Analysis of Protein Sequencing for Drug Discovery

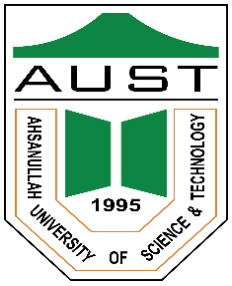
A thesis
Submitted in partial fulfillment of the requirements of the course
CSE 4100: Project & Thesis-I

Submitted By

Hasan Bin Jamal	18.01.04.070
Fatima Juairiah	18.01.04.071
Abu Tarek Rabbi	18.01.04.086
Abdullah Al Mohaimen	18.01.04.098

Supervised By

Dr. S. M. A. Al-Mamun
Professor
Department of Computer Science and Engineering



Ahsanullah University of Science & Technology

Department of Computer Science & Engineering

June, 2022

CANDIDATE'S DECLARATION

We, hereby, declare that the work presented in this report is the outcome of the investigation performed by us under the supervision of Dr. S. M. A. Al-Mamun, Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh. The present work was performed in partial fulfillment of the requirements of the course CSE4100: Project and Thesis-I, in accordance with the course curriculum of the Department for the Bachelor of Science in Computer Science and Engineering program.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualification.

_____

Hasan Bin Jamal

18-01-04-070

_____

Fatima Juairiah

18-01-04-071

_____

Abu Tarek Rabbi

18-01-04-086

_____

Abdullah Al Mohaimin

18-01-04-098

Abstract

Computational biology involves the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to study the biological system, explore insights into Pharmaceutical fields, and advancing medical technology. Here we are implementing Artificial Intelligence in the combined field of Biology and Pharmaceuticals. Our target is to explore some advanced, fast, efficient and reliable intelligent computational techniques required to expedite the Protein classification process for future drug discovery. Machine learning algorithms like Decision Trees, K-means-clustering, k-nearest neighbors will be using as great tools and techniques for the classification of protein molecules for discovering possible drug target proteins. The features, which we are using here -Protein Sequence, protein length, Molecular mass and Gene

CONTENTS

LIST OF FIGURES

Chapter 1
Introduction

1.1: Overview

A larger biomolecule formed from a long chain of amino acids is called as protein. There are mainly 20 amino acid bases which form proteins. Proteins play important functions within the body of organisms like they are main catalysts for the metabolic reactions, they transport molecules from one place to another, they replicate DNA, etc. Proteins are also important structural constituent of cells in organisms. There are four levels of protein structures viz. Primary Structure, Secondary Structure, Tertiary Structure and Quaternary Structure. A chain of amino acids in a polypeptide is often referred as Protein Primary Structure. We will work this primary structure between neighboring peptide bonds. The known proteins have been classified into protein families and super families, based on their functions, structures or sequence similarity. Whenever we come across a new protein molecule, it is necessary to know the family of that molecule and hence to classify it into one of the known families. Lot of large scale experiments and projects have been carried out in the recent past, which has resulted in explosion of biological data especially proteins and DNA. Lot of new and unknown proteins is being found. It becomes extremely difficult to know the families of such large number of proteins by using traditional ways. Hence some advanced, fast, efficient and reliable intelligent computational techniques are required to expedite the classification process. Machine learning algorithms like Bayes classifier, Decision Trees, Support Vector Machines, K-means-clustering, k-nearest neighbors are proving to be great tools and techniques for classification of protein molecules . The important aspect in the problem of protein classification is the selection of protein feature. The feature which we are using here is Protein Sequence, protein length, Molecular mass, Gene name.

1.2: Motivation:

Bangladesh is the only least developed country (LDC) that meets nearly 98 percent of its domestic demand for pharmaceutical products, with a market size of approximately $3 billion. Our indagation is to enrich our Country's Pharmaceutical field And at the wider sense escalating the human welfare by making easier, cost-effective and reliable intelligent computational techniques to expedite the classification process.

The World Health Organization warned in its 2007 report that infectious diseases are emerging at a rate that has not been seen before. Since the 1970s, about 40 infectious diseases have been discovered.

For last two decades, we have seen outbreaks of previously known diseases.

Bacteria, viruses, and other microorganisms can change their protein sequence over time and develop a resistance to the drugs used to treat diseases caused by the pathogens. Therefore, drugs that were effective in the past are no longer useful in controlling disease.

The number of common diseases like diabetics, hypertension, disorder of nervous system has been evolved significantly in previous years.

The number of adults with hypertension increased from 594 million in 1975 to 1.13 billion in 2015.

Globally, an estimated 463 million adults are living with diabetes, according to the latest 2019 data from the International Diabetes Federation.

The development of drugs by understanding the protein sequences and the remarkable eradication of these issues had created hope that diseases could be controlled or even eliminated.

More than 100 genuine and similar number of modified therapeutic proteins are approved for clinical use in the European Union and the USA with 2010 sales of US$108 bln.

Thus rapid research is going in trying to understand more thoroughly how diseases are caused and how the human immune system responds to these diseases as well as more directed research in developing and evaluating protein-based drugs and other tools to prevent malady by these agents.

So major development of drugs are Indispensable.

Thus, it motivated us to get into this field.

1.3 Introduction to Our Work:

Our work is the application of machine learning in the combined field of Bio-informatics and Pharmacology. To be more specific we are working with Computational Biology.

Bio-informatics is a field of science that uses computers, databases, math, and statistics to collect, store, organize, and analyze large amounts of biological, medical, and health information. Information may come from many sources, including genetic and molecular research studies, patient statistics, tissue specimens, clinical trials, and scientific journals.

Pharmacology is a branch of medicine, biology and pharmaceutical sciences concerned with drug or medication action

Computational biology, a branch of biology involving the application of computers and computer science to the understanding and modeling of the structures ,biological systems,

discovering new drugs and exploring the area of medical health and information. It entails the use of computational methods (e.g., algorithms, statistics) for the representation and simulation of biological systems, as well as for the interpretation of experimental data, often on a very large scale.Computational biology is a very broad discipline, in that it seeks to build models for diverse types of experimental data and biological systems and that it uses methods from a wide range of mathematical and computational fields.

We know that protein is an essential element of the body that controls so many thinks in human body.We are going to work with the primary structure of the protein sequence which is build up by the amino acids. We will analyse the sequence and find out some specific type or class of protein that is useful for the discovery of drug target proteins class and finding out whether that particular sequence is drugable or non-drugable.

As we have seen the importance of drugs and drug targets in medical science in recent days as there is so many biological disease out breaks such as COVID-19, EBOLA etc. The progress in medical sector for finding out better prevention and cure knows no bound. Most drug targets are members of families of proteins that are related phylogenetically. Examples include G-protein coupled receptors (GPCRs), protein kinases, nuclear hormone receptors, serine proteases, and ion channels. The degree to which compounds that bind to the desired target also bind to these related proteins varies greatly and depends on the conservation of the protein fold and the sequence homology of the binding site.

We have explained our whole work in four chapters. Chapter 1 contains the introduction, Chapter 2 talks about literature review, Chapter 3 gives a vivid description of our dataset and features . and Chapter 4 holds all the information about implementation process and result analysis.

Chapter 2
Background

2.1 Computational Biology:

Computational biology, a branch of biology involving the application of computers and computer science to the understanding and modeling of the structures ,biological systems, discovering new drugs and exploring the area of medical health and information. It entails the use of computational methods (e.g., algorithms, statistics) for the representation and simulation of biological systems, as well as for the interpretation of experimental data, often on a very large scale.

Computational biology is a very broad discipline, in that it seeks to build models for diverse types of experimental data and biological systems and that it uses methods from a wide range of mathematical and computational fields.

Perhaps the most important task that computational biologists carry out (and that training in computational biology should equip prospective computational biologists to do) is to frame biomedical problems as computational problems. This often means looking at a biological system in a new way, challenging current assumptions or theories about the relationships between parts of the system, or integrating different sources of information to make a more comprehensive model than had been attempted before. In this context, it is worth noting that the primary goal need not be to increase human understanding of the system; even small biological systems can be sufficiently complex that scientists cannot fully comprehend or predict their properties. Thus the goal can be the creation of the model itself; the model should account for as much currently available experimental data as possible. Note that this does not mean that the model has been proven, even if the model makes one or more correct predictions about new experiments. With the exception of very restricted cases, it is not possible to prove that a model is correct, only to disprove it and then improve it by modifying it to incorporate the new results.

This view emphasizes the importance of machine learning for constructing models. In most current machine learning applications, statistical and computational methods are used to construct models from large existing datasets and those models are used to process new data. Examples include learning to classify spam emails, to enable fingerprint access to your phone, and to recognize human speech. However, an increasing number of machine learning applications don't stop learning after their initial training. They can either learn from additional data as it becomes available, or, even choose what additional data they would like to learn from. This last area is termed active machine learning, and it promises to play a very important role in biomedical research in the coming years.

