Assignment Two

Instructions

Student need to create/develop a Docker-based container and main that container at hub.docker.com like https://hub.docker.com/r/raghavagps/gpsrdocker. All work should be done in container, we will evaluate assignment in container. User should submit instructions to use container and scripts so TAs can evaluate assignment. Each question carry 5 marks.

1. Write a python program to compute repeats and inverse repeats in a nucleotide sequence using dotplot algorithm, minimum length of repeat should be 4 nucleotides. Using this script compute repeats and inverse repeats in following sequence "ATGTGTGTCATGCTACGGTCAGGGGTGCATGCTACGTCGTGTCATGTACTG"

2. Write a python script for clustering (K-mean and Hierarchal) protein sequences in FASTA format based on amino acid composition and atomic composition. You may use python libraries like Pfeature and Scikit-learn. Using this script create different type of clusters for all protein sequences in following folder https://webs.iiitd.edu.in/raghava/eslpred2/data/PK7579-data/

3. Write a python script to compute potential vaccine candidate in a protein sequence; i) B-cell epitope using ABCpred, ii) CTL epitope using CTLpred, iii) HLA-2 binders using Propred and iv) Toxicity using Toxinpred. Student may use GPSRdocker for implementation.

4. Write a python program for predicting anticancer peptides using SVM, ANN and Random Forest. In order to develop prediction model student should uses amino acid composition and trained their model on Main Dataset (anticancer and non-anticancer peptide) give at https://webs.iiitd.edu.in/raghava/anticp/datasets.php . You may use Pfeature and Scikit-learn for developing prediction method. Please write performance (Sensitivity, Specificity, Accuracy, MCC) of models developed using SVM, ANN and Random Forest; student may use five-fold cross-validation for evaluation of models.