Big Data Mining in Healthcare
Assignment Report

Team Members:

Ankit Rana	**2018381**
Deepanshu Kaushal	**2018388**
Yo YO Amritpal Singh	**2018379**

Classification of Interacting ,Non-Interacting RNA

Problem:
Classifying RNA into INteracting and Non-Interacting manually is very tedious and time-consuming. Our objective is to create a Machine Learning Model which is accurate and also quick to predict whether an RNA is interacting or not and can do it automatically. We will use evolutionary information as well as the sequence itself to predict these ATP binding sites.

Methodology :
We used Dataset available on Kaggle having. The Dataset contains 311385 unique RNA sequences with their labels as 1 and 0 for non interacting and interacting one. Other than the sequence itself, we have used evolutionary information in the form of Binary profiling as features for our model.
We split the dataset into five parts and have used five-fold cross-validation. All results are means of each fold (unless specifically mentioned otherwise).

We have used many different models to classify the RNA type. We have also tried many different parameters to get the best results.

We have used the following features for prediction:

1) **Mass** - Mass of an amino acid doesn't seem to be really related to ATP binding-site but was actually improving the accuracy. This could be due to various reasons.
2) **Hydrophobicity** - from previous studies, we know that hydrophobicity is a crucial factor when it comes to the binding of proteins and ligands. Proteins can only fold and bind in certain conditions. Using the Fauchere and Pliska scale, we got the hydrophobic calculations.

3) **Polarity** - Charges always play a big role in protein bindings. Different amino acids have different charges are thus more/less attracted to the ATP molecule for binding. We have used the Grantham R polarity scale for each amino acid.
4) **Solvation Potential** - just like hydrophobicity, solvation in a solvent also plays an important role in protein binding. If a site is not fully exposed then the ATP will not have a place to bind. Jones scale was used to find solvation potential.
5) **Net Charge** - Other than the dipole moment, the overall charge of the surface proteins was also used as a feature. The total charge was obtained from Klein et al.
6) **Isoelectric Point** - Isoelectric point defines the pH where the new charge is zero on an amino acid. Since binding of proteins requires the exchange of ions, so the isoelectric point is an important factor.

RoadMap:

1) **Upper Sampling on Train Data**
2) **Cross Validation on Train Data**
3) **Created features matric fo train data**
4) **Found Best parameters for Random Forest Classifier by applying GridCV**
5) **Created Feature matrix for test data**
6) **Predicted labels for Test Data**
7) **Made sample file as output file with labels of Test data**

Accuracy : 0.84999 on Public leaderboard

Steps to run the code:

1) Make sure python code .py file, Train file and Test file are in the same directory
2) Run command - **python3 grp9.py RNA_Train.csv test.csv**
3)