2.2: Domain Knowledge:

2.2.1 Protein Sequence:

Protein sequencing is the practical process of determining the amino acid sequence of all or part of a protein or peptide

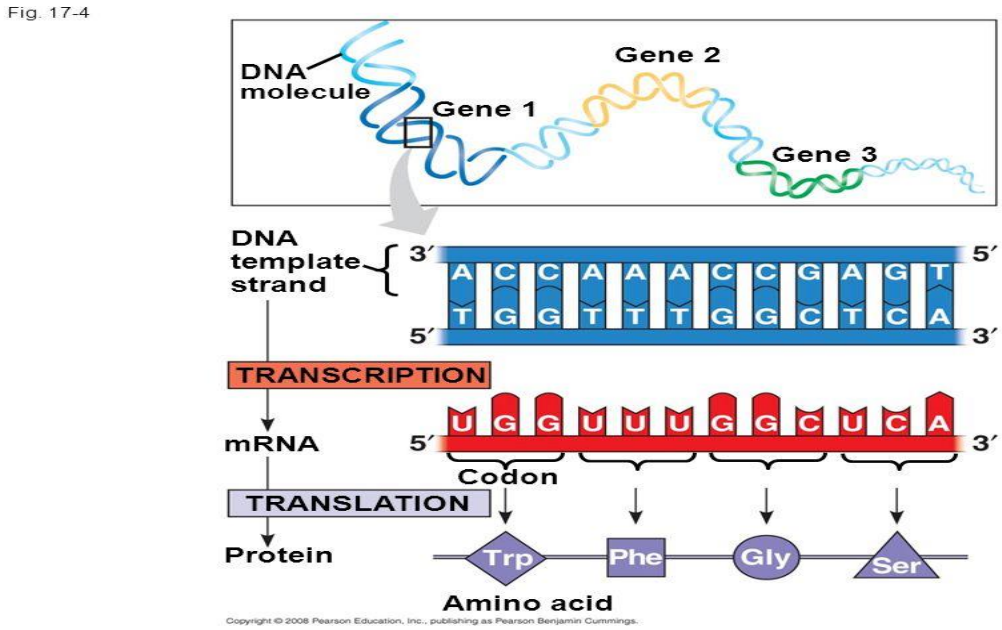


Figure : (2.1) DNA Transcription & Translation[15]

There are 20 types of amino acids that forms protein:

Lysine(lys - **K)**

Lysine is one of the most commonly mentioned essential amino acids. Foods such as bread and rice tend to be low in lysine. For example, compared to an ideal amino acid composition, wheat is low in lysine. The United Nations University carried out the research about people in developing countries where they depend on wheat for protein, and found out the lack of lysine in their diet. Not having enough lysine and other amino acids can lead to serious problems such as stunted growth and severe illness.

Threonine(thr - **T)**

An essential amino acid that is used to make the active site of enzymes.

Phenylalanine(phe - **F)**

An essential amino acid that is used to make many types of useful amines.

Methionine(met - **M)**

An essential amino acid that is used to make many different substances needed in the body.

Histidine(his - **H)**

An essential amino acid that is used to make histamine.

Tryptophan(trp - **W)**

An essential amino acid used to make many types of useful amines.

Glutamine(gln - **Q)**

Glutamine is one of the most common amino acids in the body. Glutamine protects the stomach and gastrointestinal tract. In particular, glutamine is used to produce energy for the gastrointestinal tract. Glutamine promotes the metabolization of alcohol to protect the liver.

Aspartate(asp - **D)**

Aspartate is one of the amino acids that is most usable for energy. Aspartate is one of the amino acids positioned most closely to the tricarboxylic acid (TCA) cycle in the body that produces energy. The TCA cycle is like the engine that powers cars. Each cell in our bodies functions to produce energy.

Glutamate(glu - **E)**

The kombu stock used in Japanese cooking contains glutamate. Glutamate is the base of umami and free glutamates are found in kombu, tomatoes and cheese. Inside the body, glutamate is utilized as an important source of essential amino acids.

Arginine(arg - **R)**

Arginine plays an important role in opening up the veins to enhance blood flow. Nitric oxide that opens up the veins is made from arginine. Arginine is a useful amino acid for removing excess ammonia from the body. Arginine increases immunity.

Alanine(ala - **A)**

Alanine supports function of the liver. Alanine is used to make glucose that are needed by the body. Alanine improves the metabolization of alcohol.

Proline(pro - **P)**

Proline is one of the amino acids contained in collagen that makes up skin tissue. Proline is one of the most important amino acids to the natural moisturizing factor (NMF) that keeps skin moist.

Cysteine(cys - **C)**

Cysteine reduces the amount of black melanin pigmentation made. Cysteine is plentiful in head hair and body hair. Cysteine increases the amount of yellow melanin made instead of black melanin.

Asparagine(asn - **N)**

An amino acid that was discovered from asparagus. Both asparagine and Aspartate are positioned close to the tricarboxylic acid (TCA) cycle that produces energy.

Serine(ser - **S)**

An amino acid used to make phospholipids and glyceric acid.

Glycine(gly - **G)**

A non-essential amino acid that is made in the body. Glycine is plentiful in the body. It acts as a transmitter in the central nervous system and helps regulate body functions such as locomotion and sensory perception. Glycine makes up one-third of collagen.

Tyrosine(tyr - **Y)**

Tyrosine is used to make many types of useful amines. Tyrosine is grouped as an aromatic amino acid together with phenylalanine and tryptophan.

Valine(val - **V), leucine(**leu - **L) and isoleucine(**ile - **I)**

Branched-chain amino acids (BCAAs) are a group of three amino acids (valine, leucine and isoleucine) that have a molecular structure with a branch. BCAAs are plentiful in muscle proteins, stimulate muscle growth in the body and provide energy during exercise.

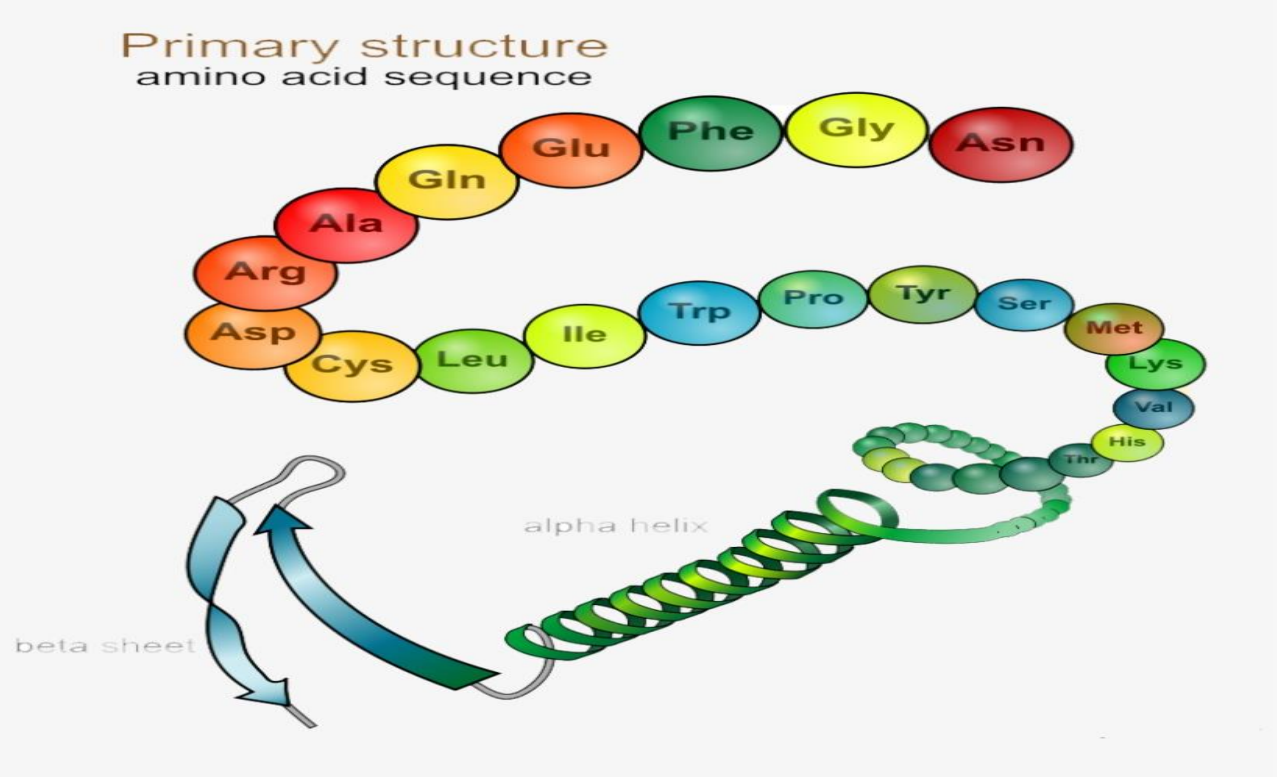


Figure : (2.2) Protein Primary Structure[9]

2.2.2 Protein family:

A protein family is a group of proteins that share a common evolutionary origin, reflected by their related functions and similarities in sequence or structure.

