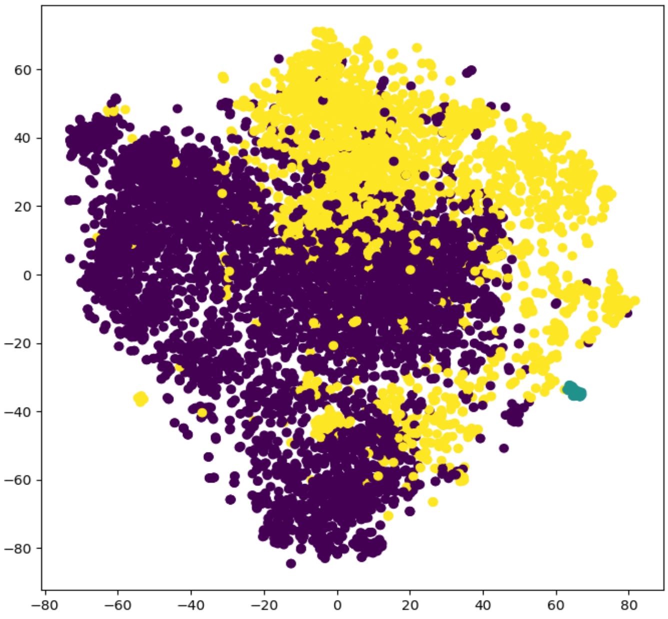
1.

- In this, we used 'Dot plot' to compare the two sequences- a protein sequence and itself and to identify the regions of nucleotide repeats.
- For given sequence, it visualized as a square matrix=nxn. n=51.
    - 51x51 matrix.
- Repeats: count the forward principal diagonals whose length is >=4. That count is gives the #repeats in the given protein sequence.
- Inverse Repeats: count backward diagonals which are normal to the principal diagonal. This count gives us #inverse repeats in the given protein sequence.

2. K-means and Hierarchical Clustering:

Pre-processing: given FASTA formatted sequence are converted into amino acid composition(extracted PFeatures) which given 20 dimensional vector for each sequence.

Clustering: To do optimal K-means and Hierarchial clustering, computed silhouette distance between the points in the clusters and done the clustering. Plotted the TSNE plot to visualize the clusters.

K-Means Clustering	Hierarchical Clustering
	

3. In this, used GPSR docker to compute the potential vaccine candidate of protein sequence by ABCpred, CTLpred, Propred, Toxinpred.

- B-cell epitope using ABCpred.
- CTL epitope using CTLpred.
- HLA-s binders using Propred.
- Toxicity using Toxinpred.

To compute them, used FASTA format protein sequence input file as input to our python script which uses perl scripts in GPSR docker and executes them produces output file.

4. Anti-cancer prediction using machine learning techniques- SVM, ANN , RF methods:

- **Pre-processing:** given FASTA formatted sequence are converted into amino acid composition(extracted PFeatures) which given 20 dimensional vector for each sequence. As we are using supervised learning algorithms, labelled the data points as
    - Label=1 for +ve data points.
    - Label=0 for –ve data points.
- splitted the dataset by 80:20 ratio.
- Done the training using SVM(Support vector machine), Artificial Neural Network(ANN) and Random Forest(RF) classifier.
- Evaluated the model by metrics Accuracy, Recall(Sensitivity), Specificity, MCC(Matthews Correlation Coefficient).

Metric	SVM	ANN	RF
Accuracy	0.941414141414	0.9595959595959	0.9656565656565
Sensitivity(Recall)	0.3809523809523	0.6904761904761	0.5952380952380
Specificity	0.9933774834437	0.9845474613686	1.0
MCC	0.5429351340016	0.7243277821112	0.7574352821772