A

Project Report on

Drug-Target Protein-Protein Interaction Network

Submitted in partial fulfilment for the award of

PG DIPLOMA IN BIG DATA ANALYTICS

CENTRE FOR DEVELOPMENT OF ADVANCED COMPUTING (C-DAC)

ADVANCED COMPUTING TRAINING SCHOOL (ACTS)

Knowledge Park, Bengaluru - 560 038

Submitted By:

S. no.	Name of Student	ROLL NO
1.	Vinod Kumar Reddy D	69
2.	Nalli Vennela	32
3.	Vandana Maurya	66
4.	Vemula Sree Mukesh	68

Under the Guidance of:

Mrs. Janaki

CERTIFICATE

This is to certify that, the project report entitled
Drug-Target Protein-Protein Interaction Network

S. no.	Name of Student	ROLL NO
1.	Vinod Kumar Reddy D	69
2.	Nalli Vennela	32
3.	Vandana Maurya	66
4.	Vemula Sree Mukesh	68

is the record of Bonafide work carried out by them in partial fulfilment of the requirement for the award of **PG Diploma in Big Data Analytics** prescribed by **Centre For Development Of Advanced Computing(C-DAC)**

Mrs. Janaki
Project Guide

Mrs. M Savitri
Course Co-Ordinator

Date:

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all the people without whom this project would have been highly impossible.

We would like to devote our first vote of thanks to our **guide- Mrs. Janaki** for her constant support and encouragement and for providing valuable insight for solving our problem statements and solving our doubts regarding **Feature Extraction** in our project. She has a great hand in the firm foundation of this project. We are deeply in debt for her valuable suggestions, scholarly guidance and constructive criticisms along with constant encouragement at each and every step for successful completion of the project.

We would also like to thank our **Project Co-Ordinator, Mrs. M Savitri** for inspiring us towards completion of this project and thank all those who assisted us directly or indirectly for their valuable time and help.

We are especially indebted to our parents for their love, sacrifice, and support. They are our first teachers after we came to this world and have always been mile stones to lead us a disciplined life.

ABSTRACT

There is a need of accurate identification of drug targets as it is a crucial part in biomedical research and drug development programs. Many studies have been conducted on analyzing drug target features for getting a better understanding on principles of their mechanisms. But most of them are based on either strong biological hypotheses or the chemical and physical properties of those targets separately.

According to some special topological structures of the drug targets, there are significant differences between known targets and other proteins. These topological features are helpful to understand how the drug targets work in the PPI network. Particularly, it is an alternative way to predict potential targets or extract nontargets to test a new drug target efficiently and economically.

In this project, we tried to extract, understand and analyze the topological features of drug targets protein in the Protein-Protein Interaction Network. We also tried to use some of the feature to predict whether a given protein will act as a drug target or not using some of the Machine Learning algorithms.

CONTENTS

Chapter 1. INTRODUCTION……...1
1.1. Proteins……...2
1.2. PPI……....3
Chapter 2. DATA COLLECTION……....4
2.1 Database……4
2.1.1 STRING (version 10.0,PPI) ……4
2.1.2 DrugBank……4
2.1.3 UniPort……....4
Chapter3. Network Topology……...…5
3.1 Network Topology……..……..…5
3.1.1 Degree……...…6
3.1.2 Betweenness……...6
3.1.3. Average Distance……..…7
3.1.4. Eccentricity……...…….....7
3.1.5. Modularity……...……..…7
3.1.6 Coreness……...……..8
3.1.7 Cluster Coefficient……..…9
3.1.8 Clique……...……..…9
3.1.9 Closeness Centrality……..10
3.1.10 Stress Centrality…….....10
3.1.11 MNC-Maximum Neighborhood Component……..…11
3.1.12 DMNC - Density of Maximum Neighborhood Component……....11
3.1.13 Radiality Centrality…….....11
Chapter 4. Tools Used……..…13
4.1 Tools Used……..…….……13
4.1.1 Cytoscape……..13
4.1.1.1 Cytohubba API……13
4.1.2 R Language……..13
4.1.3 Python……..…….14
4.1.3.1 Pandas……..…….14
4.1.3.2 NumPy……..…15
4.1.3.3 NetworkX……..……15
Chapter 5. Machine learning……..…..16
5.1 Stratified Sampling…….. 16
5.2 Machine Learning……..…16
5.2.1 Random forest……...…17

5.2.2 Gradient boosting……..18
5.2.3 KNN…….……19
5.2.4 STACKING……...……...20
5.2.5 Gaussian Naive Bayes……..20
5.2.6 XGBoost……21
5.2.7 Multi Layer Perceptron……..21
5.2.8 Grid Search……22
Chapter 6.Methodology……22
6.1Flowchart for Feature extraction in R……23
6.2 Methodology Flowchart……24
Chapter 7.Code…….…..25
7.1R-code for Feature Extraction……25
7.2 Python code……...30
Chapter 8.Results and Discussion…….….38
8.1 Results…….…..38
8.2 Discussion…….…….38
FUTURE DIRECTION..……..……..……39
REFERENCES……..……..…40

List of Figures

Figure No	Figure Name	Page No
1.1	Proteins of different structure	2
1.2	Protein-Protein Interactions	3
3.1	Degree	6
3.2	Betweenness	6
3.3	Cluster Coefficient	9
3.4	Clique	10
5.1	Stratified Random Sampling	16
5.2	Random Forest	17
5.3	Gradient Boosting	18
5.4	K-Nearest Neighbours	19
5.5	Normal Distribution	20
5.6	Naive Bayes	20
5.7	XGBoost	21
5.8	Multi Layer Perceptron	22

List of Flow Charts

1) Flow Chart for feature extraction in R 23

2) Methodology Flow Chart 24