Some of the most famous protein families:

1. G protein coupled receptors (GPCRs)

2. Aminoacyl-tRNA synthetases

3. Intein-containing proteins

4. Translation initiation factors

5. Uncharacterized protein families (UPF)

6. Ligand-gated ion channels

2.2.3: Targeted Proteins:

G-Protein-Coupled-Receptors:

G protein-coupled receptors (GPCRs), also known as seven-(pass)-transmembrane domain receptors, 7TM receptors, heptahelical receptors, serpentine receptors, and G protein-linked receptors (GPLR), form a large group of evolutionarily-related proteins that are cell surface receptors that detect molecules outside the cell and activate cellular responses. Coupling with G proteins, they are called seven-transmembrane receptors because they pass through the cell membrane seven times.

Ion channels:

Ion channels (LICs, LGIC), also commonly referred to as ionotropic receptors, are a group of transmembrane ion-channel proteins which open to allow ions such as Na+, K+, Ca2+, and/or Cl− to pass through the membrane in response to the binding of a chemical messenger (i.e. a ligand), such as a neurotransmitter.

Nuclear receptor:

In the field of molecular biology, nuclear receptors are a class of proteins found within cells that are responsible for sensing steroid and thyroid hormones and certain other molecules.

In response, these receptors work with other proteins to regulate the expression of specific genes, thereby controlling the development, homeostasis, and metabolism of the organism.

2.3: Drug Discovery with Targeted Proteins:

Major protein families as drug targets

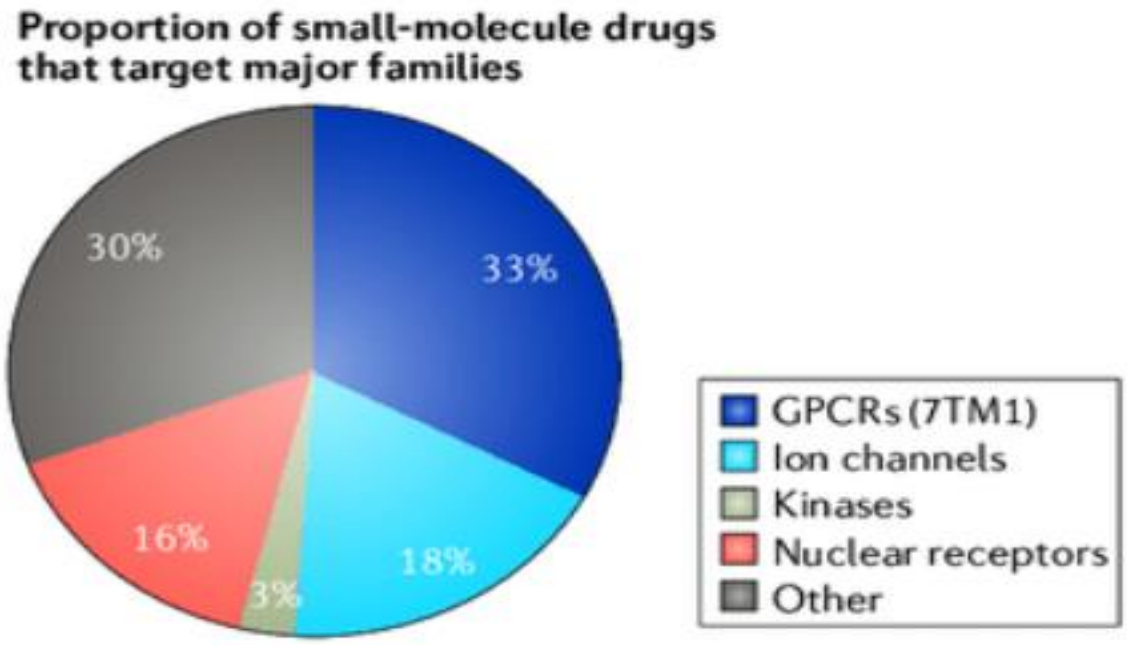


Figure: (2.3) Drug Target Protein Families[10]

1. GPCR

2. Ion Channels

3. Nuclear Receptor

4. kinases

So from the pie chart it is cleared that , GPCR and ion-channels are most frequently used protein families as drug targets.

2.3.1 Drug Discovery with GPCR:

G protein-coupled receptors (GPCRs) are the most intensively studied drug targets, largely due to their substantial involvement in human pathophysiology and their pharmacological tractability.
Here, we report the first analysis of all GPCR drugs and agents in clinical trials. This reveals the current trends across molecule types, drug targets and therapeutic indications, including showing that 481 drugs (~34% of all drugs approved by the FDA) act at 107 unique GPCR targets. Approximately 320 agents are currently in clinical trials, of which ~36% target 64 potentially novel GPCR targets without an approved drug, and the number of biological drugs, allosteric modulators and biased agonists has grown.
The major disease indications for GPCR modulators show a shift towards diabetes, obesity, and Alzheimer's disease, while other central nervous system disorders remain highly represented. Others like High blood pressure, Asthma, Schizophrenia, Heartburn treatment clonidine, bisoprolol, betaxolol, albuterol, nadolol, penbutolol, haloperidol, and olanzapine are frequently used drugs created from Alpha and Beta subfamilies of GPCR

The 227 (57%) non-olfactory GPCRs that are yet to be explored in clinical trials have broad untapped therapeutic potential, particularly in genetic and immune system disorders.
Reflective of the discrepancies between databases, the number of drugs that most frequently target GPCRs varies widely among the sources. Histamine (HRH1), serotonin, dopamine, opioid, and adrenergic receptors are the most frequently targeted GPCRs, in terms of the number of available drugs.

2.4 Literature Review

In course of searching for interesting machine learning tools for classification and prediction, we came across some publications about Analyzing Protein Sequence ([1]-[4]) and their their importance in drug discovery[5]-[7] also went through some relevant papers and websites ([8]-[12]). We were inspired very much by those works to undertake our project-thesis task. Here is a brief description of the most influential of the works we studied.

The research paper by Babasaheb Satpute and Dr. Raghav Yadav named as "Machine Intelligence Techniques for Protein Classification"[1] provided us a good guidance. From it we found important clues for looking into relevant publications and aslo gathered knowledge about extracting features from protein sequences. The journal 'Data Mining Approch for Amino Acid Sequence Classification'[2] by Dr. Sheshang Degadwala and others also helped extracting features from protein sequence. For other features selection like Residue Count, Molecular Weight we primarily relied on the research paper 'Machine Learning Models to Predict Multiclass Protein Classifications' by Yash Parikh Andaman Abdel Fattah [3] . We came to know the importance of GPCR in drug discoveries reading a number of papers like 'G Protein-Coupled Receptors as Targets for Approved Drugs'[4], 'Predicting

Ion Channels Genes and their Types with Machine Learning Techniques' [5] and others used ANN, Naïve Bayes Classifiers, Decision Trees, Random Forest, Extra Tree Classifier as their ML models. Many of them used Evaluation Matrix for comparison and accuracy checking. The paper 'Prediction of potential drug targets based on simple sequence properties' by Qingliang Li and Luhua Lai [6] and 'Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins ' by Sina Ghadermarzi1, Xingyi Li2, Min Li2and Lukasz Kurgan[7] enriched us about how to identify the drugable target proteins of GPCR. 'Clustering of chemical data sets for drug discovery' [8] by Mohamed G. Malhat, Hamdy M. Mousa; Ashraf B. El-Sisi and 'G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs?'[4] by Krishna Sriram and Paul A. Insel[8] used for the classification of the drugable proteins of GCPR to classify them into Heart Disease drug Proteins, High Blood Pressure Drug Protein or Schizophrenia disease drug proteins.

Broad Description:

Paper : "Machine Intelligence Techniques for Protein Classification"[1]-

Many large-scale biological projects and experiments like the Human Genome Project have been carried out in the recent past, resulting in the generation of mammoth biological data, mostly DNA and Protein sequences. It is challenging to manage such a massive amount of data using traditional methods; hence, many biological databases related to protein families and sequences have come into existence. Many databases contain information about proteins

and their families. It is important to know the family of proteins as they play essential functions within the body of organisms like they are the main catalysts for metabolic reactions, transport molecules from one place to another, replicate DNA, etc. Proteins are also important structural constituents of cells in organisms. Thus we need to classify proteins.

Here the known proteins have been classified into protein families and superfamilies based on their functions, structures, or sequence similarity. Whenever a new protein molecule comes across, it is necessary to know the family of that molecule and classify it into one of the known families. It becomes extremely difficult to understand the families of such a large number of proteins using traditional ways. Hence some advanced, fast, efficient, reliable, and intelligent computational techniques are required to expedite the classification process.

And so this research uses advanced Machine Intelligence computing techniques like Artificial Neural Networks, Naïve Bayes Classifiers, Decision Trees, etc., to classify the protein molecules.

The important aspect of the protein classification problem is the selection of protein features. Protein features are based on the protein surface, amino acid sequence, or functions. Various features have been used in the past for classification purposes. It is

essential to select the feature which can yield more accurate results. Here is presented an effective method to classify proteins using Intelligent Computing Techniques like Machine Learning Tools.

Paper: "Machine Learning Models to Predict Multiclass Protein Classifications"[3]-

This paper investigates three machine learning models and a comparison was conducted using different performance measures to determine which algorithm would effectively predict protein classifications based on residue count, structure of molecular weight, and protein sequences. Decision Trees, Random Forests and Extra Trees models are applied on a structural protein sequences dataset. This dataset also contains other features for extraction methods and details on protein structures. Based on the experiments that were conducted on these models, it was demonstrated that Extra Trees model had comparable results but marginally better than the Decision Trees and Random Forests models.

With different diseases that are rampantly spreading worldwide, researching cures and drugs to combat them becomes essential. With protein information collected on the Protein Data Bank (PDB) by the Research Collaboratory for Structural Bioinformatics (RCSB), we can assess the data collected on various protein information. Using such data, such as structures and extraction methods, with machine learning algorithms could provide meaning assistance to finding new and better structure-based drugs for the future.

In our research the featuers protein sequence, molecular mass of the sequence, length of the sequence and also the Gene name of that protein is included as informed in this paper.

Paper : "G Protein-Coupled Receptors as Targets for Approved Drugs[4]" -

Estimates vary regarding the number of G protein-coupled receptors (GPCRs), the largest family of membrane receptors that are targeted by approved drugs, and the number of such drugs that target GPCRs. We review current knowledge regarding GPCRs as drug targets by integrating data from public databases (ChEMBL, Guide to PHARMACOLOGY, and DrugBank) and from the Broad Institute Drug Repurposing Hub. To account for discrepancies among these sources, we curated a list of GPCRs currently targeted by approved drugs. As of November 2017, 134 GPCRs are targets for drugs approved in the United States or European Union; 128 GPCRs are targets for drugs listed in the Food and Drug Administration Orange Book. We estimate that ~700 approved drugs target GPCRs, implying that approximately 35% of approved drugs target GPCRs. GPCRs and GPCR-related proteins, i.e., those upstream of or downstream from GPCRs, represent ~17% of all protein targets for approved drugs, with GPCRs themselves accounting for ~12%. As such, GPCRs constitute the largest family of proteins targeted by approved drugs. Drugs that currently target GPCRs and GPCR-related proteins are primarily small molecules and peptides. Since ~100 of the ~360 human endo-GPCRs (other than olfactory, taste, and visual GPCRs) are orphan receptors (lacking known physiologic agonists), the number of GPCR targets, the number of GPCR-targeted drugs, and

perhaps the types of drugs will likely increase, thus further expanding this GPCR repertoire and the many roles of GPCR drugs in therapeutics.

Reflective of the discrepancies between databases, the number of drugs that most frequently target GPCRs varies widely among the sources (Figure 4). Histamine (HRH1), serotonin, dopamine, opioid and adrenergic receptors are the most frequently targeted GPCRs, in terms of the number of available drugs.

From this research paper we learned how to classify GCPR drugable protein class into different drug target proteins like for Durability for Heart disease, High blood pressure, and Schizophrenia.

Chapter 3:

Dataset Analysis

3.1 Feature Selection:

The features, which we are choosing here -Protein Sequence, protein length, Molecular mass and Gene.

Our total features will be 23. Among them, 20 features are extracted from protein sequence

The features, which we are using here -Protein Sequence(Extracted - 20 Features), protein length, Molecular mass and Gene.

Protein sequence:

Each protein or peptide consists of a linear sequence of amino acids. The amino acid sequence of a protein or peptide is useful information to understand the protein or peptide, identify it in a sample and categorize its post-translational modifications. The process of determining the amino acid sequence is known as protein sequencing.

Molecular mass:

The molecular weight (mw) or mass of a protein can be determined by summation of the mw of its corresponding amino acid sequence. Certain modifications to this sequences can result in changes to the mw.

Protein Molecular mass varies due to the formation of amino acid sequence and depends on how the amino acids are bond to each other. So same protein structure can have different weights. Here molecular mass unit is Dalton . Dalton is a unit used in expressing the molecular weight of proteins, equivalent to atomic mass unit.

Gene name:

In biology, a gene is a basic unit of heredity and a sequence of nucleotides in DNA that encodes the synthesis of a gene product, either RNA or protein. During gene expression, the DNA is first copied into RNA. The RNA can be directly functional or be the intermediate template for a protein that performs a function. A particular protein is formed or structured from a gene. That is why gene name is important to express the full characteristics of a protein.

Protein length:

Length is the number of amino acids in a protein sequence. It can be derived from the sequence. In our research we have taken protein of various length (from least 17 to highest 1000) . Protein length is an important feature of the protein because different class of protein has different protein length range.

Entry name:

Every protein has a specific entry name is uniprot. So it is easy to find out more details of a protein from uniprot.

3.2 Overview of Our dataset:

We have total 4 datasets of a total of 3400 data:

1st Datset contains - 1800 data

2nd Dataset contains - 700 data

3rd Dataset contains - 300 data

4th Dataset contains - 600 data

3.3 Dataset Source:

We have taken our data from UniProt Organization.

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR).

Website Link: https://www.uniprot.org

3.4 Dataset Description:

Dataset1:

Our 1st Dataset contains 3 types of Protein Class (GPCR,Ion-Channel,Neuclear Receptor)

ID	Entry	Protein names	Gene names	Length	Mass	Sequence
1	Q96MV8	GPCR	Palmitoyltransferase ZDH	337	39331	MRRGWKMALSGGLRCCRRVLSWVPVLVIVLVVLWSYYAYVFELCLVTVLSPAEKVIYLILYHAIFVFFTV
2	O88427	ION CHANNEL	Cacna1h Kiaa1120	2365	262030	MTEGTLAADEVRVPLGASPSAPAAPVRASPASPGVPGREEQRGSGSSVLAPESPGTECGADLGADEEQ
3	Q95223	GPCR	ARRB1	410	46360	MGDKGTRVFKKASPNGKLTVYLGKRGFVDHIDLVDPVDGVVLVDPEYLKERRVYVTLTCAFRYGREDL
4	P29275	GPCR	Adenosine receptor A2b	332	36333	MLLETQDALYVALELVIAALSVAGNVLVCAAVGTANTLQTPTNYFLVSLAAADVAVGLFAIPFAITISLGF
5	P35499	ION CHANNEL	SCN4A	1836	208061	MARPSLCTLVPLGPECLRPFTRESLAAIEQRAVEEEARLQRNKQMEIEEPERKPRSDLEAGKNLPMIYG
6	B0ZBE0	GPCR	Alpha-1D adrenergic rece	572	60463	MTFRDLLSVSFEGPRPDSSAGGSSAGGGGGSAGGAAPSEGPAVGGVPGGAGGGGGVVGAGSGEDNR
7	Q6QT55	Neuclear Receptor	AR NR3C4	895	96537	MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLQQQQQQQQETSPI
8	A5YC96	GPCR	ackr3a cxcr7b si:dkey-19	313	35070	TLYAFIFVVGLAANALVVWVNMRSQRHYHETHMYILNLAVADLCVVATLPVWVSSLAQGGHWAFGC
9	A0A3P9B6A2	GPCR	ACKR3	463	52075	MYYLSLVAFLLPFLRSSLFLALSLTPPQFHCSAAEAVINHAEWGCLCGRFFFSHLNRSVGGPGETACDPL
10	A0A3Q7UYJ2	GPCR	HRH1	487	55784	MNLPNSSCIFEDKMCEGNKTTIANPKLMPLVVFLSAISLVTVGLNLLVLYAVRSERKLHTVGNLYIVSLS
11	A0A6P5DJR5	ION CHANNEL	ZACN	410	45555	MMALRLLLHLTFLGLGTTQPLAQQQGFSVPAFDWPSSSNLNSPQEVPDLIQIPNNGSKPLVVDVQVFV
12	Q6UXB2	GPCR	C-X-C motif chemokine 1	119	13819	MKVLISSLLLLLPLMLMSMVSSSLNPGVARGHRDRGQASRRWLQEGGQECECKDWFLRAPRRKFMT
13	Q14B80	ION CHANNEL	Kcnc2	642	70503	MGKIESNERVILNVGGTRHETYRSTLKTLPGTRLALLASSEPQGDCLTAAGDKLQPLPPPLSPPPRPPPLS
14	A0A851W1M7	GPCR	Ackr3 COPSEC_R11898	361	41308	LDLTSILDFLETANLTEINWTCNNSECITVDATTCSGTLNKSALLYTLSFFYIFIFVIGLVANSVVVWVNLC
15	Q9JIS7	ION CHANNEL	Cacna1f	1985	221926	MSESEVGKDTTPEPSPANGTGPGPEWGLCPGPPTVGTDTSGASGLGTPRRRTQHNKHKTVAVASAQF
16	P62956	ION CHANNEL	Cacng7	275	31003	MSHCSSRALTLLSSVFGACGLLLVGIAVSTDYWLYMEEGTVLPQNQTTEVKMALHAGLWRVCFFAGR
17	Q5BJR8	Neuclear Receptor	Rxrg Nr2b3	463	50893	MYGNYSHFMKFPTGFGGSPGHTGSTSMSPSVALPTGKPMDSHPSYTDTPVSAPRTLSAVGTPLNALGS

Figure:(3.1) Dataset 1

Dataset 2:

The 2nd Dataset contains only the Drugable proteins of GPCR class:

ID	Entry	Status	Protein names	Gene names	Length	Mass	Sequence
1	A0A024R3C5	Druggable	D(2) dopamine receptor	DRD2	443	50619	MDPLNLSWYDDDLERQNWSRPFNGSDGKADRPHYNYYATLLTLLIAVIVFG
2	P35790	Druggable	Choline kinase alpha (CK	CHKA	457	52249	MKTKFCTGGEAEPSPLGLLLSCGSGSAAPAPGVGQQRDAASDLESKQLGGQC
3	A0A024R645	Druggable	Endothelin receptor type	EDNRB	436	48710	MQPPPSLCGRALVALVLACGLSRIWGEERGFPPDRATPLLQTAEIMTPPTKTL
4	A0A0C5B5G6	Druggable	Mitochondrial-derived p	MT-RNR1	16	2175	MRWQEMGYIFYPRKLR
5	A0A0D9SBU1	Non Druggable	Histamine H1 receptor	HRH1	487	55592	MTLPNSSCLLEDKMCEGNKTTMASPQLMPLVVVLSTISLVTVGLNLLVLYAV
6	A0A0J9YWR0	Druggable	Adenosine receptor A3	ADORA3	123	13255	MPNNSTALSLANVTYITMEIFIGLCAIVGNVLVICVVKLNPSLQTTTFYFIVSLA
7	A0A1U7QEW2	Non Druggable	Histamine H1 receptor	Hrh1	489	55786	MSLPNISSAFEDKMCEGNRTAMASPQLLPLVVVLSSISLVTVGLNLLVLYAVH
8	P35503	Druggable	UDP-glucuronosyltransfe	UGT1A3	534	60338	MATGLQVPLPWLATGLLLLLSVQPWAESGKVLVVPIDGSHWLSMREVLREL
9	P43088	Druggable	Prostaglandin F2-alpha r	PTGFR	359	40055	MSMNNSKQLVSPAAALLSNTTCQTENRLSVFFSVIFMTVGILSNSLAIAILMK
10	P43354	Druggable	Nuclear receptor subfan	NR4A2	598	66591	MPCVQAQYGSSPQGASPASQSYSYHSSGEYSSDFLTPEFVKFSMDLTNTEITA
11	A0A1U7TVX5	Non Druggable	Histamine H1 receptor	HRH1	487	55391	MTLPNSSCLLEDNMCEGNKTTMANPQLMPLVVVLSTISLITVGLNLLVLYAV
12	P35504	Druggable	UDP-glucuronosyltransfe	UGT1A5	534	60071	MATGLQVPLPQLATGLLLLLSVQPWAESGKVLVVPTDGSHWLSMREALRDL
13	A0A1V4JVS7	Non Druggable	Histamine H1 receptor	HRH1	475	54359	MSKNTTVNFTNSQLALLGLFLGSISIITIVMNILVLCAVKTEKKLQTVGNLYIVS
14	A0A1Y8EK52	Druggable	D(2) dopamine receptor	DRD2	7	789	MDPLNLS
15	A0A218UNW4	Non Druggable	Histamine H2 receptor ((	HRH2	374	41787	MRIISDTMDPCYNHTSSQKSNHTSSQVFPLQVLVGFCLFTLIVVTLCGNIIVCL

Figure:(3.2) Dataset 2

Dataset 3:

3rd Dataset is the reviewed data of Proteins in GPCR class that is used as drugs for curing the Heart Disease, High Blood Pressure, Schizophrenia:

ID	Entry	Status	Protein names	Gene names	Length	Mass	Sequence
1	P08588	Druggable for Heart Disease	Beta-1 adrenergic recept	ADRB1	477	51323	MGAGVLVLGASEPGNLSSAAPLPDGAATAARLLVPASPPASLLPPASESPEPLSQQW
2	Q9H244	Druggable for Heart Disease	P2Y purinoceptor 12 (P2Y	P2RY12 HORK3	342	39439	MQAVDNLTSAPGNTSLCTRDYKITQVLFPLLYTVLFFVGLITNGLAMRIFFQIRSKSN
3	Q6ZMP9	Druggable for Heart Disease	cDNA FLJ16773 fis		394	43309	MTAGRSQERRAQEMGRGSVQGLDLKGDLEFFTAPMLSLRSFVFVGVGSGLTSSHIF
4	X5DNB4	Druggable for Heart Disease	Adenosine receptor A2 (F	ADORA2A	412	44707	MPIMGSSVYITVELAIAVLAILGNVLVCWAVWLNSNLQNVTNYFVVSLAAADIAVG
5	P24530	Druggable for High Blood Pressure	Endothelin receptor type	EDNRB ETRB	442	49644	MQPPPSLCGRALVALVLACGLSRIWGEERGFPPDRATPLLQTAEIMTPPTKTLWPK
6	Q8IVQ6	Druggable for High Blood Pressure	Palmitoyltransferase ZDH	ZDHHC21	265	31385	MGLRIHFVVDPHGWCCMGLIVFVWLYNIVLIPKIVLFPHYEEGHIPGILIIIFYGISIFC
7	P32121	Druggable for High Blood Pressure	Beta-arrestin-2 (Arrestin	ARRB2 ARB2 ARR2	409	46106	MGEKPGTRVFKKSSPNCKLTVYLGKRDFVDHLDKVDPVDGVVLVDPDYLKDRKVFV
8	P05305	Druggable for High Blood Pressure	Endothelin-1 (Preproendo	EDN1	212	24425	MDYLLMIFSLLFVACQGAPETAVLGAELSAVGENGGEKPTPSPPWRLRRSKRCSCSS
9	A0A286YF(	Druggable for High Blood Pressure	Alpha-1B adrenergic rece	ADRA1B	35	3712	MNPDLDTGHNTSAPAHWGELKNANFTGPNQTSSNS
10	P49407	Druggable for High Blood Pressure	Beta-arrestin-1 (Arrestin	ARRB1 ARR1	418	47066	MGDKGTRVFKKASPNGKLTVYLGKRDFVDHIDLVDPVDGVVLVDPEYLKERRVYVT
11	Q9BXX6	Druggable_For_Schizophrenia	Dopamine D3 receptor (F	DRD3	90	9656	MASLSQLSSHLNYTCGAENSTGASQARPHAYYALSYCALILAIVFGNGLVCMAVLKE
12	B0ZBE0	Druggable for High Blood Pressure	Alpha-1D adrenergic rece	ADRA1D hCG_164	572	60463	MTFRDLLSVSFEGPRPDSSAGGSSAGGGGGSAGGAAPSEGPAVGGVPGGAGGGGG
13	P04899	Druggable_For_Schizophrenia	Guanine nucleotide-bindi	GNAI2	355	40451	MGCTVSAEDKAAAERSKMIDKNLREDGEKAAREVKLLLLGAGESGKSTIVKQMKIIH
14	F8WAN1	Druggable for Heart Disease	Cytospin-A	SPECC1L	911	101514	MKKASRSVGSVPKVSAISKTQTAEKIKPENSSSASTGGKLVKPGTAASLSKTKSSDDLL
15	Q6UWT4	Druggable for High Blood Pressure	Uncharacterized protein (	C5orf46	87	9693	MAVSVLRLTVVLGLLVLFLTCYADDKPDKPDDKPDDSGKDPKPDFPKFLSLLGTEIIEI

Figure:(3.3) Dataset 3

Dataset 4:

Unknown protein sequences that will be classified

ID	Entry	Status	Gene names	Length	Mass	Sequence
1	A0A218V9S6	unreviewed	F2RL1_0 RLOC_00(	390	43945	MVARRGLCLLLWCALLGAAAAAGENDGSSKPKGRSFIGYKAQNANNSEELYEVDEFVAEVL
2	S5TLS4	unreviewed	CNR1	472	52858	MKSILDGLADTTFRTITTDLLYVGSNDIQYEDIKGDMASKLGYFPQKFPLTSFRGSPFQEKMT
3	Q3TU81	unreviewed	F2rl1	399	44752	MRSLSLAWLLGGITLLAASVSCSRTENLAPGRNNSKGRSLIGRLETQPPITGKGVPVEPGFSI
4	A0A250YGM0	unreviewed	CCR2 Ccr2	370	41815	MKDTYLQISRSLLSTSQSLLESINDSSEEVTTIYDYGSSEPCYKPKVRQVAARLLPPLYSLVFIFG
5	A0A286XGT3	unreviewed	F2RL1	399	43966	MLGSRAAWLLGGALLLAAAASGSSTGPGAGVNKTSKGRSLIGKNHERLPVTKKGVTVAPGFA
6	Q0VBE5	unreviewed	Ghsr	364	40969	MWNATPSEEPEPNVTLDLDWDASPGNDSLSDELLPLFPAPLLAGVTATCVALFVVGISGNLI
7	D4A8L8	unreviewed	Fpr3 Fpr2	351	39461	METNYSIPMSGSEVMVNDSTISRVLWILTMVVLSITFVLGVLGNGLVIWVAGFRMAHTVTT
8	A0A250YHJ2	unreviewed	ADRB2 Adrb2	418	46894	MGQPGNDSDFLLAPNGSQAPGHDITQERDEAWVVGMAMVMSLIVLATVFGNVLVITAVA
9	G5B2D1	unreviewed	ADRB2 GW7_0982	418	47078	MGHLGNGSDFLLAPNTSYAPDRNVTQERDEAWVVGMAIVMSLIVLAIVFGNVLVITAIAKF
10	I3LZ09	unreviewed	CCR2	373	42619	MEDKDALPQFIHNILSTSHSLFVRSNKGSSEESTTTYDYDYSSPCYKPDVKHIGALLLPPLYSLV
11	A0A1S3FXY0	unreviewed	F2rl1	397	44267	MRSPRVAWLLGGAVLLAAAASGGVTGEGANQTSRGRSLIGRNKGHSSVTVKGVKGEPGFTV
12	I3NAX4	unreviewed	F2RL1	397	44370	MRSRSLMWLLGGAILLAASASCSQTTRGSNRTSKGRSLIGKPNKPPSVTGKGVTVVPDFSVD
13	Q8VBU7	unreviewed	Adrb2 rCG_46851	418	46960	MEPHGNDSDFLLAPNGSRAPGHDITQERDEAWVVGMAILMSVIVLAIVFGNVLVITAIAKFE
14	X5DNB4	unreviewed	ADORA2A	412	44707	MPIMGSSVYITVELAIAVLAILGNVLVCWAVWLNSNLQNVTNYFVVSLAAADIAVGVLAIPF
15	F6WW26	unreviewed	F2R	428	47631	MGPRRLLLVAAGLSLCGPLLSARTRGRRPGSKATNDTVDPRSFILSNSHDQFEPFPMEEGYE

Figure:(3.4) Dataset 4

3.5: Dataset Preprocessing:

3.5.1 Feature Extraction:

We extracted features for every protein sequence in each family. In each protein sequence, we computed the distance of each amino acid residue from the first residue. Then took the mean of all distances of a particular amino acid residue from first residue. For example, if a residue 'A' occurs 4 times in a particular sequence, then for 'A' we will get 4 distances for 'A' i.e. for every occurrence of 'A' we will compute its distance from the first residue. Then we will take the mean of those 4distances as 12 which will be the feature value for 'A'. Thus, for every protein sequence, we will get 20 feature values as there are 20 different amino acid residues that occur multiple times in a sequence. Thus we prepare a dataset of size 2000 X 20 i.e. for each protein there are 20 features and we have 2000 total number of proteins

Sequence

M I K T A L L F F A T A L C E I I G C F A

Number of Alanine(A) = 4

Vector distances of A

5	**10**	**12**	**21**

Average Vector Distance = (5+10+12+21)/4 = 48/4 = 12

Number of Leucine(L) = 3

Vector distances of L

6	**7**	**13**

Average Vector Distance = (6+7+13)/3 = 26/3 = 7

Average vector distance of amino acids

Alanine (A)	Methionine (M)	Isoleucine (I)	Leucine (L)
12	**1**	**13**	**7**

Sample dataset:

Sequence
MRRGWKMALSGGLRCCRRVLSWVPVLVIVLVVLWSYYAYVFELCLVTVLSPAEKVIYLILYHAIFVFFTWTYWKSIFTLPQQPNQKFHL
MTEGTLAADEVRVPLGASPSAPAAPVRASPASPGVPGREEQRGSGSSVLAPESPGTECGADLGADEEQPVPYPALAATVFFCLGQTTRI
MGDKGTRVFKKASPNGKLTVYLGKRGFVDHIDLVDPVDGVVLVDPEYLKERRVYVTLTCAFRYGREDLDVLGLTFRKDLFVANVQSFPI
MLLETQDALYVALELVIAALSVAGNVLVCAAVGTANTLQTPTNYFLVSLAAADVAVGLFAIPFAITISLGFCTDFYGCLFLACFVLVLTQS
MARPSLCTLVPLGPECLRPFTRESLAAIEQRAVEEEARLQRNKQMEIEEPERKPRSDLEAGKNLPMIYGDPPPEVIGIPLEDLDPYYSNK
MTFRDLLSVSFEGPRPDSSAGGSSAGGGGGSAGGAAPSEGPAVGGVPGGAGGGGGVVGAGSGEDNRSSAGEPGSAGAGGDVNGTA
MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLQQQQQQQQETSPRQQQQQQQQGEDGSPQAHF
TLYAFIFVVGLAANALVVWVNMRSQRHYHETHMYILNLAVADLCVVATLPVWVSSLAQGGHWAFGQAACKLTHLLFSVNLFASIFFL
MYYLSLVAFLLPFLRSSLFLALSLTPPQFHCSAAEAVINHAEWGCLCGRFFFSHLNRSVGGPGETACDPLPDSVPSDPPWIRTDPDRAAL
MNLPNSSCIFEDKMCEGNKTTIANPKLMPLVVFLSAISLVTVGLNLLVLYAVRSERKLHTVGNLYIVSLSVADLIVGAVVMPMNILYLLM
MMALRLLLHLTFLGLGTTQPLAQQQGFSVPAFDWPSSSNLNSPQEVPDLIQIPNNGSKPLVVDVQVFVSNVFNVDILRYTVSSTLLLRI
MKVLISSLLLLLPLMLMSMVSSSLNPGVARGHRDRGQASRRWLQEGGQECECKDWFLRAPRRKFMTVSGLPKKQCPCDHFKGNVKI
MGKIESNERVILNVGGTRHETYRSTLKTLPGTRLALLASSEPQGDCLTAAGDKLQPLPPPLSPPPRPPPLSPVPSGCFEGGAGNCSSHGG
LDLTSILDFLETANLTEINWTCNNSECITVDATTCSGTLNKSALLYTLSFFYIFIFVIGLVANSVVVWVNLQAKMTGYETHLYIFNLAIADL

Table: (3.5) Feature Extraction

Sample Feature extraction from our dataset:

Alanine	Arginine	Aspartic Acid	Asparagine	Cysteine	Glutamine	Glutamic Acid	Glycine	Histidine	Isoleucine	Leucine	I
155.9230769	114.933333	225.272727	233.3571429	142.6666667	180.7	214.1111111	195.625	162.6	164.6428571	149.2352941	
1226.832402	1130.50877	1304.89286	1072.697368	1094.793103	1170.257143	1280.661972	1124.45122	1124.746479	1122.378641	1167.739837	
205.0909091	203.84	228.580645	248.5294118	176.5714286	194.1428571	237.5	173.4545455	210.2	226.125	194.8205128	
143.75	210.75	152.555556	190.75	160	206.2727273	136.8571429	178.6315789	257	176.9473684	145.6888889	
859.7913043	879	922.765957	904.8764045	857.1025641	955.6862745	913.0787402	951.5043478	925.52	974.1366906	896.225	
277.1	359.711111	273.875	223.5555556	321.5882353	407.5	339.8461538	191.3174603	357.6666667	280.9333333	293.2037037	
382.7831325	472.414634	461.162162	567.3157895	487.5185185	336.1403509	416.5	386.2045455	521.4736842	657.0434783	468.7317073	
143.2666667	149.8125	183.444444	152.6	173.2	146.3333333	200.8	177.6428571	156	160.1333333	157.9069767	
192.5151515	242.73913	236.6	227.8421053	224.4375	245.4285714	203.6666667	194.6875	217.6153846	271.75	220.9459459	
224.9166667	263.76	244	247.4	264.6666667	276.2222222	275.3333333	256.7826087	274.7	255.21875	201.8627451	
260.0344828	209.947368	172.294118	165.3684211	265	174	230.6470588	257	206.2727273	219.6666667	191.03125	
68.8	65.1538462	55.3333333	68.33333333	77.83333333	80.22222222	48.33333333	48.28571429	80.66666667	5	45.8125	
320.6590909	355.195122	278.678571	364.2692308	321.6875	305.6	271.0833333	271.3103448	255.75	348.1212121	332.1549296	
198.9090909	228.222222	146.111111	153.6111111	173.0714286	211.1428571	198.625	162	196.75	173.0645161	169.4772727	
1043.15942	997.783784	1103.28235	931.0933333	903.175	1206	1061.007194	934.6241135	947.3125	979.5508475	972.5848214	
127.7777778	153.8125	199.714286	125.625	109.3333333	169.625	123.875	116.8125	157.2857143	153.6666667	107.962963	

Table: (3.6) Feature Extraction

Code for Extraction

```
import pandas as pd
import numpy as np
# Then loading csv file
    df = pd.read_csv('Dataset1.csv')
# converting ;FRUIT_NAME' column into list
    a = list(df['Sequence'])
# converting list into string and then joining it with space
    b = ' '.join(str(e) for e in a)

#for a single amino acid in a sequence in this case Alanine
    for i in range(len(b)):
        resedueCount=resedueCount+1
        if b[i] =="A":
            countA= countA+1
            newlistA.append(resedueCount)
        if b[i] == " ":
            total_distanceA = sum(newlistA)
            if countA!=0:
                avgdisA=total_distanceA/countA
                newAvglistA.append(avgdisA)
            else:
                newAvglistA.append(0)
```

This particular part of the code will calculate the average distance of Alanine amino acids in a sequence. Thus for 20 amino acids in a sequence is also calculated.

3.5.3 Null value handling

A null value in a relational database is used when the value in a column is unknown or missing. A null is neither an empty string (for character or datetime data types) nor a zero value (for numeric data types).

For Handling null value, we have deleted the specific row.

df1 = df.dropna()

3.5.4 Label Encoding:

Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering. The main target of encoding is to covert the strings or alphabets in machine readable numeric values.

```
#Applying Label Encoding

from sklearn.preprocessing import LabelEncoder

categorical_cols = ['Gene names']

# instantiate labelencoder object

le = LabelEncoder()

# apply le on categorical feature columns

df1[categorical_cols] = df1[categorical_cols].apply(lambda col: le.fit_transform(col))
```

Chapter 4
Implementation of the Algorithms and Result Analysis

Approach to Achieve Our Goal
4.1: Supervised ML model

Step 1:

At first, we classified Proteins into one of these three protein families – GPCR (G – Protein Coupled Receptors), Ion- Channels, and Nuclear Receptors.

Here Dataset-1 is used as train and test set. Then we used this model to test again for unknown protein sequences. K_Nearest_Neighbour ML model is used for step -1.

4.1.1: K-Nearest Neighbor (KNN) Algorithm for Machine Learning

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Step 2:

Here we checked whether a GPCR protein can be used in making drugs or not. We classified dataset-2 GPCR Proteins into Druggable or NON-Druggable. First, we train and test dataset 2 with ML model Decision Tree. Then we used this model to test the durability of unknown protein sequences predicted as GPCR Proteins in step-1.

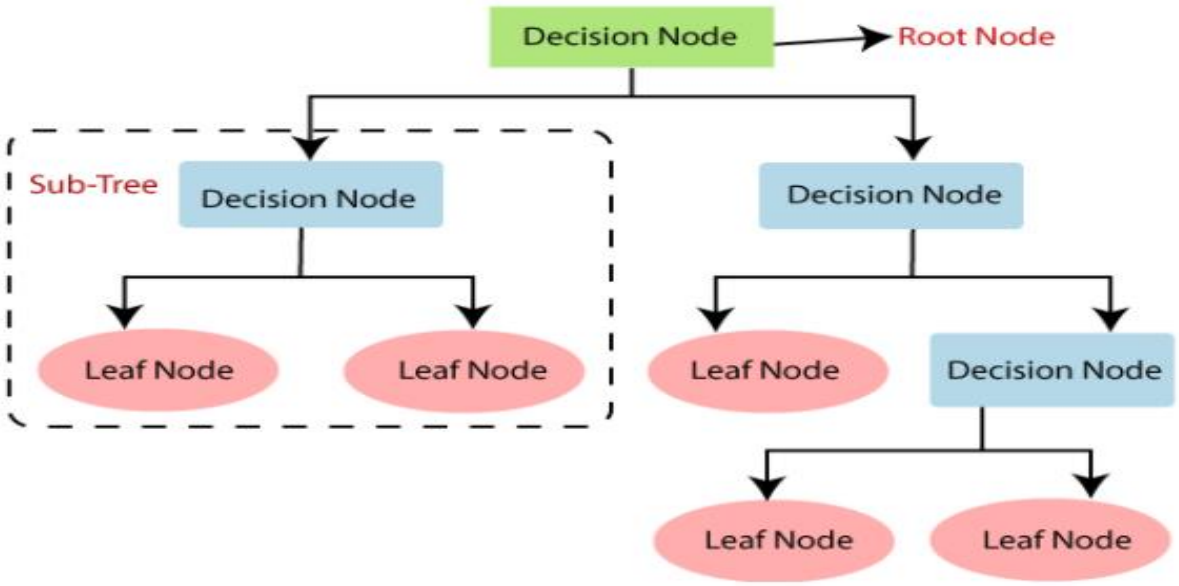


Figure: (4.1) Decision Tree

4.1.2: Decision Tree Classification Algorithm

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Below diagram explains the general structure of a decision tree:

Step 3:

In the last stage (dataset 3), Druggable GPCR Proteins are further classified into 3 drug classes - Durability for Heart disease, High blood pressure, and Schizophrenia. We train this dataset with the unsupervised ML model K_means_clustering.

Using this ML model we test the unknown sequences taken from step-2 predicted as Druggable GPCRs'.

So for the unknown proteins, firstly we classified them in GPCR, Ion channels and nuclear receptors. The Proteins which are classified into GPCR are further classified into Duggable and Non-Druggable. The proteins, which are classified as Druggable, was again tested their durability that can be used for making which types of medicines or drugs.

4.2: Unsupervised ML model

4.2.1: K-means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. A cluster refers to a collection of data points aggregated together because of certain similarities. A target number *k*, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.
Every data point is allocated to each of the clusters by reducing the in-cluster sum of squares.

In other words, the K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.
The *'means'* in the K-means refers to averaging of the data; that is, finding the centroid.

How the K-means algorithm works

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

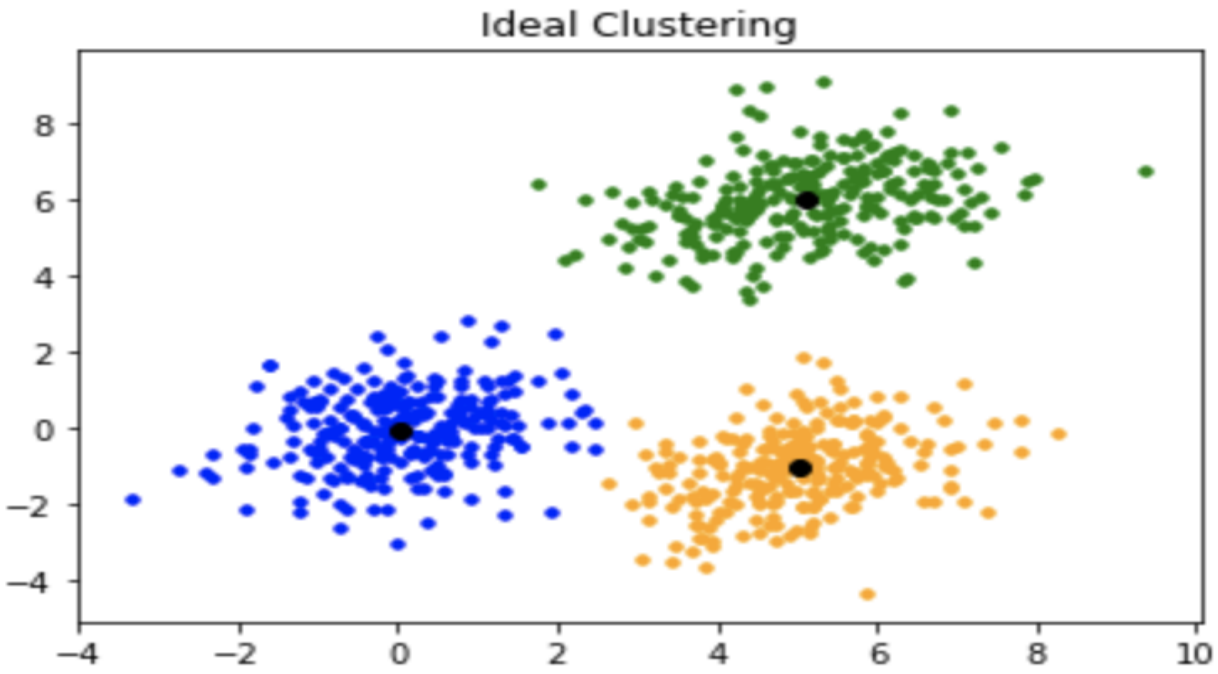


Figure: (4.2) K means Clustering

4.2.2: Gaussian Clustering

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

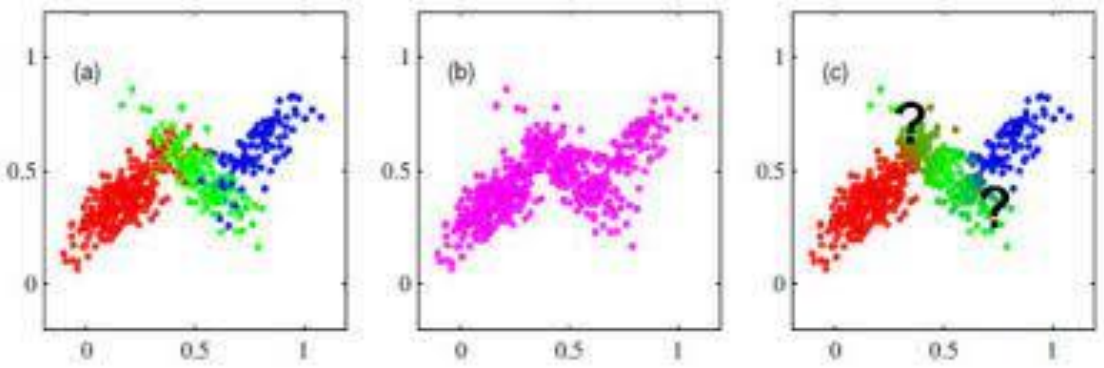


Figure: (4.3) Gaussian Clustering

4.3 Result Analysis:

4.3.1 K_Nearest_Neighbour:

Confusion Matrix

	Actual GPCR	Actual Ion Channel	Actual Nuclear Receptor	Total Predicted
Predicted GPCR	120	11	13	144
Predicted Ion Channel	11	84	18	113
Predicted Nuclear Receptor	10	15	121	146
Total Actual	141	110	152	

Classification Report

	precision	recall	f1-score	support
GPCR	0.82	0.87	0.84	139
ION Channel	0.79	0.75	0.77	123
Nuclear Receptor	0.77	0.75	0.76	91

Accuracy			0.80	353
Macro avg	0.79	0.79	0.79	353
Weighted avg	0.80	0.80	0.80	353

4.3.2 Decision Tree classifier:

Confusion Matrix

	Actual Drugable	Actual Non-Drugable	Total Predicted
Predicted Drugable	46	4	50
Predicted Non-Drugable	12	75	87
Total Actual	58	79	

Classification Report

	precision	recall	f1-score	support
Druggable	0.78	0.90	0.83	50
Non Druggable	0.94	0.85	0.89	87
Accuracy			0.87	137
Macro avg	0.86	0.88	0.86	137
Weighted avg	0.88	0.87	0.87	137

4.3.3 K Means Clustering

Silhouette Score 0.733

Chapter 5
Discussion and Conclusion

BIBLIOGRAPHY

1) Babasaheb Satpute and Dr. Raghav Yadav named "Machine Intelligence Techniques for Protein Classification", 2018 3rd International Conference for Convergence in Technology (I2CT),06-08 April 2018.

2) Degadwala, D. S. ., & Vyas, D. . (2021). 'Data Mining Approach for Amino Acid Sequence Classification' . International Journal of New Practices in Management and Engineering, 10(04), 01–08.

3) Yash Parikh; Eman Abdelfattah. 'Machine Learning Models to Predict Multiclass Protein Classifications' 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 10-12 October 2019.

4) Krishna Sriram and Paul A. Inse. 'G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs?' 2018 Apr; 93(4): 251–258

5) Ke Han, Miao Wang, Lei Zhang, Ying Wang, Mian Guo, Ming Zhao, Qian Zhao, Yu Zhang , Nianyin Zeng and Chunyu Wang. 'Predicting Ion Channels Genes and Their Types With Machine Learning Techniques'. 03 May 2019

6) Qingliang Li and Luhua Lai. 'Prediction of potential drug targets based on simple sequence properties'. BMC Bioinformatics volume 8, Article number: 353 (2007). 20 September 2007

7) Sina Ghadermarzi1, Xingyi Li2, Min Li2and Lukasz Kurgan 'Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins '.Front. Genet., 15 November 2019

8) M. G. Malhat, H. M. Mousa and A. B. El-Sisi, "Clustering of chemical data sets for drug discovery," 2014 9th International Conference on Informatics and Systems, 2014, pp. DEKM-11-DEKM-18.

9) Website: https://www.seekpng.com/ima/u2w7r5u2w7w7i1o0/

10) 2.Rita Santos, Oleg Ursu, Anna Gualton, "A Comprehensive Map of Molecular Drug Target",Nature Reviews Drug Discovery 16(1), December 2016, DOI:10.1038/nrd.2016.230,

11) Website:https://www.ajinomoto.com/aboutus/amino-acids/20-amino-acids (ACCESSED : 18/6/2019)

12) Website: https://cbd.cmu.edu/about-us/what-is-computational-biology.html (ACCESSED : 18/6/2019)

13) Website: https://www.mcgill.ca/biochemistry/about-us/information/biochemistry#:~:text=Biochemistry%20is%20the%20application%20of,the%20chemistry%20of%20living%20systems. (ACCESSED : 18/6/2019)

14) Website: https://medlineplus.gov/genetics/understanding/howgeneswork/protein/#:~:text=Proteins%20are%20large%2C%20complex%20molecules,the%20body's%20tissues%20and%20organs. (ACCESSED : 18/6/2019)

15) https://owlcation.com/stem/protein-production-a-step-by-step-illustrated-guide?fbclid=IwAR2Pro0ZGgzwu_JRL24CEFktVBuneheJSrCLkjdYZekdnREvBfoah_UfDqQ

Appendices

Appendix A

Step -1: K means clustering:

```
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors=4)
classifier.fit(x_train, y_train)
y_pred = classifier.predict(x_test)
```

Appendix B

Step -2: Decision tree

```
print('Decision Tree classifier :')
from sklearn.model_selection import train_test_split
from sklearn import tree
#import matplotlib.pyplot as plt

X_train, X_test, Y_train, Y_test = train_test_split(x_2, y_2, test_size = 0.2, random_state=1)

dt_clf = tree.DecisionTreeClassifier(max_depth=10)
dt_clf.fit(X_train, Y_train)

#y_score = dt_clf.score(x_test, y_test)

y_pred_2 = dt_clf.predict(X_test)
```

```
#print(y_pred)
print('Accuracy :')
#print(dt_clf.score(x_test, y_test))
percentage = "{:.0%}". format(dt_clf.score(X_test, Y_test))
print(percentage)

y_pred_2 = dt_clf.predict(X_test)

#confusion matrix decision tree
from sklearn.metrics import confusion_matrix, precision_score
#print(precision_score(y_pred, y_score))
#print(confusion_matrix(Y_test, y_pred_2))
print(classification_report(Y_test, y_pred_2))
```

Appendix C:

Step -3: K-means Clustering

```
#Applying Algorithm
from sklearn.cluster import KMeans
x = df2.iloc[:,:]
kmeans = KMeans(3)
kmeans.fit(x)
identified_clusters = kmeans.fit_predict(x)
```

Appendix D:

Step -4: Gaussian Clustering

```
from sklearn.mixture import GaussianMixture
GaussianMixture_model = GaussianMixture(n_components = 3)
Guassian_pred = GaussianMixture_model.fit_predict(x)
print(Guassian_pred)

countH=0
countB=0
countS=0

for cluster in range(len(x_test_druggable_unknown)):
    if identifiedbyguassian[cluster]==1:
        countH+=1
        print(cluster+1 ,index[cluster],' :Druggable for Heart Disease')
    elif identifiedbyguassian[cluster]==0:
        countB+=1
        print(cluster+1,index[cluster],' :Druggable for High Blood Pressure')
    elif identifiedbyguassian[cluster]==2:
        countS+=1
        print(cluster+1,index[cluster],' :Druggable for Schizophrenia')

print('Proteins found for Heart Disease:',countH)
print('Proteins found for High Blood Pressure:',countB)
print('Proteins found for Schizophrenia:',countS)
```